

**LOGNET**  
**Innovating, Developing & Delivering**  
**The Defence Support Network**

on

Thursday, 12<sup>th</sup> March 2020

---

Transcribed from the Audio Recording

---

**CHRIS KUTARNA:**

Thank you very much. You know, there's a chapter in Age of Discovery on pandemics; if anybody wants a copy of the chapter. *[Laughter]* But I thought that I would take my turn addressing the microbe in the room, and just say that I actually think that it's brilliant that we are meeting at a time like now, in a context like now, because this is the perfect moment to be evaluating our blind spots, including our blind spots when it comes to artificial intelligence. Next slide please. It works. Human communication, did you see that? Anyway. Because we are, you know, still in the opening chapter of a massive and rapid adoption of artificial intelligence technologies across every sector of society, from how we share and consume information, to our relationships with one another, and, of course, to how we provide and ensure security.

One of my good friends in industry, he's the head of the data science and AI group big for one of the global consultancies, got about 800 data scientists that work under him, and I said, "Look, I'm coming to this meeting today, and I would love to have a fresh, current example of the typical work that you're doing on AI, for your large industrial clients today," and he said, "I've got the perfect example for you, Chris," and so he gave me a slide; this is a real slide, I've anonymised it appropriately, but it is for automating a single function inside one of the large banks here in London, okay? Are you ready for this slide? So it's a complicated slide, [REDACTED]



the organisations, that throw the biggest budgets, or the most data at the problem. It is going to be those who waste the least time, and the least resources zigzagging, sort of like a pinball, between those two extremes, and instead can chart some kind of sensible, thoughtful pathway between them. Does that make sense? Okay, so that's what I want to help us to do.

I've got 40 minutes, that's lots of time. I mentioned briefly that I'm at Oxford, so I'm a fellow at the business school, and do a lot of work with senior executives on just making sense of the present; forget about the future, making sense of the present. I've started to publish a monthly thought piece, a monthly essay, to people that I've worked with in the past, and we've started to call these essays my maps, and everybody finds them really helpful, in part because I keep talking about these things that end up happening, like pandemics. So some of these people have come to me and they have started to talk to me as a futurist, and every time somebody tells me, "Chris, you're a futurist," I just sort of have to hang my head, and sigh, because they don't get it, that in a time like now, seeing the future is not the problem; our problem is, "Can we see the present clearly?" That's the hard part. I mean, can we answer questions like, "What aren't we seeing?" "What's wrong with how we're looking at the world?" "What's out of focus for us?" "Where are our blind spots?" and again, in a moment like now, it's worth contemplating that, you know, for leaders, and senior leaders, it just may be that the most important value that you bring in your leadership roles, is in helping some things not to happen, by asking some of those good questions, about "What aren't we seeing?"

So, that's what I want to help us to do, when it comes to AI, and what I want to do is try to share with you three truths, that we tend not to get, when we talk about AI, and those three truths are first, that AI is not substitute; it is a simplification. Second, that the model is always missing something; and third, that whatever we model becomes our reality. I'm going to unpack each of those in turn, but let's just start with the first, the idea that AI is not a substitute, it is a simplification. I know that this is a sophisticated room, so maybe not here, but outside the doors of this room, in, sort of, broad public discourse, AI is seen as a substitute for human intelligence. And if you want to understand that, a fun game you can play, you can actually do it for any word, and learn a lot about how society thinks, but if you do a Google image search for "Artificial intelligence," and just see what some of the top images are, that come up, you get a visual glimpse of how the broad public thinks about artificial intelligence, and what do we see in this picture? That artificial intelligence is beginning to develop the capacity to think like humans do, to reason, to make decisions, to exercise judgement, to bring intuition. We're not just talking about a tool here, we are talking about a thinking machine. Right?

That's the popular discourse outside these doors, and even if we don't share it, it can cloud and infect our thinking in this room, because the reality, the on the ground reality of AI is, it's far more boring, frankly. *[Laughter]* This is a snippet of code for a machine learning algorithm that I'm playing with right now, to do handwriting recognition. A lot of fun, and admittedly this is complicated; it's a deep learning algorithm, "Woo," and it's driving some pretty complicated math and maximisation functions, I grant you that. But if you look under the hood of anything that is called artificial intelligence in the world today, that's all it is. It is code, it is a database, it is driving some kind of math function. It's basically trying to take reality, strip it down to a matrix of numbers so that we can solve a math problem.

That's all that it is, and yet every time we manage to do something that seems humanlike with this process, we leap towards giving it humanlike qualities. Remember a few years ago, Google's AlphaGo beat the world's Go champion, anybody remember that news story? In popular media, on cue, all of these stories with things like, "Look at the human intuition that the machine is demonstrating." I don't know, you tell me, if I gave you these lines of code, and a big database of numbers, and I said, "Whoever comes back to me with an accurate solution is the winner," is that human intuition? Is that anything like our faculty of intuition? We love to talk about... we love to anthropomorphise AI. "It's playing Go," no. To play is an act of human intelligence. What this algorithm was doing was stripping that act down to a computational task, that machines could perform. It's a very different thing. So, AI is not the substitute, it is a simplification.

This is more than an academic distinction. I predict, at the risk of sounding like a futurist again, some of the big mistakes, some of the big problems in industry, in government... hopefully not in defence, are going to come from failing to see the difference. For example in workforce automation, it's going to lead to some really big mistakes. Some of the... now I'm kind of talking outside of school, but some of the leading work on understanding workforce automation came out of Oxford University, came out of a research unit that I used to be a part of, called the Oxford Martin School. There was this famous study, I guess now was about ten years ago, maybe not quite ten years ago. This was the study that said... it concluded that about 50 percent of current jobs are going to be automated away by 2050. Have any of you ever heard any number kind of like that? Any statement kind of like that? Raise your hand if you have. Wow. Okay, did you know it all came from this report? I kid you not. If you look at other studies, and you go into their references, they all refer back to this report, because they did such a comprehensive job at looking at this question. They took the entire universe of US jobs, and for each job, they made a determination, "How likely is it that this job is going to be automated away?" and then you rolled it all up into this universe, and gave a kind of predictive factor, "Here's the 50 percent number," like, more than half of all jobs, or, almost half of all jobs [were sort of? 00:14:36] 50 percent, 60

percent likelihood or more, when they ran that algorithm. These are the nine job characteristics that they identified, that helped them to figure out, “Is this particular job going to be automated away or not?” But don’t worry about that dry detail for the moment, we’ll get back to this slide.

What I want to share with you is what a couple of clever students did, which is, they created a fun website, as a front end for the conclusions of this report, called, “willrobotstakemyjob.com”. Has anybody ever been to this website? So, it is the same report but, kind of, you know, millennial-ised. I’m not a millennial. Okay, so, you type in your job, and it tells you how likely you are still to have it in 2050. Now, I’m a political scientist. I’m just saying. I’m totally safe. I suspect, because nobody understands what value that we bring. Right? Like, “Why would we invest in getting rid of Chris?” Different story for... you know, one at the top end of the spectrum, something like a procurement clerk. You're doomed. *[Laughter]* And, you know, frankly, just listening to the job title, it kind of... we feel like, “Yes, okay,” like, what is a procurement clerk? Actually, people in the room here could probably tell us. I suspect it has to do with understanding the inventory, and inventory levels, and how quickly they’re being drawn down, and when we need to order more supply... it does sound like a kind of... the sort of task that you could imagine a machine doing for you.

The big blind spot with examples like this, is that the task is not the same as a human performing the task. So, let me give you what I think is maybe one of the most brilliant examples in all of history of this. It happens to be about a procurement clerk in the late 19<sup>th</sup> century, this is about 1890s, in Liverpool. This is [Ed Morel? 00:16:50]. Ed was a procurement clerk, and he noticed something really... what to him was quite unusual in the data of the ships that he was loading. He noticed that there were all of these ships going to the Congo that were loaded with nothing but guns and bullets, and they were coming back with nothing but rubber. Ed Morel was the man who blew the whistle on what some historians now call the Congolese genocide, and he then led the international effort to stop it, by about 1907. The point is that when he blew the whistle, he was not doing anything that was part of the patterned, repetitive job that was his task; he was bringing something completely other, completely human, to that same problem. Does that make sense?

So, society-wide, you start to worry about when we’re looking at spreadsheets like this, at decisions like this, are we thinking about these broader human qualities, when a person is doing the task? So, AI is not a substitute, it is a simplification.

It’s interesting though, how pervasive the idea that AI is a substitute is in our thinking, and at the risk of losing friends, I have to admit that I’ve found evidence of this even within the Ministry of

Defence's own definition of artificial intelligence. I looked it up, "What do you define artificial intelligence as?" [REDACTED]

[REDACTED] So, "Artificial intelligence is the ability of machines to perform tasks normally requiring human intelligence." This definition makes me uneasy, because to me it sounds like we're not... it doesn't sound like there's a trade-off here, and I'm to tell you that there is always a trade-off. There is never a pure case of doing more with less, there's always going to be a trade-off, and the hard question is, does the trade-off matter? How much does it matter?

I would be far more comfortable... I know we won't, but I would be more comfortable if we didn't talk about artificial intelligence at all, if instead we talked about pseudo-intelligence, which would be something like stripping an act that requires human intelligence down to a computational task that machines can perform. Why I would prefer that is because then I feel like the risk of that blind spot is smaller, because this definition, instead of opening up a potential blind spot, it is begging the question, "Okay, so, what might we have stripped away?" and that is the first really valuable question that you can ask whenever you're evaluating an opportunity to bring AI into your business, is "What might we be stripping away here? Might it be something relevant?" Okay, so that's number one.

Truth number two: The model is always missing something. All right, this is going to be fun. So, the model is always something. The procurement clerk example, good example of that, right? Some value that was performed that the model, if you were to automate that job away, probably wouldn't account for. But I want us to have, maybe, an even example, and one that I'd like us to have some fun with, and create together. So, what I'd like us to do for this example is we're going to build our own machine learning algorithm, right now. Has anybody here never built a machine learning algorithm? *[Laughter]* Oh, okay, great, so you'll be able to go home today and say, "Hey, guess what I did?" we're going to build our own algorithm, and then use it as an example.

To build an algorithm we only need four things. First, we need an agent. Can I have a brave volunteer? Is somebody willing to be the agent in our algorithm? It would involve coming to the front, I admit. Excellent. [REDACTED] Everyone, our agent. *[Applause]* Thank you very much. I'm Chris, you are?

**KEV:**

I'm Kev.

**CHRIS KUTARNA:**

[REDACTED]

[REDACTED]

[REDACTED]

**CHRIS KUTARNA:**

Kev, excellent. All right, so, Kev is our agent. One. The second thing that we need, to build an algorithm, is an environment in which Kevin can [play in? 00:21:38]. So, let's say the front of the room is our environment. Okay? It's a warehouse, let's say. So, we've got this warehouse on both sides of the stage, and inside the warehouse we've got these cones. All right? These represent the things that we might want to move around the warehouse. Makes sense? Okay, Kev, could you just, sort of, four on each side, a few over there, a few over there, make the environment a bit messy, so that we're creating work for you to do, is what you see. Yes. Yes, so maybe four on that side of the stage, and then come over here, and then four on this side of the stage.

**KEV:**

Like that?

**CHRIS KUTARNA:**

Perfect.

[REDACTED]

[REDACTED]

**CHRIS KUTARNA:**

I think at Lego they call this, "Serious play." Excellent. Okay, so we have an agent, Kev, we've got our environment, this warehouse, the third thing we need is... you can come back, Kev, thanks... the third thing we need is a policy. So, what are the rules that Kev is going to follow? Rule number one is, "Your job is," now that you've helped me set it up, "You're going to bring all

of these cones back to me," okay? That's the job. Rule number two, "You can only pick up one at a time," only carry one at a time. All right? And then rule number three, and this is the most important, is, "We've got to have a human in the loop," okay? So, I'm the human in the loop, and if I say, "Stop," then my agent's going to stop whatever he's doing, and come back to the podium, until I tell him to go again. Okay? Clear?

**KEV:**

Yes.

**CHRIS KUTARNA:**

Does that make sense? Okay, so... oh, one more thing, Kev needs a reward function. Right? Why are you even doing this? Right? So, I happen to have a reward function, gummy bears, and... and this is the way it's going to work: each cone from that side of the room, that Kev brings to me, he gets five gummy bears. Each cone that Kev brings from that side of the room, he gets one gummy bear. All right. And your function is to get as many gummy bears as you can.

**KEV:**

Okay.

**CHRIS KUTARNA:**

All right? Clear? All right, so I'm going to press play in a moment, and we're just going to see what happens. *[Laughs]* Okay? Excellent. Kev, go. Stop. You've got to come back. Okay, you can go again.

**KEV:**

Okay.

**CHRIS KUTARNA:**

Stop. Go.

██████

[REDACTED]

**CHRIS KUTARNA:**

Excellent. Okay. You've proved the point, thank you very much, Kev.

[REDACTED]

[REDACTED]

**CHRIS KUTARNA:**

All right, great. Give Kev a round of applause. *[Applause]* All right, so this is our example now, about this point, that the model is always missing something. So, two things about our environment, that our agent didn't know. One, is that this is inside the warehouse, and this is outside the warehouse. He didn't know that. And two, it's snowing outside right now. All right? And I didn't build this agent to operate in the snow, so I didn't know what was going to happen, that's why I kept telling it to stop whenever it stepped outside. Right? That's what was happening in my reality.

What do we take away from this? Two things. First, it's possible for the agent to learn the wrong things. Kev learned two things that I never wanted him to learn. One was to avoid going outside, when in fact that's what's most useful to me, and two, Kev learned to avoid the human in the loop, and you can imagine in certain, especially defence situations, that would be a really bad thing for your autonomous agent to learn. Right? So, badly programmed algorithms, no offence, Kev, they can learn bad behaviours really easily. Does that make sense? Okay.

But the second, and, this is, kind of, the broader takeaway, is that the model is always missing something. All right, Kev is not operating in "The world," he is operating in an environment that we have specified, and that is always going to be something less than "The world," that we are living in. Right? And for a couple of reasons. One, I mean, we can't possibly specify everything; the best we try to do is to specify what is relevant, and that can be hard to figure out, but then, two, there are some stuff that may be relevant that isn't... it's not even quantified; there is no data for it. So, therefore, we cannot make it part of Kev's function, to take it into consideration.

Now, in our pretty simple example, there are definitely things we could have done. We could have said, "Oh, yes, weather is relevant, so, you know, let's stick a weather sensor on Kev's head," if he doesn't mind, "and add a bit of code, that says 'If snow, then don't go outside,'" but in the real world, this becomes really difficult, trying to specify the model, so that it consider everything that is relevant.

Remember this? So, because I'm, kind of, concerned about how ubiquitous a single narrative about automation has come out a single report, I continue to nit-pick on this report. The authors are not big fans of me doing this, but I think it's good for the world, so I keep doing it. One of the things that I did is, I went through every single prediction, job by job, and I circled the ones that said, "This makes no sense whatsoever." My favourite example: fashion models. 98 percent chance of being automated away. Did you know, in ten years, there will never... there won't be any more runways, right? I happen to have a friend who is an executive in the fashion industry, so I call her up, and I said, "How much are you spending on automating away your runway models?" She laughed, she said, "We're not spending anything on that," I said, "Then I have a report that I need to send you, because you're missing a big opportunity."

Obviously, what's really going on is that the algorithm is missing something about reality. Okay? So, what is it missing? Well, now we go back to this boring slide. These are the nine job characteristics that, according to the algorithm, help it decide, "How hard is it to automate this job?" right? Things like manual dexterity, "Does it require fine art skills?", social perceptiveness, it, kind of, all makes sense, until you really go through all the list, and you say, "Well, you know, one of the things that isn't on the list, is the ability to strike a pose like that." And yet, in the reality that that job operates in, that is a relevant consideration, right? So, this is the second question, that I think you can always ask to find some value, some perspective, some judgement to bring, to the choice to bring AI into anything, which is, "Okay, but what is missing from the model?" and if the person who made the model says, "Nothing," do not hire them. Right? Something is missing. My job is to figure out, "Might it be something relevant?" and it's very difficult to do.

I went back to the friend who gave me this slide, and I said, "Okay, ~~Silva~~, you've worked on hundreds of these cases, big complex AI problems: how hard is it, you know, to get to the model that fully specifies what's relevant? How long does it take?" and he laughed, and he said, "Like, that work is never done," it's never done, but what he has, over hundreds of cases, identified, is a, kind of, formula to help him set his own expectations: 10-20-70. Ten percent of time, and the effort, and the value, comes from the algorithm; 20 percent of the time, and the effort, and the value, comes from the data that feeds into the algorithm; 70 percent of the time and the effort, and the value, comes from working with the experts, to figure out, "What are the relevant pieces

of reality that are not in the model yet?" and that work is never done. And I think that you'll find that that's true for you as well.

All right, number three, and now I'm getting a bit philosophical, but I wanted to talk about this third, deeper truth, that I think if we can grasp it, was really going to help us to avoid some of the blind spots, as we rapidly adopt AI over the next 20 or 30 years, and it is that what we model, actually becomes our reality. I'm talking about bias. And I'm not talking about the kind of obvious bias that... we've all heard the stories, we've consumed them in the popular media, there is a story in the Guardian a few years ago, that if you Googled, "Unprofessional hair," these were the results you got. Why? Because people are racist. Right? The algorithm doesn't know that, you've just searched for a term, and it's given you the most popular results associated with that term, and over the last few years, Google has worked really, really hard at suppressing the obvious racism, and, sort of, other unwanted results from its search algorithm. It's had some success; three years ago, if you started to google search, and you entered, "Are women..." what was the first auto-suggest result, to complete your search? *[Laughter]* Where did that come from? There's a lot of misogyny in the world. Right? It's just a statistical correlation. Jump forward to today, I tried it yesterday, you get, you know, a far more... it's still a strange... but, you get a different list of auto-suggests. You also get at the bottom, this link to, "Report inappropriate predictions," right?

But that's not the bias I'm talking about. That's not the bias I'm worried about. I'm actually more deeply worried about a more fundamental bias in the nature of algorithms, which is that they are maximisation functions; they optimise for something. They're all trying to get the gummy bears. A definition of gummy bears. And you can see... it's hard to think about... with... you can see the consequences of that deep bias, when you look at the attempts to remove the obvious bias. So, really fascinating.

If you go into Google search today, and you search for, "Unprofessional hair," this is what comes up, and it's not that much different than it was a few years ago. Now, it's not because of the racist associations; those links have been suppressed by google, no, now it's because of that Guardian news story, that created so much conversation, and popularity about the problem, right? Because this is what Google's algorithm does, it's looking for the popular result, and so it gives that result, and then it leaves the algorithm, and it enters our real world, where it becomes popular for different reasons, and then it can't get unstuck. So, we see it with social media all the time, the algorithm is trying to maximise my time and attention to the app, content that does not help that objective has no value, and therefore I see less and less of it, it becomes less and less a part of my reality. What we model becomes our reality.

Now, why is this important? Well, maybe it helps with an example. Say, this is a bit of a naïve example, but let's say you wanted to build a recruitment AI, to help recruit the perfect soldier. All right? You're going to recruit the perfect soldier, and you're going to use an algorithm to help you, and the naïve thinking is that, "Well, if the human recruiter picks someone here on the line, the algorithm is going to pick someone over here," right? Closer to the optimal. And what happens, the intention is that the more the algorithm runs, the more it learns, the better and better it gets, the closer it gets us to that optimal recruitment decision. Then what happens within the reality of the organisation, is that, you know, the more people that get hired by the algorithm, the more people you have inside the organisation who think that the algorithm is making great recruitment choices. Right? More and more people just like me. The problem with that is, this, obviously, it's this big word, "Resiliency," that I heard all morning, right? Who were the great optimisers in history? The dinosaurs. Right? And so that is the third question, that we really need to learn how to ask ourselves as we're bringing artificial intelligence into our systems, which is, "What biases are we baking in when we do that?" Because we are baking in biases.

The flipside is, if we can see that truth, and ask that question, that puts us in this brilliant position, to ask a really interesting question, which is, "What patterns could the AI help us to break?" and you can start to imagine using AI not just pursuing one objective, which is this optimal that doesn't exist, but, kind of, thinking about two objectives at once; both selecting for the optimal, and mixing things up. Which, again, in our HR example, in a rapidly changing environment, you need to mix things up, because, you know, if you're only selecting for one trait, and then there's some environment change, and you know, suddenly, people with geeky computer programming skills become really valuable, but you've bred them out of the population, you're in trouble. But if, instead, what your algorithm is doing is identifying your patterns, and saying, "Hey, you might want to mix this pattern up, maybe you don't recognise that this is a pattern in your behaviour," then that can add a useful element of mutation to your population.

Now, too much mixing up, you know, that's cancer, that's not valuable, but we know that there's got to be some kind of healthy balance between these two forces; a balance between recognising the pattern I'm trying to follow and reinforcing it, so I can do it better, and recognising the pattern I'm trying to follow and tell me, "Hey, time to mix it up a bit," and we know that that Goldilocks zone exists, because that's where mammals came from, right? That's where we came from, that's where evolution happens. And this, I think, is really the big, societal potential of AI, is if we can use these technologies to help us spend more time in this zone, between pattern fixing and pattern breaking, that is where innovation speeds up, that is where learning speeds up, that is

where we develop the resilience as a society to deal better with the, kind of, Black Swan events that we have going on right now all around us.

Kind of heavy for the first session after lunch, eh? But, wrap things up. So, I've tried to offer three truths to help us see the AI present clearly: it's not a substitute, it's a simplification; the model's always missing something; and whatever we model reflects back into our reality, and if we're not careful it becomes our reality. I've also offered you three questions that you can use to protect, to guard against that blind spot: "What are we simplifying away here, in this automation opportunity that's been put on my desk?", "What is missing from this model?", "What biases are we baking in, what patterns might we break?"

Taking a step back to the big perspective for moment. I mean AI is here, it's coming, it has enormous potential as a new general purpose technology in society, I mean, enormous potential. There is a path to a brighter and more beautiful world, and AI may be one of the important vehicles that gets us there. The question is, how zig-zagging is that path going to be? Right? I mean, are we going to get lost in the hype, are we going to get derailed by the hysteria, and the answer to that question is a function of your human intelligence. In how we think about, and how thoughtfully we adopt these technologies, and so the irony is that AI is not here to replace any of us, it is here to demonstrate to all of us how indispensable human thought and human reflection is in a moment like now, where everyone outside the doors to this room are just racing to jump to the conclusion that, "Computed equals better." Right? That is the biggest blind spot, and those of us who see that blind spot, you are the ones who are going to benefit the most from artificial intelligence, and achieve AI superiority. I hope it's us. Thanks.

**CAPTAIN (RN) DAMIAN EXWORTHY[?]:**

Chris, thank you so much.

**CHRIS KUTARNA:**

Thanks. Kev, you forgot your gummy bears.

**CAPTAIN (RN) DAMIAN EXWORTHY[?]:**

Chris, thank you very much indeed, that was outstanding.

*[Recording ends]*