Mike Harre – 'Can Artificial Intelligence Really Be Smarter Than You?'

Moderator: Welcome to the podcast series of *Raising the Bar Sydney*. Raising the Bar in 2016 saw 20 University of Sydney academics take their research out of the lecture theatre and into 20 bars across Sydney, all on one night. In this podcast, you will hear Mike Harre's talk, *Can Artificial Intelligence Really Be Smarter Than You*? Enjoy the talk.

Mike Harre: Good evening everyone. Thank you very, very much for coming out tonight. This was a really great opportunity for us, as academics, to come out and talk to people about what we do for research, because we don't often get an opportunity to take what we do in our labs and let people see what makes us enthusiastic about what we do in our research and in our daily lives, and why it might be important to the way in which we evolve as a society over the next 20 or 30 years.

So in the introduction, I was introduced as someone who works in the Complex Systems Research Group at Sydney University. So my job is to combine artificial intelligence with economic theory, and how people make decisions. How real people make decisions using real psychologies, and can we find ways to put that into a computer. This is a really interesting problem to try and solve because as far as we know, there is nothing as complicated as the way in which our processes interact with other people to make decisions in the way that they do. So with that in mind, I want to just talk very briefly about what artificial intelligence means in very general terms, and then I'm going to talk about some of the interesting challenges that we're confronted with, and how that's giving rise to some really challenging questions about who we are and what we do, and how we're going to live the next several decades.

Artificial intelligence, very, very generally, what do we mean when we talk about this? Generally speaking, when we talk about artificial intelligence, we're talking about a piece of code, a piece or algorithm which runs on a computer, and it tries to emulate some aspect of the way in which we think, the way in which we make decisions, the way in which we generate our behaviours. When we think about things in those terms, it's only a very, very recent thing when we realised there might be a way to actually simulate that.

It was only really at the beginning of the 20th Century where we had some very interesting and exciting ideas coming out of mathematicians. So a guy called Church and a guy called Turing were interested in the way in which we can mathematically describe how things make decisions, how things process information. They came up with this idea that there is nothing in a piece of algorithm, a little piece of mathematical process that couldn't simulate the way in which our brains work. It hasn't been proven. It's important we distinguish between what we can prove and what we can demonstrate. We can demonstrate this is true, but we haven't proved it. This was a Church-Turing thesis. It says we can take our brains and we can simulate them in computers. We take that for granted. As artificial intelligence researchers, we take that for granted. If we take that as our starting point, what are the sorts of things we can do with those artificial intelligences running on our desktops or our laptops? What sort of things can we do?

There are two broad categories that we have in AI. We have, broadly speaking, what are called expert systems, and we have learning or adaptive systems. Those are two very broad categories. There are dozens of ways in which we might actually do this in practice, but those are the two broad, very high label categories.

What's the difference between these? In an expert system, we have literally a human. We find ourselves a human who is very good at something that we do as humans, and we set them down and we get a computer programmer, and we get the human to describe to the computer programmer how they take information and turn that into a decision. They've got to describe their thought process, their decision process, to the computer programmer. The computer programmer then turns that into a piece of code, and that piece of code then runs on a computer and in principle, the human expert and the computer can get the same input, go through the same series of logical operations, and churn out the same answer. That's the theory. Now it turns out that we know that we really do not know what we know. Now, how do you understand that? Have you ever really tried to explain something to someone who refuses to understand?

If you take an expert, we're all experts in lots of things, right? Just naturally we're experts because we do things all the time. Imagine what it was like to try and explain all the details about opening a door. Someone says, “Could you please open the door”? Okay, how would you go about breaking that down into things that you could then describe to a computer programmer who would then tell a robot, for example, a completely ignorant robot, how to stand up, go to a door, and open a door. Struth it turns out to be incredibly hard, right? You have to first of all take into account, all the different ways in which you might be asked to open a door.

What sort of ways could you be asked to open a door? You'd be asked to open a door when you're sitting in a car. You could be asked to open a door when you're sitting in a room like this. You could be asked to open a door when you're in a jail. You could be asked to open a door when there's no door in the room. The problem is that all the different circumstances under which you might be asked that same question would have to be listed by the expert, and then all of those particular instances would need to be turned into code, into a whole bunch of ways in which you actually stand up, you walk, or whatever you have to do, you recognise a door, door handle, turn the handle, push or pull, and that is actually very, very hard to do. We found out very early on that that's almost impossible, except in very, very narrow circumstances.

For a long time, that was really all we could do, and then we started working on things which are adaptive, things that learn through examples. This is the second category. This is adaptive learning systems and in this case, we're not trying to simulate the actual high-level process, the logical process that an expert can describe. Now we're trying to simulate the underlying structure of the human brain that gives rise to that sort of logic. Now we really want to understand how neurons change and adapt over time, and result in that sort of logic that an expert has.

What do these do? These take a whole bunch of examples, thousands and thousands of examples. They go through an artificial neural network, which is a bunch of nodes and links between nodes in a neural network that's encoded in a computer, and it gets some output.

We tell it, for example, we've got a whole bunch of pictures of dogs and cats, and if I give you an example of one of those, whether it's a dog or a cat. Turns out that this is also quite hard, but we can do it, and we can categorise all sorts of images.

This is one of the greater advantages that we're getting from this. We don't need to have an expert. We can actually force it to learn how to do this. I thought it would be funny to take an AI that is available readily on the Internet, or rather a piece of software that I've got, run by Wolfram. They've got an artificial intelligence which has already categorised a whole bunch of images, and I thought I'd throw a couple of images that I had in my photo library at this. I took some faces and some sceneries, and things like that. Then I gave it a picture of my pushbike. I gave her the picture of my pushbike, and it was supposed to categorize it into whether it's a person, or a particular piece of machinery, or whatever it is. I gave it a picture of a pushbike, and it came back and told me it was a crossbow.

It's a crossbow? It had mistaken it for a cross bow, but if you looked at the photo, it was a front-on shot of my bike, and the handle bars were like that. You can see the top bar going to the back, and there's a seat at the back. You could've imagined the seat might be the shoulder pad, and the handle bars might be the cross member from the crossbow, and maybe that's why it saw it as a crossbow.

A couple of days later I went back and did the same thing. I put the image in again, and I got the same answer. Now people don't behave like that. People don't learn like that. We learn much, much faster. In effect, we can learn from zero examples. How do we learn from zero examples? If I told you that having a car accident is a bad thing, you don't need to have a car accident to know a car accident is a bad thing. Some of us might have had multiple car accidents. We don't learn very fast sometimes.

But for a car to learn, or an AI that goes into a car to learn, it would need to go through hundreds of examples of accidents in all sorts of different circumstances. Does it hit a tree, does it hit a bridge, does it go through snow into a snow bank. All those different variations, it would have to experience those different variations, and thousands of them before it can actually learn. This is really not how we learn. We learn very, very quickly, one instance in some cases. If you put your hand on a hot plate, you know never to do that again afterwards. This is the thing. AI really doesn't do very much of what we do, except in very, very narrow circumstances.

Some concrete examples, some really interesting concrete examples. 1997 we have a world champion in chess in Garry Kasparov, and we had IBM, who had decided it was going to take on one of the great challenges of artificial intelligence, and it was how to beat a standing world champion in chess. A very humanistic thing to understand how chess works. IBM spent a few years developing this combined hardware/software platform, and it does some incredibly powerful computations, and what did they give this computer as a way to start thinking about chess? It gave them a whole bunch of opening positions, a whole bunch of end positions, and then, because that's not enough, you have to have a way of evaluating whether a good position for black or white is on the board.

How do you do that? The middle game is really complicated. You need to be able to evaluate whether you're in a good place or a bad place.

In chess, humans do that in a very, very subtle way, and we don't really know how we do that. They had to come up with an approximation to that, and they did it by getting a committee of experts together, and constructing an equation, which took the state of board and estimated a version of who's ahead in the game. What did that mean? That meant that you could take a search engine that literally goes five, ten moves deep, millions and millions of these examples, and evaluate them, over and over and over again, many, many times a second. And then you choose the best one out of those, and you'd start moving down that path. Okay. That's a system that didn't adapt though. Once you've finished coding the software, that was it. It would just keep on doing the same thing over and over again. It didn't learn.

Now this is quite different from the way in which humans behave. Humans walk up to, particularly Kasparov, or any really good expert, walks up to the board, instantaneously recognises about five moves to make, and that's it. Often, those five moves are enough to be a really good strong move. That's all you need. That early perception of the board is not even conscious. It's literally less than a second or two. How does that happen? Obviously much, much practice, and all those sorts of things. All that comes into play. Then over the top of that, maybe you've got five or even ten moves that you're going to explore, but you don't even branch off very much from those options. This conscious bit then takes over. Those five moves that I unconsciously thought were good ones, I'm now going to explore those in a very narrow way. That's how humans think about, that's conscious processing overlaying unconscious processing. That's where we get an expert system. That's where we are with expert systems.

Roll on forward to early this year, 2016; 2016 we've got Alford Go. Now this is a vastly different cup of tea. What have with got with Alford Go? Alford Go is a learning adaptive system. It's a neural network. What they did with Alford Go is they wanted to play this game called Go. Go is an ancient game, two-and-a-half thousand years or more it's been around. It's been around for a long time.

Now with Go, just to give you a scale of the difference between chess and Go, in chess you can search all the branches of possible moves, not even all of them, sorry, you can search a very large number. Now if you took all the possible states in chess, and for every single one of those states in chess, you had a complete different set of states just as large as the original state, that's the sort of multitude. It's even worse than that, but it gives you an idea about just how much more complicated Go is. We couldn't do the searching algorithm at all, it was never going to work.

That's why they gave it a neural network. What did they do with a neural network? They gave it a whole bunch of examples. They gave it hundreds of thousands of examples that it'll learn from. With hundreds of thousands of examples, what they found is that it would become a good player, but it wouldn't become a strong player. In order to make it a stronger player, they said, “Okay, we're going to take the original one that's just learned how to play using examples, and we're going to copy it, and we're going to get it to play against itself”. It played against itself for a long time. Once it played against itself for a long time, then suddenly it started to get really, really strong. It got strong enough to beat the then, still currently, world champion in the game of Go. So now we've beaten two of the very most difficult games we have available to us using two different forms of artificial intelligence.

What have they not told us though? With these experiments and with these test cases, what have they not told us? They haven't told us anything about the way in which we do it. They don't tell us about the way in which we think about these problems. I've talked about two of those things that we do, so we talk about patent recognition, or early perception. We know that from a vast array of psychological studies that have been done by a bunch of very talented people since about the 60's or so, about the way in which we recognise large scale patents very, very quickly.

Then we have, over the top of that, this sort of strategic analysis. How you look deep into the sequence of moves you want to make? But you also need something else. You also need to have some understanding about the other player. This is one of the really important things. A good chess player is able to have a very narrow selection of moves because the other player has a very narrow selection of moves. If there were more moves available and the other player explored all those other different moves and you didn't, you would struggle. But because we're similarly cognitively arranged, we're playing against humans; we're able to explore the same number. We're sort of comparative in that sense. There's this interesting play between how we understand our strategies in terms of other people's strategies, and what is the best response? It turns out that that sort of thinking is identical to the way in which we play game theory in economics.

Game theory is all about how I make a decision based on how I think you're going to behave in response to my decision. In order to think that way, as clearly as possible, I need to have some understanding about you. Economics gives us an insight into the sorts of problems that we try and solve. This is probably the strongest thing economics has ever done. It's been around for about 200 years, this notion, in various forms. For all the problems that economics has at the moment, this is probably the greatest insight economics ever had into the way in which we think. Economics realised very early on that our decisions are influenced by what we think other people are going to decide and that's quite cool. That didn't turn up in other fields. It didn't turn up in psychology until very, very late in the piece. It didn't turn up in psychology until early to mid 20th Century.

We now know from psychological studies that this is part of our early development processes. We get to understand other people. We get to build models internally of other people, and we get to exercise that model, exercise that as part of our brains that use that very, very early on. If we don't do it, you can actually see people who don't understand the internal workings of other people. If it gets to a very, very extreme case, we see it in autistic spectrum disorders, where a misunderstanding of the internal states of someone else's thinking hasn't developed as well as it might have. This is really important. This gives us a very interesting psychological insight into the difference between AIs, and the way in which we psychologically behave.

That's narrow AI. That's AI that works in very, very focused cases. We can do that. In fact, we can do it so well now that we can beat humans, which is pretty astonishing in its own right. The amount of computational power we have to throw at these problems and how good we are at that. But you can't take Alford Go and put it into a car. Cars are one of the really interesting things we're playing with at the moment. If for no other reason than cars confront us with some really interesting ethical dilemmas.

One of the great things about artificial intelligence is it raises questions about who we are and how we want to make decisions. How does that work? We've got autonomous cars. What sort of questions would be asked ethically? There's a really famous example from philosophy. It's called the trolley car problem. The idea is that you have a train, and it's out of control, and it's coming down a slope. There is five people at the bottom of that slope, and they're going to have this train run into them. You are standing off to the side with a little side track coming off and your little switch, and you can choose to throw that switch, and instead of hitting those five people, it's going to swerve aside and it's going to hit one person.

If you do nothing, you might feel like you don't have any moral responsibility because you don't do anything, and you let the trolley, let the train hit those five people. If, on the other hand, you say, I'm going to throw the switch, I'm going to actively participate in this, but I have to actively choose I'm going to hurt one person. If I actively choose to hurt one person, most people will say, for the greater good, who do I hurt less? Most people say they would choose to go down the path of hurting one person. Now we can play with this idea a little bit, and we can say, well, maybe it's not a stranger on the track. Maybe instead of me being by the switch, maybe it's me on the track, and then if I pull the switch, it's not going to run into those five people, it'll run into me. So who would voluntarily pull the switch to hurt themselves, given that choice?

What does this mean for artificial intelligence? When we put AIs in cars, we have the option in some sort of philosophical sense, of having AIs that choose to potentially run into a crowd of people in a pedestrian crossing, or swerve into oncoming traffic. You swerve into oncoming traffic; you've got a chance of hurting yourself. If you run into five people on a pedestrian crossing, there's a good chance you'll be fine, but those five people are going to be hurt.

We end up with a very similar philosophical question about what we want our AIs to do. They asked people what they would like people to do, and we have the answer. People, when they're making their decisions by the cold light of day, potentially in a car yard, what would they rather have an AI do in a car? They would rather have an AI choose to injure the fewer number of people, which means, in practice, you'd have a car that would run into a tree or into oncoming traffic, rather than hurt someone else, or a group of people. The trouble is, if they say that, it's also true that they say that they're not going to buy a car with an AI in it, because no one wants to be the guy that gets hurdled into a tree because his AI would rather save these five people than him.

We're asking some really interesting questions about the way in which we frame these. What used to be abstract philosophical questions about trolley cars turn out to be really practical questions about the way in which we make decisions about the sorts of AIs that want in our everyday lives.

There's more pertinent and far more interesting problems that we can also ask about. What we want to do if we go to war, for example. Let's imagine that a country that you belong to declares war. Your country has the capability of developing an artificial intelligence that can be put into an armed robot.

It knows that it can do that, it knows that armed robots can go off and wage war on your country's behalf, while you stay at home safe. All right. The alternative to arming artificial intelligences is for you to send either yourself or your loved ones to war. Which would you rather do? It's a challenging question because we don't really want to arm AIs. This is a very common debate around what we would do with them. If we got them, would you rather kill the AI, or destroy the AI, or destroy people?

These are really interesting questions. We're not really looking for an answer to these questions, but they confront us with very serious concerns that we want to reflect on. What do we want in our societies? In thinking about these sorts of questions, we naturally lead ourselves to the idea of what are the risks that we want to be able to address? Instead of talking about the negatives, let's talk about the positives. There's a World Risk Report that comes out every year. It's in its 11th edition, and the 2016 one makes for some relatively sobering reading.

In the 2016 one, it has a list of all of those threats that we face, and it measures those threats according to likelihood and impact. You might know, just off the top of your own heads, what a bunch of those might be. Large scale forced migration of people, financial crises, the inability to address global warming issues. We are currently in the sixth mass extinction of life on earth. These are real things that are going on right now. These are real risks we're confronted with. Curiously, AI's not on that list. There is no one out there saying that the most immediate mid-term threat we're confronting with is actually AI taking over the world. But we do have very, very real risks that we're confronted with, and AI can help us.

How can AI help us? I want to illustrate this with one of the main problems that we get with these risks. It's being emphasized now in this Global Risk Report. These risks are inter-related. They're coupled together in a way that they haven't been coupled together previously. How do we illustrate that as an idea? Let's take finance and economics. Finance and economics is studied by itself by economists and financial economists, and the idea there is you want to redistribute capital around the economy to invest in projects that produce all sorts of goods and services within the economy. That's a complicated and difficult thing to try and understand. We're trying to reinvent it at the moment, but it's a complicated thing to understand.

We also have the ecological environment. The minerals and natural resources we have in the world. We have coal, we have oil, and we have rain forests and natural plantations, and those sorts of things. We know that those ecological concerns that we have are coupled to the way in which we have biological resilience in the world, and we have people studying that, in and of itself. Again, that's an incredibly complicated thing to try and understand. Then we have the climate, and the climate is fluid dynamics. There's the fluid dynamics of air flowing across the globe, and it's also the fluid dynamics of the water transporting heat energy across the globe. Of course, the problem is that these three things that have historically been studied by groups of scientists independently, are now tightly coupled together. Our economics is influencing the way in which we extract minerals and resources from the ground and from the forests, and that is having an impact on the fluid dynamics of the climate.

While those three groups of scientists have talked in the past, nowhere near as much talking has gone on as might have happened, as we might have hoped for at this point. How can we try to hope to understand an even more complex system when we struggle with those as independent systems? AI gives us a way to do this. AI gives us a way, because in its richest and successful form, AI is able to crunch staggering amounts of data, enormous amount of data, looking for correlations, looking for patents, looking for emergent behaviour. From that, we can start to understand the way in which these risks are genuinely being influenced by individual decisions that are being made by corporations and companies, and the way in which we behave and interact with our environment. AI can help us with that. It can help us build these models. It's an incredibly rich and powerful tool and technique we can use.

I just want to finish by talking about why the idea of keeping AI open is the best thing for us. There is genuine concern, the worst case scenario for AI, the domination of the world, the robot overlord that keeps us under its jack boot, I think there's a low risk of that. That's my punt, okay? But there is definitely risks with AI, and we shouldn't shy away from that. But the importance that comes from AI is that if we keep it open. If we keep it a public domain idea, and we fund it with public money, we get research involved, we get government involved, we build centres to understand this.

We build white hat centres. Those of you who know what a white hat is, it's those people out there that using tools and techniques from IT, attack systems to understand how well they work in a positive way. White hat versus black hat. White hat is an important way in which we can defend ourselves against any risks that arise. Keep it rich, keep it open, keep it explorative. Keep it vibrant, and if we do so, we have the best possible outcome that if there is ever a threat, we will know enough about the problems that we're confronted with, that we might already have the solutions, the tools in our products that we need. That's what I'd like to see. That's my perspective. I'm a positive AI supporter with some caution caveats. Thank you very much for listening.

Moderator: Thank for listening to the podcast series of *Raising the Bar Sydney*. If you want to hear more Raising the Bar talks, head to raisingthebarsydney.com.au.

End of Recording.