Adrian Weller Elson Ethics Lecture 26.09.2018 OCT 2018 STA - DEPTH - 70 MINS
[MALE RESPONDENT]
[Other comments:]
McGowan Transcriptions - Digital Transcription Specialists

Ladies and gentlemen, as Dean of Windsor, I have great pleasure in welcoming you to St. George's Chapel, for this year's Elson Ethics Lecture. It is encouraging to see here tonight so many friends of St. George's House and the College of St. George. I'm particularly pleased to have the privilege to welcome Ambassador Edward Elson, and members of his family who are with us this evening. It is as most of you will know, Ambassador Elson's generosity that enables us to host this event annually, and we are enormously grateful to him. It is, you might say, a point of principle that the Elson Lecture each year should grapple with a topic of national or international significance. This evening's subject is truly global in its implications. To a lesser or greater degree, we're all aware of artificial intelligence, whether through the advocacy of those who believe that AI will transform life for the better on this earth and beyond, or through the premonitions of those who claim to foresee the end of human civilisation as we know it. There are of course many shades of opinion between the two. Whatever our view, there's no doubting that artificial intelligence is already changing society, and will surely continue to do so in ways that as yet we can hardly imagine.

Tonight's speaker is Doctor Adrian Weller of the Alan Turing Institute, the national institute for data science and artificial intelligence. Who better could we have found to address us on this subject? Doctor Weller is also a Senior Research Fellow in Machine Learning at the University of Cambridge, and at the Leverhulme Centre for the Future of Intelligence, where he leads a project on trust and transparency. Previously, Adrian Weller held senior roles in finance. He received a PhD in computer science from Columbia University, and an undergraduate degree in mathematics from Cambridge University, where he was a student at Trinity College. This is not his first visit to St. George's House. Recently he led a consultation on faith and artificial intelligence, introducing us to a hugely interesting and challenging area, and one which we're sure will be included in the future programme of St. George's House. Now, following this evening's lecture, Martin Stanford will conduct a short interview, giving you an opportunity to ask any questions you might have of the speaker. We shall then adjourn to the Dean's Cloister for a drinks reception. For now, as he takes for his title Trust, Transparency and Artificial Intelligence, it's my genuine and great pleasure to invite Doctor Adrian Weller to deliver the 2018 Elson Ethics Lecture.

[Applause]

Many thanks for that kind introduction. Mr. Dean, Ambassador Elson, ladies and gentlemen, it's a great honour and a great pleasure to be giving the Elson Ethics Lecture this evening. I'm very grateful to Gary McKeone and St. George's House for inviting me, and I recognise that as an academic who specialises in machine learning and artificial intelligence, that is AI, I imagine that some of you may be wondering why

I'm here to talk about ethics. As the Dean kindly said, I believe it's because St. George's House recently graciously hosted the first of what we hope will soon become a series of consultations on faith and AI, to discuss related ethical issues. When we consider the ethics of new technologies, we have much to learn from the deep wisdom of our faith traditions, and we must listen to their communities and the broad public. We should aim to move beyond partisan politics and dangerous tribalism, to learn to dwell together and to try to ensure a bright future of hope for us all.

AI does present great hope for society, with tremendous opportunities for us all to prosper and live better, healthier lives, but tonight I will also highlight important concerns. I'll argue that as AI systems are deployed across society, we must take measures to be sure that we can trust them. When we trust something, we make ourselves vulnerable. We need to be sure that that trust won't be abused to take advantage of us. Following Baroness Onora O'Neill, I suggest that we need reliable measures of justified trustworthiness. Onora spoke about three characteristics we should require of people. Those are competence, honesty, and reliability. Those of course are key, but I'm going to speak about AI systems. We should remember that AI tools are never ultimately responsible for anything. There's always a person behind them who puts them into production, and that person must be held accountable if something goes wrong. That's a very important point. I also note that there are significant concerns about technological disruption to jobs and the nature of work, but for this talk I'm going to focus mostly on AI systems themselves, which present particular opportunities and challenges.

I'm going to argue that there are some key measures of trustworthiness of AI systems which we must require. I'll speak to several of these measures and my claim is that if we can provide these measures, then well-founded trust can emerge. I'll start by trying to demystify AI. Broadly speaking, AI is about making machines do things which appear intelligent. Don't worry, I'm going to avoid getting into precise definitions where actually there is no clear consensus anyway. Instead, I'm going to tell you some things that AI can do, what it can't do, and I'll also tell you what I think we must require that it should do to enable well-founded trust in the systems we develop.

It's an exciting time for AI. We've seen great improvements in the capabilities of AI systems, and that's led to deployment around us in our everyday lives. I'm sure that many of you are familiar with these things here. There's an Amazon Alexa device on the left. On the right there's Siri, as many of us may use on our phones. These systems seem to listen to us and appear to understand what we're saying. They even sometimes can give us somewhat witty responses.

Actually, these systems don't really operate in a way that's similar to human intelligence. They instead operate in a much more brittle fashion, which I'll discuss soon. While indeed their performance is remarkable, it's also remarkable that we as consumers are happy to buy these products and place them in our homes to listen to everything we say. This of course raises important questions about privacy and trust, which I'll talk more about later.

Recommender systems. These learn about us, and then can sometimes make helpful recommendations for books, for grocery products, or movies, or even news stories. Increasingly, we see the world filtered through a digital lens, controlled by companies. Companies which are motivated to show us things we'll click on or keep watching. They appeal to our short-term desires, rather than our long-term best interests, or those of society. I'll call this concern one of influence, as algorithms target us and nudge us in different directions.

In computer vision, particularly for tasks of objects and face recognition, we're already at the point where in ideal situations, AI systems are sometimes able to rival humans for the ability to recognise objects or faces, and they can even do a reasonable job at image-captioning.

Please look at the image at the top left, if you can make that out. That image has been automatically labelled by an algorithm, and the caption that it provides is, 'Little girl is eating piece of cake.' That sounds great. It's a lovely picture, and that's a perfectly good caption. If you move over two places to the right, staying in the top row, there's a picture which has the caption, 'Woman is holding bunch of bananas.' All good so far. But now take a look on the bottom row. Start at the left. I'm going to point at how these systems can sometimes start to go wrong, and it reveals some of the brittleness of the systems. On the left, we've got a very cute picture of a baby, and the baby is holding a toothbrush in front of him. The caption says that a young boy is holding a baseball bat. Obviously that's not quite right. What I suspect is going on here is that this algorithm was trained on many thousands of labelled examples, that is, pictures where a human provided a caption. In the set of data that the algorithm was trained on, it happened that there were quite often cases where a baseball bat was shown, but it didn't see many cases of a toothbrush. This was the closest thing it found to come up with a good caption. If we move over to the bottom right, there's a much more worrying example. Here you see a bit of a ghastly image of a plane which almost crashed into a main road. The caption, however, simply says, 'An airplane is parked on the tarmac at an airport.' This shows that these algorithms really have no notion of common-sense reasoning, and they can fail in worrying ways.

Many of the breakthroughs and applications that we've seen are based on something called deep learning. Many of you have probably heard that phrase. Deep learning systems are neural network models similar to algorithms which were developed decades ago, but there have been a few changes. There have been some architectural and algorithmic innovations which have improved those systems. There are two other themes here that I want to emphasise, that have led to much greater capabilities. One of those is vastly larger data sets. We just have much more data than we used to have, and we also have much more computational resource than we used to have. Just those two things have led to enormous improvements in the capabilities of these algorithms. That's important to note, because looking forwards those trends are likely to continue. Even if we don't manage to improve the algorithms, and I suggest that we are likely to because there's enormous investment in doing that, but even if we didn't, the trends of more and more data, and more and more compute, are likely to continue, leading to improvements in at least these narrow tasks of perception which these sorts of algorithms are good at.

Deep learning does have limitations. The algorithms are very data-hungry, that is, they often require millions of labelled examples. That is, labelled by a human. They're very computationally intensive. They can be poor at representing uncertainty. They're easily fooled by adversarial examples. That's an important aspect where they're not as robust as we'd like, and I'll show you more about that on the next slide. They can be tricky to optimise. Also importantly, it can be difficult to understand exactly how they manage to reach the conclusions that they do, so sometimes they're described as uninterpretable black boxes, lacking in transparency. I'll speak more about that later on.

First I mentioned adversarial examples. This was really the first illustration of this problem, this is from 2015. What's going on here is that a deep learning system has been trained on many, many examples, to be able to tell the difference between many different thousands of object classes. After it's been trained, it's shown this image on the left, and you can see underneath, that happily, the algorithm does think that it's seeing a panda, and it's actually quite confident about that. It has 58 per cent confidence, which is quite high if you keep in mind that there are many thousands of possible labels it could provide. That's quite good. What's worrying is if we add just a tiny amount of this weird picture in the middle, we add just .007 of this strange image which to us just looks like weird colourful noise, then the resulting image we get is shown at the right - to us it looks identical to the image on the left. It's imperceptibly different. But when that's shown to the algorithm, it now has 99 per cent confidence that it's seeing a gibbon. That's not good.

There's nothing special, either, about a gibbon. We could have picked any other class. For example, you could have picked a chair or a table or a tree, and then if we were careful, we could have chosen a different weird image in the middle, such that if we'd just added a little bit of it, we would have fooled the algorithm again. You might think that it would be easy to avoid these problems, but actually no, it turns out at least for now, that these kinds of adversarial examples are a somewhat universal phenomenon, which is worrying, and let me show you an example in the real world. I'm going to show you this video a few times and explain what's happening. First please look at the right-hand side. Look at the right panel. What's happening is, we're in a car which is driving towards a stop sign. I'll keep showing this a few times. We're driving towards a stop sign. At the bottom, you'll see what an AI system thinks it sees. Imagine we're trying to train an autonomous vehicle to know what it's going towards. On the right-hand side, the good news is that the AI thinks it's seeing a stop sign, which is quite right, and therefore it knows to slow down as it approaches the stop sign.

What's worrying, though, is if you look at the left-hand side, as we get close to the stop sign, you'll notice that a naughty person has stuck some black and white tape on the stop sign in special places. I don't know if you can see that. I'll show you again. Because of that black and white tape, you'll see underneath, at the bottom, in fact the AI system doesn't see a stop sign, it sees a speed limit of 45. It only sees the stop sign when it's got very close, too close to stop in time. This is a real-world problem and something which active researchers are working on. Of course this is really necessary if we're going to be able to trust these kinds of systems in the real world. We need to be able to find systems which are reliable and robust to these kinds of tricks. I should point out that there are many valuable aspects of machine learning and AI that go beyond deep learning. There's exciting work going on in all these areas, with great real-world applications, but overall, the machine learning research paradigm has always been to train on data in the lab and then assume that you're going to get very similar conditions when you go out in the wild.

That's often not the case. Think of autonomous vehicles, which must face all sorts of conditions in the real world. Can we insure that an AI system can reliably navigate unexpected challenges that it hasn't seen before? Here I'm showing a challenging environment even for an experienced human driver. The point is that we can't train an AI system on every possible setting. The system must be able to generalise appropriately. It needs to have some ability to use common sense. Let me speak a bit more about that, and return to this idea of trying to label pictures. Here we're showing a nice countryside picture, and again we've got an automated caption underneath which at first looks fine. It says, 'A close-up of a hillside next to a rocky hill,' but if you look at the tags underneath it says, 'Hillside,' fine, but then it says, 'Grazing,' and 'Sheep.' Those are a bit worrying because there aren't any sheep in this picture. What's happening, I think similar to what we saw before, is that the sorts of pictures that this algorithm was trained on, that looked like this, often do have sheep here, and actually if you squint a little bit and look at this picture, you might imagine that perhaps there are sheep here.

The next tag is perhaps a little bit more perplexing, 'Giraffe,' but again if you look at the left part of the picture, maybe and you screw your eyes up a little bit, perhaps you can imagine that you could see a giraffe there. What's important is that when we look at this image, I suggest the thing which prevents us from being fooled is that we wouldn't just look at the static image. If it was a real setting, we would try to move our head around, we'd try to see how things are moving, and relate that to what we think the objects are. As far as our current level of what we're capable of with AI, we can't do that yet. It's a very interesting and challenging open problem. We saw the mention of sheep there. Here we've really got sheep. These are baby sheep, lambs, and we've got pictures of a boy and girl holding a lamb, but because this algorithm doesn't often see children holding lambs, it's labelled the images as a man holding a dog, or a woman holding a dog. Lastly, if we again really do have sheep, but they're painted orange, now the AI system thinks that they've become flowers!

Let me show you a different example of a failure of common-sense understanding. This is from an article written by Douglas Hofstadter, who some of you may know, wrote a great book called *Gödel, Escher, Bach*. He wrote this in an article in January, discussing a state-of-the-art system, Google Translate. Google Translate can do a very good job sometimes, but here I'm going to show how it can go wrong. At the bottom left, we've got an English paragraph, and it starts by saying, 'In their house, everything comes in pairs.' That sets the scene for what's going to be important about this paragraph. 'There's *his* car and *her* car. *His* towels and *her* towels.' When this gets translated into French, the problem is that in French, the possessive adjective changes its form depending not on who does the possessing, as in English, but it depends on the gender of what is possessed, the object, here the car or the towels. In French this becomes, 'Il y a sa voiture, et sa voiture.' The same word, 'sa.' It's completely lost the sense of what was in the English. 'Ses serviettes, et ses serviettes.' Again, what's going on is that the algorithm is doing a perfect job of translating locally, but the general understanding of really what's going on has been completely lost.

We've seen that current systems are narrow, and lack a common-sense understanding, but still they can perform very well. How can we help to understand how they really work to help us know when to trust them? This leads us to the issue of transparency and interpretability. Complex systems are hard to understand. Often we'd like to go under the hood to see what's going on underneath. But what really is transparency? I suggest that in some ways, transparency is like fairness. Both are ideas that almost everyone likes the sound of, yet both things can mean different ideas to different people in different contexts. I have to warn you that I'm going to build to a bad joke, so apologies for that in advance. Here's a picture at the bottom left. The picture is of a man on a swing. He's looking at a young girl, and he's saying, 'Fair is fair, Amanda, now push me already.' Clearly his idea of fairness is different from that of the girl’s. We might say that fairness is the sort of thing where you know it when you see it. On the other hand, perhaps transparency is the sort of thing… you know when you don't see it. Sorry, that was the bad joke!

More seriously, there are many different kinds of transparency. Here I'm just going to talk about some of them. For the developer of an AI system, it's very important for them to really try to understand how the system operates so they know when will it work well, when will it work badly? How can they try to improve it? How might they try to debug the system? For a user of an AI system, it's often important instead to understand one particular prediction or decision, so that user can understand and check how the system came to its conclusion. Often it's important to enable meaningful challenge. As an example, suppose a bank turns me down for a loan. I want to know, well, how did it come to that conclusion? Can I challenge it? Even more importantly, we're starting to see AI systems be used in criminal sentencing. If an algorithm tells me that I should go to jail for six years, I'd really want to understand, how did it come to that conclusion? Did it follow proper process? I'd want it to explain it in a way that would enable meaningful challenge, if I want to disagree with it.

A different kind of explanation might be important to provide an expert, perhaps a regulator, with the ability to audit a prediction or decision, particularly if something bad happens. Imagine an autonomous vehicle crashes. We might want to require, perhaps, for it to store all the data in the final few seconds, similar to the way a black box works on an airplane, so that the auditor could step through every part of those final few moments, and could try to understand what went wrong in order to try to assign accountability and liability. Each of these types of transparency motivates different measures which can be hard to define precisely.

Let me give you a taste for how transparency is being addressed technically. There are two main themes. One theme is to restrict the model class to be simple enough so that it would be easy to understand. For example, decision trees, and I'll show an example of that in a moment. The second theme instead says, don't restrict your model. Really try to train the best complex model you can to do whatever it is you want to do, and then use other tools to help understand the first model. Let's look at theme one. This is going to be a simple example of a decision tree. What I've done here is I've taken some examples from a tutorial at an important machine learning conference which is called the International Conference on Machine Learning. ICML. This is a tutorial by Been Kim and Finale Doshi-Velez.

This conference, ICML, was held in Australia, which explains the examples used. Let me talk you through what's going on at the right and explain how the example works. We have an input as shown on the right, which here is an owl, and ICML, the conference. We come to the first decision-point shown in the green circle. The question is, 'Is the animal an owl?' The answer is yes. When the answer is yes, you go to the left. You follow the left branch. Then the next question is, 'Is the conference ICML?' Again, the answer is yes, given the data we have above. We go to the left, and the final answer says, 'Raise your left hand.' I hope you don't mind if I ask you, if you don't mind, would you mind raising your left hand to show that we've followed what's going on? Thank you very much. See there are other actions to perform, depending on what the answer is. Let me try to ask if people can look at this and have a think, and then if you don't mind, try to do the correct thing after you've worked your way through. I hope it's not too disrespectful to ask you to do these actions in this very distinguished setting.

[Applause]

Good. Well, I hope you realise that even for this very small decision tree, it's actually not that easy to do this. Let me talk you through what's going on here, because actually some people didn't quite do the right thing. The data at the top was ICML 2017 Australia, kangaroo, sunny. We come to the first decision point. 'Is the year less than 2015?' Well, given the data above, the answer is no, because it's 2017. Then, 'Is the continent Australia?' The answer is yes, so we go to the left there, and the answer would be, 'Stomp.' It's tricky. Even with these simple examples, it's tricky. I won't ask you to do this next example, or the one after, but I hope I've made the point that even these simple seeming model classes can quickly become hard to interpret easily. In addition, it's hard to get a good understanding of what's really going on. As an example, which of the features are most important here in making the decision? It's tricky.

Let me turn to theme two. Theme two is where we build a complex model to achieve our objective, and then we use different tools to try to explain it. Within this theme, one popular family is called saliency approaches, and that's what I'm going to talk about here.

Here the first model was trained to tell the difference between huskies, we've got an example on the left of a husky, and wolves, as shown on the right. You can see that they look fairly similar, but there are subtle differences between huskies and wolves. We trained the model on lots of data like this, lots of labelled images, and then given a test image, we're going to see what our classifier says and then we can use a saliency method to identify which parts of the image are most salient or important in helping that first model reach its conclusion. Here's an example. After training, our model was shown this image on the left. It's really a husky. But the model's misclassified it as a wolf. Can anyone try to guess why? What is it about this picture that makes the algorithm think that it's a wolf?

M: [Unclear phrase 0:30:52.5 - too far away from recording device to be heard]

Excellent suggestion. The suggestion is that maybe it's snow in the background. That is a very, very good suggestion. Quite right. The automatic explanation here, exactly right, focuses on the snow. Very good. What's happened is, it turns out that on almost all the images on which this classifier was trained, the wolves had snow in the background, whereas the huskies didn't. What's important is, if we hadn't used our tool to figure out what was going on, what we typically would have done is we would have held out some data from our training set, we would've trained our model, and then we'd have tested it on the test set. It would have done very well and then we might have been very confident that the algorithm worked. We might have taken it with us when we went on an Arctic expedition, where it would have gone horribly wrong. It would have classified everything as wolves, because we have snow everywhere. These kinds of approaches to model interpretability are very useful for helping us understand our models and our data.

Let me next point out a potential problem with model explanations. We can differentiate between the intended audience of an explanation, and the likely beneficiary. Imagine we're using an online bookstore like Amazon, and it recommends a book to us. It used to be that you could click on the recommendation and ask it why did it recommend that to us. Interestingly, you can't do that anymore, but typically what it would say, it would go something like this. It would say, 'We think you might like book A, because in the past you bought book B, and you said you like book C.' Now we know that really what's going on in the algorithm is probably much more complex than that, but that's the explanation they're giving us. Why are they giving us that explanation? Well, I suggest that we should keep in mind that the motivation for them giving us an explanation is, they really want us to click on buying that first book A, so we need to be a bit cautious about just accepting their explanation. This presents a difficult problem. What we'd want in an explanation is what we might call a faithful explanation, that is, loosely speaking, something which gives us the truth, the whole truth, and nothing but the truth, but that can be difficult to define precisely. Let me give you another reason to be concerned.

This is a famous study from psychology that happened in the 1970s. This was back in the days when many people would go to a library to take out books, and if they couldn't take the book out, sometimes they would like to make photocopies of the pages that they wanted. There's a photocopier, and there's a line, a queue for the photocopier, and the experiment was trying to get people to push into the line. There were three different groups of people who would try to push in. The first group would just say, 'Please would you let me in,' and they wouldn't give a reason. It turned out that 60 per cent of the time people would let them in. Fair enough. The second group would give a reason. They would say, 'Please could you let me in, because I'm in a big hurry,' and then perhaps not surprisingly, compliance went up to 94 per cent. People would let them in much more often. But what's very interesting, and perhaps a bit worrying is that if people gave a fake reason, that is if they gave a reason with no information content - what they would do is, they would say, 'Please could you let me in, because I need to make a photocopy,' which was true of everyone in the line, obviously. Even if they gave that reason, the compliance rate was still about the same, 93 per cent. There seems to be something about people where, if we're given an explanation for something, we're more likely to go along with it. We need to be careful again, we need to try to ensure that when an algorithm gives an explanation, we really want it to give a full and faithful explanation.

Just to share a thought about that, I'd like to compare that to what human's do. How reliable are human explanations? Michael Gazzaniga has done a lot of famous experiments with people, trying to figure out whether people really know what's driving them to make a decision.

Actually it turns out that it's difficult for people to know what makes them make a decision. Often it's unconscious processing. Just to summarise this last paragraph here, he says that the reality is that listening to peoples' explanations of their actions is interesting, and in the case of politicians, can be entertaining, but often it actually is a waste of time. People tend often just to make up an explanation to be socially convenient or useful at the time, even if they may not be aware about that. If we want algorithms to just do a job that's as good as humans, we may be able to do that, because the bar is not that high, but I do think we should try to expect more of algorithms.

This next example I'm afraid is going to be slightly technical. Please bear with me just for a couple of minutes. This is looking at transparency from a different perspective. This is a very nice example of what's called Braess's paradox, due to Frank Kelly. The idea is that here we've got a network flow problem. Think about six cars driving in at the left. They're driving into that Node S at the left. Then the cars can either go up to U or down to V, and then they'll continue on their way to T, and then exit to the right. What's shown in red is the cost of taking a particular road. Here it's the delay of going down a road.

Sorry there's an equation there, I know, but don't be nervous about it. What it's showing is the delay as a function of X, where X is the amount of flow through that road, so that SU road says that the delay is 10X. What that means is that if one car goes down it, that car experiences a delay of 10 time unites, if two cars decide to go down it, then both of them experience a delay of 20. If three cars go down, then 30. This reflects, I suggest, the realistic idea that if lots of cars try to go down the same street, they're all going to be slowed down. It's a somewhat realistic assumption. In economics this reflects the idea of what are called externalities. Externalities are costs which are imposed on others, for things which I choose to do. It's known in economics that when you have a system with individual agents or people who are trying to do what's best for themselves, when there are externalities involved, that system may converge to a point which is not the best for everyone. What happens here is you let people keep moving around, choosing which path they want to take, until we settle down to what's called a Wardrop Equilibrium. That's a situation where no one can do better by changing their path, and that equilibrium setting for flow is what's shown in blue. Three cars go up, three cars go down, and they continue on their way. If you compute the delay, you'll notice that everyone has the same delay of 83 time units.

What is interesting here, and how this relates to transparency, is this. Suppose it turned out that all along actually there was an additional road that goes from U to V, but no one knew about it; and now it's made transparently available to everyone. The engineers amongst you might be thinking, well, gosh, we've made more capacity available in the system, this must be a good thing. Actually, I suggest maybe a better way to think about it is that we have given additional power to these selfish agents to try to optimise their own best interests. Because of these externalities this can lead to the system performing in a worse way. What happens is, initially some of the people who went from S to U took this U V route and they were better off, but then some of the people that went from S to V realised it was better to re-route and go to S U and come down. In the end, the equilibrium settles down to a place which is worse for everybody! Again, revealing the extra path, full transparency, makes everyone worse off. This is important, because there are sometimes regulations, like MiFID II you may know about, which is for financial markets in Europe, which insists that certain information is made transparent to everyone. We should just bear in mind that sometimes when you require that, it can lead to the system being worse off for everybody.

Let me just recap here some of the things that we've learnt about transparency. There are many different types of transparency with different motivations, and we need better ways to measure and provide them. Sometimes an explanation can be abused as a manipulation channel. In some settings, more transparency can actually lead to less efficiency. I should also say that we need to recognise that sometimes transparency is a means to an end, and not a goal in itself. For example, with autonomous vehicles, sometimes people say that we need to have transparency to understand how the algorithms work before we let these cars go out on the road, but perhaps we care more about qualities such as safety and reliability. Would we rather have cars driving around where we know how they operate, and they kill 100,000 people a year, or would we rather have cars where we don't know exactly how they work, but we are only getting 1,000 fatalities a year? Sometimes it may be worth investing resources directly into areas we care about, such as safety and reliability, rather than putting too much attention only on interpretability.

So far we've talked about privacy, influence, reliability, and transparency. Next I'm going to mention another critical concern, and that's fairness. The AI systems we develop must treat all people fairly. They must not discriminate against any individual or minority subgroup. This can present significant technical challenges, particularly if we're learning from historic data which reflects past human bias. This is already important in many commercial settings, such as selecting whom to interview when making a hiring decision, or making a loan, and even more importantly of course in criminal justice. We must treat all people fairly. Just to illustrate a very simple example where this is trickier than you might think, this is what might be considered a racist soap dispenser! Here's a soap dispenser, and you saw, when a light-skinned person put their hand underneath the soap dispenser, soap came out, but now here's a dark-skinned man putting his hand underneath, and no soap is coming out. I'm sure this wasn't intentional, but if we don't think carefully, these things can easily happen. What we'll see is, he's now taking a white tissue, and when he puts that underneath the soap dispenser, the soap is now going to magically come out.

This may seem silly, but it is worrying, and this has raised a call for more diversity among the people who are making and devising these systems. We need to try to make sure that a broad cross-section of society is involved in designing these systems which are going to affect all of our lives. Similar ugly effects to these can be observed even in cutting-edge systems. This is a famous example from 2015, where Google released an image classifier, which did a very good job of classifying many images across the top, but this image in the middle at the bottom, of course, was very worrying. This classifier was mistakenly classifying some people with dark skin as being gorillas, and of course, this caused a lot of upset and controversy. It turned out that this was not easy to fix. In fact, years later, earlier this year, this was the best that they could do. Google ‘fixed’ its algorithm simply by removing the category gorillas from its classes. It couldn't find a better way to do it. Sometimes it can be quite challenging to address these concerns.

Finally, let me talk about privacy and security. There are great opportunities to learn from our data. Health care is a great example. If we could pull together all our health care data, there are great opportunities to diagnose conditions at scale, cheaply, and perhaps even to find new cures for diseases. Other examples would be that we could learn many things from people's travel patterns, to help us do wonderful things like design better traffic light systems, or even help to design better cities. If we can examine transaction level data on payments of individuals and companies, this would allow high-resolution monitoring of the economy. All these great applications require data, yet users must feel comfortable that their private data is safe and secure. There's exciting ongoing work in technical areas such as, I'll just throw out some words which you may not remember, but these are important new areas of research here: differential privacy and cryptographically secure multi-party computation, to enable appropriate data-sharing, while giving users peace of mind, and the privacy and security of their information.

I'll start to conclude here. It's an exciting time for AI. Increases in data and computational power, together with improved algorithms, yield great opportunities, and even current narrow techniques can be used immediately in areas such as medical diagnosis, optimising transport, or other infrastructure systems, but we're still very far from general AI, or from the ability to do common-sense reasoning. Don't be misled by what you might have seen in shows like Humans, or Westworld, or Black Mirror. We're nowhere close to being able to do anything like that. We have seen a rapid increase in algorithmic systems in areas which directly impact our lives, such as hiring decisions, criminal justice, or ads. Increasingly we are seeing ads and media more generally through an algorithmic lens controlled by companies. As I mentioned before, these companies are incentivised to optimise for what we'll click on, rather than what's in our long-run best interests. Systems to enable personalised ads and messaging provide an ability to nudge many people a little bit, but that can have major consequences for selling products, and even for perhaps more important matters such as influencing elections.

There's nothing new about marketing or trying to convince someone of your point of view, but the scale that's now possible, and the level of personalisation is new. Of course, we should be wary about interfering with free speech, but we do have a history of thinking about situations where we might want to control messaging. For example, we outlaw hate speech, and we prevent the marketing of certain goods such as alcohol to minors. I suggest that we do need to think carefully about how to address the concerns of AI-powered influence on society. For effective deployment, we need measures of trustworthiness, including reliable performance, fairness, appropriate influence, privacy, and transparency. We're pursuing all these areas at the Turing Institute. But while technical approaches can push ahead the frontier of what is possible, we need to talk about these points effectively with the public, and to bring together interdisciplinary thinkers to address them, particularly since there are often trade-offs among these different desirable properties. We need together to decide what's right. If we do that correctly, then through informed discussion, we can develop AI systems to benefit us all. Finally, as we reflect on how algorithms should behave, it can give us fresh insight into our own weakness and our own morality.

We've been thinking about our own values for millennia, but now ethics has become a pressing issue for AI systems which are already deployed in ways that affect millions of people's lives. By thinking about how to improve AI systems, we can also sometimes improve ourselves. Thank you.

[Applause]

Ladies and gentlemen, on behalf of St. George's House, thank you very much indeed Doctor Weller for a fascinating lecture. Let's see if we can dive a little deeper into some of those things. Apart from the Turing Institute, of which you're part, who do you think is going to make the rules or set the global kite mark of the standards that we want in this area of work?

That's a great question. I should emphasise that Turing is not going to set the rules. Turing is doing work to try to bring together interdisciplinary experts to try to make progress on what is technically feasible, and what we might want, but of course, the decisions need to be made together as a society. Thinking about…

I was thinking maybe sometimes governments or regulators put a break on private enterprise. Do you see that happening in this area?

I should say that, to my mind, in the UK, the government has been admirably thoughtful about these issues. We've seen a lot of interest over the last few years: we've seen many different select committees in the Lords and in the Commons set up to examine these issues. We've seen wise reports from organisations like the Royal Society and British Academy, who brought out a report last year on data use and management, and the government commissioned a report from Dame Wendy Hall and Jerome Pesenti, which came out last year, that made certain recommendations. There was a general consensus that what was perhaps needed was a stewardship body to help consider these matters in the long term, not have reflex regulatory responses which I think would be the wrong idea, but to think carefully about these issues. The government did just set up a new Centre for Data Ethics and Innovation. They've established the chair, and they'll be appointing the rest of the board just next month to consider exactly these issues.

That's on a global scale? Are there similar things happening elsewhere?

Great question. That's UK. Other countries are also doing things, but we hope and believe that potentially in the UK we can help to be a leader in trying to set these ethical standards. We have a history of trying to lead the way, and successfully leading the way in some respects, with respect to regulation across finance and other areas. We have the British Standards Institute which plays a very important role in setting standards, and is looking at these issues. There's also, internationally the IEEE, which is similarly looking to set standards for these kinds of issues. Also I'd say that the big companies are trying to be helpful. They set up something called the Partnership on AI, bringing together all the tech companies to try to be proactive in figuring out what would make sense.

But we're in that usual dilemma of profit versus opportunity, aren't we, and presumably if somebody gets ahead, they want a return on their investment.

That's a great question. I think that specifically is why the UK government chose to call this new centre the Centre for Data Ethics and Innovation, to try to balance them. What I would suggest is that if we can do it right, we're going to enable the right kind of responsible innovation. I think people will demand and require that, and what we hope, what I hope to do is to help participate in establishing what is technically possible, so that we know the best place to set those ethical guidelines, which will best foster the right kind of innovation.

You've talked a lot about the algorithms. A couple of questions for me and then ladies and gentlemen, please do put your hands up and I will come amongst you with the microphone so you can ask your own. At the moment, am I right in thinking that those algorithms in terms of computer code or the theory behind them, they start off as a human idea, and they are interpreted by a machine, or are we now at the point where the machines can write their own algorithms to do the next job you've asked of it, or them?

That's a great question, and I think many people find it quite scary to think about computers writing code for computers, and could that mean that they'll run away and keep getting better and better and better? What I would say is for the vast majority of cases, it's humans writing code, but we are starting to develop techniques to help us write code more efficiently, and we're starting to write programmes which will look at what humans are doing and try to infer from that how they may be able to write a programme to do something similar. We're starting to get to that point.

In which case, is there a moral algorithm you can throw on top?

Yes. Yes.

If you wanted to.

Yes, well, this is a very interesting area. I suggest it's very - people for decades have been trying to come up with a simple code of ethical principles, like Asimov's three laws of robotics that some of you may be familiar with. Whenever someone tries to do that, inevitably you find that there are always cases which are not covered by the rules. It's very difficult to try to cover all the cases, and perhaps it's impossible. Then some people thought that what we might try to do instead would be to have a computer system watch humans, and try to copy what we did, but a big problem there is unfortunately, perhaps setting aside some of the good folk here, there aren't many people who behave in exactly the way that we would want a machine always really to behave. We need to be careful about that. Typically, I think what we might want is to try to see if we can train an algorithm do something like what it learns humans want other humans to do, rather than perhaps what humans do do, and we're not able to do that yet, but it's an interesting direction.

One example that I've heard reference to is, let's take one of our autonomous cars, and for some reason it's going to head for, it's presented with an instant where it's going to crash, and there are a group of six schoolchildren to the left of it, and with all due respect to the age of this audience, three senior citizens to the right of it. Does it knock over the old people, because they've had a good life,

and leave the young people alone, or vice versa? How does it make that decision?

That's a great question. This is an example of what are sometimes called trolley problems, where you have a really tough dilemma like this. Of course, who knows what the right answer is. We might all debate and come to different answers ourselves. Given that, I think it's sometimes a bit unfair to try to suggest that the algorithm must make the right decision, because really, if for example we were to say that, we would want the algorithm at least to do something within the realm of what a reasonable person might do, there's a very broad range of things that it might choose to do, and they would all be okay. On the other hand, I do think that it can make sense to hold the algorithms to a higher level of account than humans, and one reason for that is because if a human gets this sort of thing wrong, there's only so much damage which will happen. Of course, it could have terrible consequences, but it's still just going to be one case. But if we program an algorithm in a way which is wrong, it could happen thousands or millions of times, and really lead to a lot of damage, so we should be very careful about these questions. At the same time, I would just suggest that these specific questions, while they are philosophically very interesting, I don't think are going to be all that important practically.

I think that the vast degree of benefit we're going to get, at some point, from these vehicles - if we can get them to work - is that the number of deaths are going to drop by an enormous amount. Instead of 100,000 people killed a year, maybe we'll only get 10,000 people killed a year. Of course they're important questions about exactly which ones, but I think that big drop-off, or that big improvement in safety, might be a bigger issue than exactly what happens in those settings.

That short-term pain for long-term gain could be argued, in that [over speaking 0:56:06.9]?

You could argue that, but I would hope we don't get short-term pain.

Either - no pain. Yes. Ladies and gentlemen, any questions from your good selves? Let me come down the aisle. If you'd like to just indicate - we have a couple of microphones if you'd like to put a question. Yes, we have one at the back here. I'll just come and pass the microphone to you. Where was the hand? Do have the microphone, sir.

M: Thank you. My name is Chris Rees, I'm president of the British Computer Society. You said that we're still far from general AI, or common-sense reasoning. Are we ever going to get there? May I suggest that actually it's not something to worry about, because it's impossible, because there's no conscience, and there's no consciousness in machines.

Great question. You're raising very deep issues. Will we ever get to general intelligence? We don't know.

People are very interested in working towards that. They're starting to take small steps. As an example, can we train a machine to learn in one domain and then try to abstract some kind of concept which can be used in another domain? Sometimes this is called transfer learning. We're making initial progress in that direction, but I think you're referring more to the kind of widespread general intelligence which humans have, and you mentioned the word 'consciousness.' There's a lot of debate in the community about whether or not this will ever be possible. Some people say, 'Well, a human is just made up of atoms arranged in some particular fashion, and that's evidence that if we can arrange atoms correctly, we can achieve human-level intelligence. We see it in humans.' Other people say, 'No, we're much more than just the mechanistic sum of our atoms.' This is a long philosophical debate, and I won't try to go further in it now.

Any other thoughts from our audience? While we're just thinking about it, do indicate and we'll bring a microphone to you. You must have heard this before, but Professor Stephen Hawking was quoted as saying, 'The development of full artificial intelligence could spell the end of the human race.' Was he right to be so pessimistic?

I think, if I recall correctly, that the full quote is, 'It could be the best thing or the worst thing for the human race,' which isn't, I know, an awful lot more optimistic, but at least it opens the door to optimism!

You throw the coin I suppose, toss a coin and see what happens.

There are many people who have legitimate concerns about the long-term consequences of intelligent machines, particularly if we can make machines which start to be self-motivated. We don't necessarily need to have anything like consciousness. We perhaps only need to have machines which are very competent at doing particular tasks, in a way which could eventually become problematic for us. You could imagine, someone might be able to make a system which takes instructions. Perhaps it's an autonomous vehicle, and you'd just ask it if it could please, take me as quickly as possible from Windsor Castle to Paddington Station, and it might dash off at very high speed, going through traffic lights, smashing through windows, going through the middle of buildings, paying no heed to what's going on, because it was following exactly what it was told to do without the type of common sense needed to make sure that it didn't go wrong. I wouldn't dismiss those concerns. What I would say is that at least for the foreseeable future, we're far away from algorithms being out of our human control, but I think it is important to work on those issues.

There are some people working on them, and my particular thinking about this, which is shared by some others, is that it may be difficult to know exactly the capabilities of systems far in the future, but if we start today working on methods which we know are going to be helpful now, such as the ideas I mentioned, fairness, transparency, privacy, et cetera, those are needed now, and they will also likely be helpful down the road.

I think we have a question on the blue microphone.

F: My name's Katharine Scarfe Beckett, I'm here with Lily Innovation Advisors, a small company. Given that we grant human beings 18 years of generalised experience and access to tutorial bodies driven in their experience during those 18 years by a desire for sleep and food and an interest in the opposite sex, do you think there might be any way to compress those 18 years, and if so by how much, and why, in order to achieve some sort of more human common sense in machines, or is the plan simply to shortcut all of that and assume that we can get there without that kind of experiential basis?

To check I understand, the question is, might we be able to compress 18 years of human experience into a shorter time frame, for a human or for a machine?

F: For a machine.

This raises a range of issues. One interesting question is, how should we think about algorithms which are learning? Should we think of them in some way as being like children, or should we think of them as being like an animal where we still bear responsibility for them, but we recognise that while they’re learning, they perhaps might do wrong? Overall with that analogy, I think we need to be very careful. I think we always need to hold to account whoever it is deploying an algorithm. I don't think we should go down the path which some people have suggested, that allows us to say, 'Well, no, it's not my fault, my algorithm did it,' by ascribing personhood to an algorithm. I think it's a very dangerous idea which we should not do, but specifically with regard to learning, first we don't know how to train an algorithm to get all the sorts of abilities which even a one year-old has, never mind an 18 year-old, so it's difficult to know exactly how that would work, but assuming that we were able to do that somehow, I would imagine that it would be very likely that we would be able to compress it.

AI systems are already much better than humans at all sorts of particular things. One of the things that they're much better at is being able to absorb a tremendous amount of information in a short amount of time. I think it would be very likely if we could get to that point, that you would be able to compress it very dramatically, and even if initially it took fifty years, given that the speed and power of machines doubles roughly 18 to 24 months, and that seems to be a trend which is continuing, although it may not, but if it does, even if it started out being 100 years, after just a few years it would be down to one year and then less and less. Does that make sense?

Just briefly, is it worth tagging on our chess machine, which you and I spoke about beforehand, which, I was astonished, because I'm not a very good player of chess at all, but it is, the DeepMind computer, or that project managed to teach a game - it had never played the game chess before, and it went from no knowledge of the game to a grandmaster in the space of a day, or a matter of hours. That's quite extraordinary a bit of machine learning, isn't it?

Absolutely. Do you want me to comment on that? Just in case people aren't aware, there was a very impressive development that came out from DeepMind, a company based in London, where they first made a system which could beat a very good Go player at Go. Go is a game you play on a board, and you just have black and white pieces, it's a game of deep strategy. It was sort of on the list of games which computers were not better than people at playing. If you were to go back a few decades, people thought that playing chess was the highest form of human intelligence, but once computers got better than humans at playing chess, we recalibrated, and we decided, actually chess was a very simple thing! It was nothing to do with real intelligence, and there were much better things that we could do. One of the games which survived, which computers could not beat humans at, was this game of Go. That was until around a year or so ago, when DeepMind came up with a computer system which could beat the world's best Go player. Again, though, people immediately recalibrated and said, 'Well, again, Go is just one of those games, you play it on a board, it's got logical rules, you can see everything that's in front of you. Of course it's something that a computer would get better than a human at. It's just like making lots of computations.'

I think that's a bit harsh, because I think there was a lot of cleverness in what they did to make the program which beat the world's best player at Go. Indeed, there were some of the moves which it played, which were sometimes described as demonstrating creative genius, and of course there are deep questions around whether or not that really is possible for a machine, but it was very interesting. Then what they did is, they improved their algorithm again, and they made this algorithm called AlphaZero, which was no longer specialised for Go. You could give it the rules to a whole range of games that could just be specified very specifically, games where you could see everything that was going on. They're called games of perfect information. It would play against itself for a while. As you mentioned, it was typically only a few hours to get to human levels of ability, but we should qualify that and note that it was a few hours where it used many, many, many different processors. Maybe thousands of processors at once. It was actually using a lot of computational power for a short amount of time.

Still it was remarkable, because that machine, that algorithm, after a few hours, was much better than their previous specialist algorithm that only played Go. Also they could turn that algorithm on to chess, and again, it became the world's best chess player, to such an extent that you can see online, if you look up AlphaZero on YouTube, you can see, they've got some of the games that it played that they published. It's a bit, arguably a bit fishy, because they only published a few of the games out of many that it played. Anyway they chose some games where it played some very brilliant moves, which have to some extent really revolutionised the way that humans think about chess in certain situations. It is very interesting, and very impressive.

Thank you very much indeed for that. Just a very quick light-hearted thought on this famous conversation, I'll see if I can play it for you. [Plays video clip from *2001* film] You recognise that? Will that ever happen to us?

I missed the first bit.

The HAL supercomputer basically saying, 'I'm sorry, Dave, I'm afraid I can't do that.'

It's… Yes… I'd suggest this is another form, really, of that question about the long-term. I would say that this specific problem, it is serious, and it is being discussed. It's sometimes called the off-switch problem. Very smart people, for a long time, were saying, 'Don't worry about machines getting smart, because if they're going out of control, all you have to do is turn them off.' But then someone pointed out that as soon as your algorithm really starts to develop any kind of general intelligence, it will figure out various, what are called instrumental paths to achieving its goals. It will realise that certain things are important for it to be able to do the things we'd all want to be able to do, and it will realise that one of the very important things for it to be able to achieve anything, is for it to be on. It will notice that if you try to turn it off, it will probably try to stop you turning it off. People are putting serious effort into trying to figure out how to control for this.

One idea, which seems to be quite effective, is the idea that you make sure that the algorithm can never get very confident. In particular, it doesn't get too confident about what it thinks you want it to do. Then if it sees you reaching to turn it off, it will think, aha, I was wrong in what I thought this person wanted me to do, and I will let it turn me off.

Keep it in the dark. At that point, before they turn the lights out, I think we're going to offer your thanks, Doctor Adrian Weller, thank you very much indeed.

Thank you.

I'm going to invite Gary McKeone now, our programme director here at the house, to give a vote of thanks.

[Applause]

M: Mr. Dean, Ambassador Elson, ladies and gentlemen, it's my privilege as programme director to offer a short vote of thanks to Doctor Weller and to Martin Stanford for this evening's lecture and interview. The Warden of the House, Canon Finlay tells me that this is the moment to summarise and respond to what we've just heard. As someone who can barely turn on a mobile phone, I think it would require a very special algorithm to do that, so that's not what I'm going to do at all, but what I will say is that this evening's lecture has been hugely informative, highly intelligent, and indeed, entertaining. It seems to me that with artificial intelligence, we're clearly on a journey, and like the best journeys, let's hope, well, the travel will certainly be exciting, but let's hope that that final destination will be for the greater good. The Elson Lecture as you know has a very distinguished pedigree, and Adrian, tonight you have added significantly to that pedigree, so before we adjourn to the Dean's Cloister for a short drinks' reception, please join me one last time in thanking Doctor Adrian Weller for delivering the 2018 Elson Ethics Lecture.

[Applause]

[END OF TRANSCRIPT]