Module 2.2 Artificial Intelligence and Explanability

I'm Stefan Buijsman a philospher at the TU Delft and today we are talking about Explainability. Imagine that you've been stopped by border control simply because an algorithm assigned you a high-risk score. Your answers to the questions posed by the system were deemed unsatisfactory. And while you can ask for an explanation, but it's unlikely that you'll ever get one. The algorithm was, namely, one that goes through millions of calculations before delivering the answer. Calculations that, moreover, have been adjusted based on thousands of examples while the AI was being trained. All the officers using the system get to see are the calculated scores, and even if they had access to the algorithm itself it would be impossible to figure out why it settled on exactly that score. It's a frustrating situation, and one that is all too common with modern AI systems. Efforts to make these algorithms more transparent are underway, but the problem is largely unsolved at the moment.

It is possible, for example, to calculate what the influence is of different input parameters. Let's say we have an image and an algorithm that detects the objects in that image. When the AI outputs that there is a sheep in the picture, an explainability method known as a saliency map can then highlight which parts of the picture were the most important for the algorithm's answer. If the saliency map shows that the algorithm mostly looked at what we also recognize as a sheep, then all is good. If, however, it highlights the grass it's standing on, having learned that sheep are mostly found in grassy areas, then something is off. And sometimes that can help us understand what might be going on. We can draw a link with the grass, and likewise we can guess what happened when the algorithm claims there's a cow in the picture, when the saliency map highlights a black sheep. Probably the AI wasn't trained to recognize black sheep, and so it picks an animal that's similar and has black spots more often than sheeps do. But there's no guarantee that we can interpret these saliency maps, or that these guesses are even in the right direction. It can happen just as easily as with the bottom pictures showing with two birds, highlights the right most bird to say that there is indeed a bird on this picture, whereas the other bird in the picture, the one on the left, is highlighted to say that there is a 'person' there. Why these two birds led to such different conclusions in the AI, nobody knows. And so the technical challenge is still a very tough one to solve. Even though we do have some tools available then, they can at best gives an educated guess into the reasons for an algorithm's output.

Does this always have to be a problem though? Do we care that our algorithms are hard to explain if they are accurate enough? Does it, for example, have to be understood on what basis an AI in a factory picks out defective products if it does so reliably?

Or, for a more high-stakes example, is it necessary to understand medical algorithms when they are better at diagnosing illnesses than most doctors? Is it allowed to use these AI tools even when we don't know exactly why they present us with a certain diagnosis if they perform well?

One argument you'll often hear in support of this idea that accuracy is enough because people aren't great at explaining why they make decisions either. There are unconscious biases that affect our decision-making and sometimes we aren't all that sure how or why we do something either. In fact, if we could perfectly explain what procedure we follow to make these decisions, so the idea goes, then we would have been able to write transparent algorithms that go through these same steps. Since we can't offer these explanations, and need cognitive scientists to uncover aspects of our own decision-making, why should we expect algorithms to provide perfect explanations? Aren't we introducing a double standard here?

I personally think that we can still point to relevant differences between AI algorithms and people. Doctors may not be able to explain everything about their decision-making process, and neither can police officers at a border control point, but they can offer you a set of reasons that can be evaluated. Were they relevant and sufficient for the decision that was made? This is something that experts would be able to determine, and a basis for holding the relevant people accountable for the decisions that they have made. That kind of verification and accountability becomes incredibly difficult with algorithms that cannot, as of yet, offer such reasons suitable for rational deliberation. When the decisions are important, and have a big impact on our lives, it seems reasonable to want access to such explanations.

Another consideration that is relevant, is that without explanations algorithmic scores can be hard to evaluate even for experts. Doctors presented with an AI risk score might not know when to disagree with the AI, even when algorithms can be blatantly wrong. A good example is NarxCare, an algorithm used in the US to estimate patients risk of having an opioid addiction. If the risk score is too high, doctors shouldn't prescribe pain killers to these patients. And if they ignore these scores, they can in some states lose their jobs. So what to do if the AI is wrong, and as has actually happened it assigns a high risk score based on medication given to a patient's dog? Without explanations, and with only a risk score, the doctors are likely to deny the patient access to important medication. And this is in fact what happened. So that even if the algorithm is generally reliable, the overall system still needs support for when the AI makes mistakes in individual cases.

Explainability, then, is a real challenge for algorithms. But not only is it a technical question of how we can make these algorithms more interpretable. There are also broader questions about the role we want explainability to play, the situations in which we think it is necessary and should then perhaps avoid using advanced AI for the moment and those where we can do without explainability. When is accuracy enough, and when do we require more than just good performance? Those are some of the questions we are tackling here in the MOOC and thank you for watching.