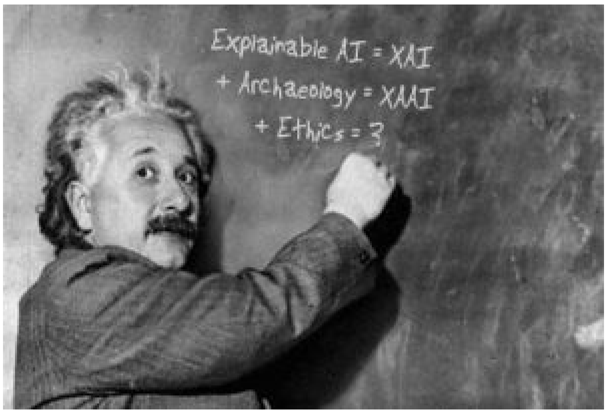
Explainability in digital systems

written by Jeremy Huggett | 22/01/2019



Created via http://www.hetemeel.com/

Some time ago, I suggested that machine-learning systems in archaeology ought to be able to provide human-scale explanations in support of their conclusions, noting that many of the techniques used in ML were filtering down into automated methods used to classify, extract and abstract archaeological data. I concluded: “We would expect an archaeologist to explain their reasoning in arriving at a conclusion; why should we not expect the same of a computer system?”.

This seemed fair enough at the time, if admittedly challenging. What I hadn’t appreciated, though, was the controversial nature of such a claim. For sure, in that piece I referred to Yoshua Bengio’s argument that we don’t understand human experts and yet we trust them, so why should we not extend the same degree of trust to an expert computer (Pearson 2016)? But it transpires this is quite a common argument posited against claims that systems should be capable of explaining themselves, not least among high-level Google scientists. For example, Geoff Hinton recently suggested in an interview that to require that you can explain how your AI systems works (as, for example, the GDPR regulations do) would be a disaster:

> “People can’t explain how they work, for most of the things they do … If you ask them to explain their decision, you are forcing them to make up a story. Neural nets have a similar problem. When you train a neural net, it will learn a billion numbers that represent the knowledge it has extracted from the training data … . If you put in an image, out comes the right decision, say, whether this was a pedestrian or not. But if you ask “Why did it think that?” well if there were any simple rules for deciding whether an image contains a pedestrian or not, it would have been a solved problem ages ago.” (Simonite 2018).

This triggered some strong responses on ethical as well as practical grounds (e.g. Jones 2018). So maybe we need to unpick what is actually in dispute here since, as Cassie Kozyrkov (Google’s Chief

Decision Scientist) has observed, people are frequently talking past each other and often actually talking about different things (Kozyrkov 2018a). For example, many (including Hinton, above) see an equivalence between the working of the human mind and that of a neural net: if we don't understand how the wetware works, why should we expect to understand how the software operates in arriving at its conclusions? As Kozyrkov suggests,

> "If you refuse to trust decision-making to something whose process you don't understand, then you should fire all your human workers, because no one knows how the brain (with its hundred billion neurons!) makes decisions." (Kozyrkov 2018a)

However, this isn't the full picture. Heather Roff (Leverhulme Centre for the Future of Intelligence), for example, points out that we are able to interrogate the human mind and so

> "… to claim that humans are inherently opaque and non-transparent and that justifies us using other intelligence that are actually more opaque and inherently nonhuman-like in their reasoning as a justified argument is a false equivalence. Humans have a theory of mind. AIs right now do not. I don't have a sense of what another being like me may think, if I'm an AI. I DO have that as a human being. And this excuse — as an attempted justification at using tech that we don't understand fully — is a red herring." (Jones 2018).

Ann Cavoukian (Privacy by Design Centre of Excellence, Ryerson University) agrees:

> "… there is a meta-algorithm in the brain that is able to view the process of decision-making and collect the sequence of features that were involved in the decision, and based on those, output the explanation. Again, this cannot be done with existing deep learning because the features are implicit, meaning that they are buried in the parameter values" (Cavoukian 2018; Jones 2018).

So if the equivalence argument falls as a justification for taking AI solutions at face value, what else is at issue? Another common claim is that a system that is essentially designed to be autonomous – such as a self-driving car – where we can't manually program all the options so it teaches itself and generates a vast number of complex decision-making processes that we frankly don't understand, is a system which does not require explainability. It just is, and we humans trust it because it works, without needing to ask how or why. For example, Cassie Kozyrkov asks:

> "Imagine choosing between two spaceships. Spaceship 1 comes with exact equations explaining how it works, but has never been flown. How Spaceship 2 flies is a mystery, but it has undergone extensive testing, with years of successful flights like the one you're going on. Which spaceship would you choose?" (Kozyrkov 2018a).

She suggests that Spaceship 2 is her preferred option, as careful testing is a better basis for trust.

For my part I'd prefer Spaceship 1, but wait to fly until it has undergone extensive testing! And that is perhaps the point – it is the mystery behind Spaceship 2 that is its problem. As Will Knight points out in the context of driverless cars, if it goes and does something unexpected, in the absence of explainability how can you find out what happened and why and fix the problem (Knight 2017)?

Kozyrkov sees the key distinction as lying between applications that generate inspiration for human decision makers (e.g. Kozyrkov 2018b) and building safe and reliable automated systems where performance matters most (Kozyrkov 2018a), and that there is consequently a trade-off between explainability and performance. Similarly, it is often said that explanations may be unnecessary where the decisions are not crucial or where there are no unacceptable consequences (e.g. Doshi-Velez and Kim 2017, 3). This is perhaps an ideal end situation – for the reasons described above, it seems unwise to assume that black-boxed automated systems can be implemented without some degree of explainability during their development phase. While archaeology may not be mission-critical, we should not black-box archaeological systems that classify or categorise data without requiring some understanding of the basis on which they draw their conclusions. So we could perhaps see most archaeological systems as falling into the 'inspirational' category – for the most part we aren't talking about systems that are fully automated black boxes and instead anticipating systems that fall into the decision support category in which the ability to provide an explanation remains critical. Interestingly, we've been here before as archaeologists – back in the 1980s Arthur Stutt developed what he called an Argument Support Program for Archaeology; an expert system which, through incorporating argument and debate between user and knowledge base and modelling different viewpoints, essentially sought to provide justification for – and hence explain – its conclusions (e.g. Stutt 1988; Patel and Stutt 1989).

A remaining question is what actually constitutes a machine-derived archaeological explanation that would be considered acceptable to a human? At what level is appropriate? And are such explanations simply stories spun by machines rather than anything more substantive? Something to consider another time …

References

Cavoukian, A. 2018 'Response to AI Explainability', https://www.ryerson.ca/content/dam/pbdce/papers/Response_to_AI_'Explainability'.pdf

Doshi-Velez, F. and Kim, B. 2017 'Towards a Rigorous Science of Interpretable Machine Learning', *arXiv*:1702.08608v2 [stat.ML] https://arxiv.org/abs/1702.08608v2

Jones, H. 2018 'Geoff Hinton Dismissed The Need for Explainable AI: 8 Experts Explain Why He's Wrong', *Forbes* (20th Dec). https://www.forbes.com/sites/cognitiveworld/2018/12/20/geoff-hinton-dismissed-the-need-for-explainable-ai-8-experts-explain-why-hes-wrong/

Knight, W. 2017 'The Dark Secret at the Heart of AI', *MIT Technology Review* (11th April). https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/

Kozyrkov, C. 2018a 'Explainable AI won't deliver. Here's why', *Hacker Noon* (16th Nov). https://hackernoon.com/explainable-ai-wont-deliver-here-s-why-6738f54216be

Kozyrkov, C. 2018b ‘What Great Data Analysts Do – and Why Every Organization Needs Them’, *Harvard Business Review* (4th Dec). https://hbr.org/2018/12/what-great-data-analysts-do-and-why-every-organization-needs-them

Patel, J. and Stutt, A. 1989 ‘Beyond Classification: the Use of Artificial Intelligence Techniques for the Interpretation of Archaeological Data’, in S. Rahtz (ed.) *Computer Applications and Quantitative Methods in Archaeology 1989. CAA89* (BAR International Series 548; British Archaeological Reports: Oxford), pp. 338-347. http://proceedings.caaconference.org/files/1989/30_Patel_Stutt_CAA_1989.pdf

Pearson, J. 2016 ‘When AI goes wrong, we won’t be able to ask it why’, *Motherboard* (6th July). http://motherboard.vice.com/en_uk/read/ai-deep-learning-ethics-right-to-explanation

Simonite, T. 2018 ‘Google’s AI Guru Wants Computers to Think More Like Brains’, *Wired* (12th Dec). https://www.wired.com/story/googles-ai-guru-computers-think-more-like-brains/

Stutt, A. 1988 ‘Second Generation Expert Systems, Explanations, Arguments and Archaeology’, in S.P.Q. Rahtz (ed.) *Computer and Quantitative Methods in Archaeology 1988. CAA88* (BAR International Series 446; British Archaeological Reports: Oxford), pp. 351-368. http://proceedings.caaconference.org/files/1988/23_Stutt_CAA_1988-II.pdf