Q&A
20 October 2021
How can Explainable AI help scientific exploration?
Professor Carlos Zednik (Eindhoven University of Technology)

Q1: The abstract of your talk says explainable AI tools can be used to better understand what a "big data" model is a model of. In scientific research, why wouldn't the scientist who made the model know what it is a model of?

Good question, in a very broad sense, of course they know. So, if you're putting patient data into a machine learning model, then the system is going to extract something from that patient data. And you could say, well, it's a model of that patient data. So, at some very broad level of analysis, you can say of course you already know what the model is a model of. But what you don't know, and this is what machine learning is really good at, is that the data might contain regularities that maybe we did not know of in advance. So, we did not maybe know that there is a correlation between, for example, sleep apnoea and diabetes. Maybe the correlation between those things was unknown (although I think in this particular case it was), but what the system has learned to detect is, or track is, exactly the fact that sleep apnoea in patients can be used to predict adult-onset diabetes, and so in that sense our model is a model of this relationship that we did not know in advance – even for the medical practitioners. So, in a broad sense, yes, we know what the model is a model of, but in a very specific sense we do not.

Q2: Transparency in this context is also about building/reinforcing Trust. Doesn't this require that anything brought out through transparency needs to be understandable to the user?

So again, user here is ambiguous. Remember there is this discussion of agents and stakeholders in the ML ecosystem. For users in this case, I would think users are something like decision subjects. The people were affected by the decisions, so these might be the people who are denied credit from a bank because of the use of a certain AI system. And if they then, assuming the GDPR has those teeth, (which is a matter of debate) say to the bank; "Hey, I need to know why I was denied a loan", then the explanation that is given by the bank should be of course understandable to the user. And that's exactly why these explanations are agent-relative, and these explanations should cite these epistemically relevant elements that are appropriate to the agent or to the stakeholder. I would argue that for an end user, I suppose a layperson, the appropriate ERE's are precisely features of the environment, that are sensible or meaningful or easy to interpret, such as income level or, more problematically, race or gender or home address. So, of course, depending on who is requesting the explanation, the explanation should be understandable to that person. An interesting sidebar here is these explanations now cite features, that are not actually in the black box at all, so an "explanation" of a "black box system" in this case is not opening the black box; it's citing features of the environment that are being tracked.

Q3: My question is regarding robot emotional intelligence, not whether machines have any emotions but whether machines can be intelligent without any emotions?

First of all, I don't really know either what intelligence is or what emotions are. I like to think in terms of systems that are able to behave in flexible ways and adapt to their environments, and at least for human beings and presumably animals, emotions are one way of doing that. Emotions are a mechanism for adapting to certain situations when we're in a certain emotional state that might be a way to protect us. For example, in a vulnerable situation, or to run away, or to be angry and feel strong and so on. So, these are a kind of emotional response to deal with unpredictable environments and in the sense that we want to develop systems that can rival our levels of adaptivity and flexibility and so on, we might need to implement similar mechanisms such as those. Whether we want to call them emotions or not, I don't know. Whether those systems then will have this feeling of emotion, this kind of conscious aspect of the emotion, I don't know. And I'll be straightforward, I'm the wrong person to ask there. I'm usually interested in more measurable aspects of behaviour and cognition than the ones that we cannot measure.

Q4: If weak emergence of high-level features is possible, would this undermine at least some of those XAI strategies – presumably, we would no longer be identifying the way lower- and higher-level features relate?

So not all those XAI methods aim to relate the low- and high-level features, some XAI methods just aim to characterize those high-level features. So, if we can characterise the representational structures in a system, we might not need to know how those representational structures are implemented. In particular; parameters, or variables. Of course, that's a tough thing to do, but that's one way to answer the question. Another way to answer the question is to try to characterise the behaviour in a compact way. So, a method that I didn't talk about, like LIME, just tries to linearly approximate the system's behaviour, and so we don't need to look at the underlying structures, parameters, and variables of the of the network. In this case, we just need to look at the behaviour and approximate this. For certain stakeholders in certain situations, certain contexts, that might be enough, but you're right, the kind of complexity you mention makes the task of course difficult for other methods.