Response to AI Explainability

Ann Cavoukian, Ph.D.

Geoff Hinton's view that we should regulate 'after the fact' is based on the premise that current deep learning algorithms, at times comprising up to a billion parameter values (which to a human observer will appear to be a massive number of random numbers), are far too complex to make sense of, especially when it comes to "self-explaining" why they made certain decisions. He then draws the analogy that if the algorithm could self-explain, it would be like asking a human being to explain its decisions in which they would simply make up a "story" (or an explanation) – inferring that the story may be fiction since we are not knowledgeable of the actual inner workings of our own brain. Therefore, we should base our regulations on algorithmic outcomes, since attempting to infer why an algorithm made a decision as a pre-emptive exercise is very difficult, if not impossible.

I believe Geoff may be correct when it comes to humans explaining decision which have emotive content. The emotional (limbic) system is far below the level of consciousness, and if asked "why did you throw that plate against the wall, in apparent anger," you may bring up various childhood harms, workplace grievances, spousal abuse, or whatever. However, it may be characterized as a "story," since it would consist of your interpretation of the conditions leading to your actions. And that may very well be viewed as "fiction" in comparison to the true neuronal network reasons.

However, if I ask a human being a non-emotive question such as, "why did you categorize this image as a dog?" their explanation may comprise, "because it has four legs, a tail, a cylindrical body, ears, a fur coat and a snout, all spatially positioned in a way – and this is key – such that my previous experiences with such images as this would have been categorized as a "dog." And their explanation would be correct.

Two things are going on here: first, the human was able to externalize the features that make up a "dog." However, with current deep learning algorithms, although they may initially decompose an image into relevant features, and then recompose those features back into an image for categorization, these features are implicit to the algorithm, buried in the myriad numbers of parameter values. Current algorithms cannot externalize these features and use them to explain a decision. What is needed are algorithms that construct "wholes" from previously learned "parts/features" such that the features are also external to the algorithm that is making the decision. Current work on hierarchical generative models based on the composition of 'external' features may be a potential solution to this problem.

However, there is a second process taking place: 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, and moreover, any one parameter value may affect features associated with categories other than "dog," in the above example. So Geoff is correct in that, with current deep learning models, it appears to be an indecipherable patchwork.

Although I was a regulator for many years (three terms as privacy commissioner), I don't share Geoff's view of after-the-fact regulation. Technology is simply moving far too fast, and regulations, in this day and age, are a lagging remedy. We must be pre-emptive and proactively build-in explainability. However, to implement useful explainability will require different artificial architectures from the existing ones.

For the last ten years, my colleagues and I have proposed that we must develop artificial agents by combining existing methods with embodied cognition and evolutionary computation, in a virtual environment. In the last few years, the technology to test this has arrived, thanks to NVIDIA, Magic Leap, Intel, IBM, and all the video game developers. But it's a paradigm shift that most machine learning developers have not yet been schooled in. However, for those forward thinkers who want an introduction to embodied cognition, as a psychologist, I recommend two books by psychologists: "Louder Than Words – The New Science of How the Mind Makes Meaning," by Benjamin K. Bergen, and "Surfing Uncertainty – Prediction, Action and the Embodied Mind," a more technical treatment by Andy Clark. Clark references much of the work by Geoff Hinton, but in the context of generative models. Enjoy!