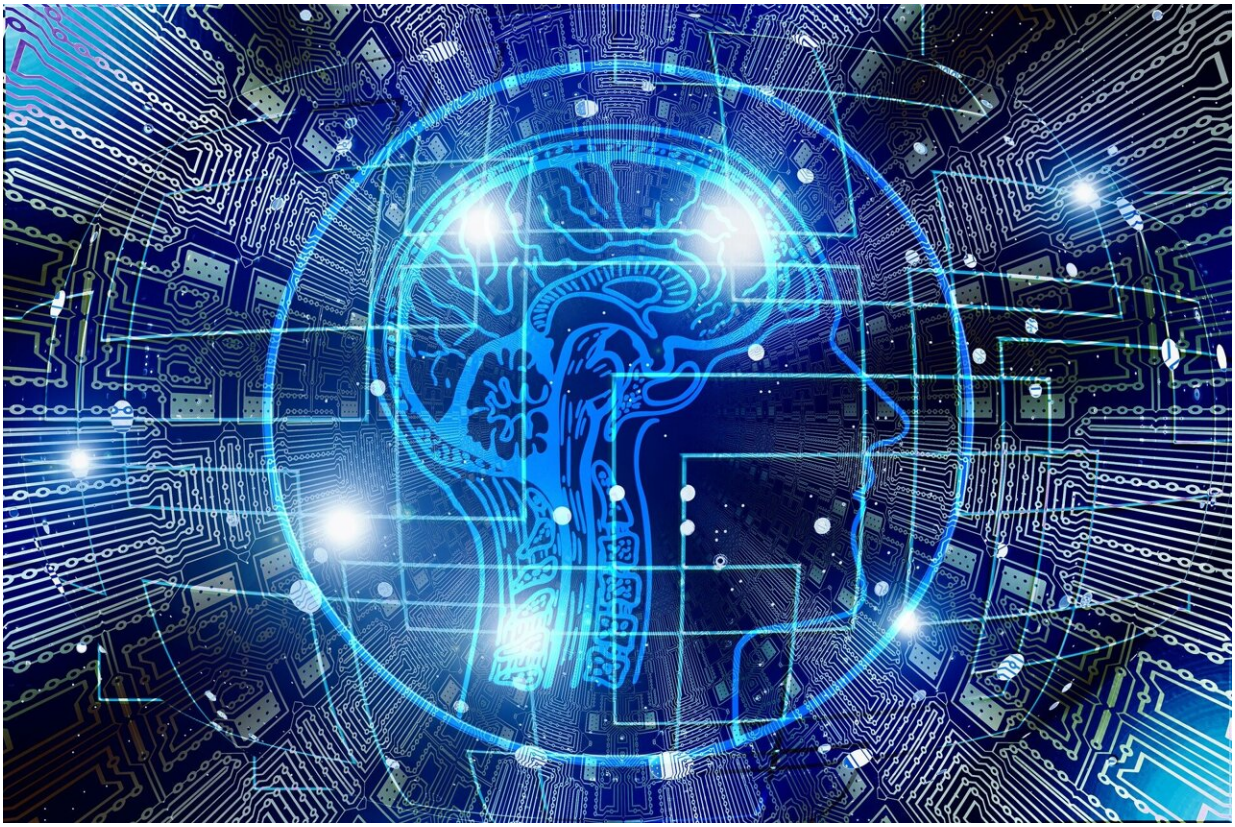
Defining the unexplainable in artificial intelligence

June 26 2023, by David Bradley



Credit: Pixabay/CC0 Public Domain

The term "artificial intelligence," usually abbreviated as AI, means many things to many people. Initially, the phrase was used to allude to the potential of machines, computers, specifically, somehow gaining

sentience on a par with human consciousness. This notion inspired a lot of philosophical debate about what it means to be human and whether or not a machine can have self-awareness. The same notion was at the heart of a lot of science fiction throughout the twentieth century and to the present day, although the idea of entities other than humans having human-like consciousness has been around for millennia.

As we enter a new phase in the development of AI technology, the concepts surrounding what we mean by that term are changing. We now consider neural networks that can be trained genetically to undergo machine learning and to take on certain properties we now refer to as AI.

However, many of these tools, computer algorithms backed by enormous information databases do not come close to displaying consciousness and they many such as the now infamous large language models come close to behaving like a human. When prompted with text, they can produce a seemingly authentic response that is, superficially at least, coincident with the response a human might give to that same prompt.

Of course, these models are only as good as the training they have been given and the algorithms they run to generate their responses. At this point in the history of this kind of AI we are fast approaching the notion of a "black box" AI. A system that given a prompt, generates a response that even the programmers and developers of the system cannot predict. Such systems and their responses reaching the point where they cannot be explained, although this is not to suggest that the system is in any way approaching the sci-fi singularity of self-awareness, emotional behavior, and any kind of concept of right or wrong.

We develop and train the algorithms, ask it to make a prediction, and we take the responses. The problems may well arise when those responses are used to make important decisions across society, in economics and finance, in industry, across healthcare and medical research, in the wider

realm of science, in politics and most worryingly in the military machine.

If the programming and training are unexplainable, then we or machines prompting AI systems for a response may get what turns out to be a very wrong response. If we have given such prompt-response systems control of important systems, then we may come unstuck when a prompt generates an entirely inappropriate response in a healthcare environment, in a factory, or on the world stage.

Fabian Wahler and Michael Neubert, writing in the *International Journal of Teaching and Case Studies*, recognize the importance of defining and understanding AI and where it might take us, sooner, rather than later. They have homed in on a definition of explainable AI that might be used in future work by both practitioners and academics alike. The work seeks to remove the ambiguity of current definitions and to increase trust and reliability in decision making by making black-box systems understandable, interpretable, and transparent to human users.

More information: Fabian Wahler et al, A scientific definition of explainable artificial intelligence for decision making, *International Journal of Teaching and Case Studies* (2023). DOI: 10.1504/IJTCS.2023.131664

Provided by Inderscience