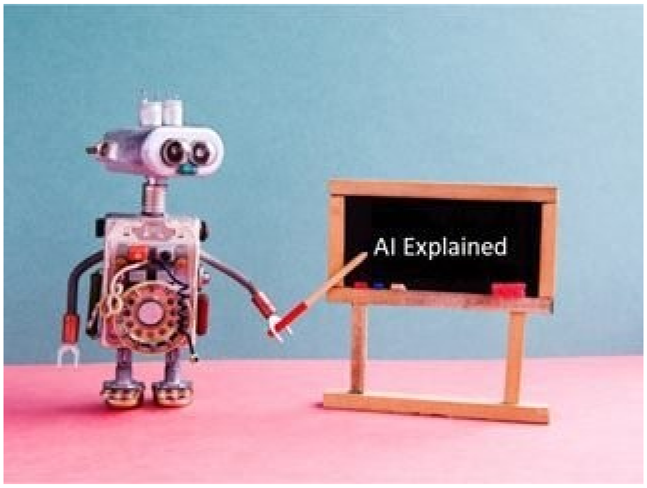

What Is Explainable AI (XAI)?

www.cmswire.com

6 mins read

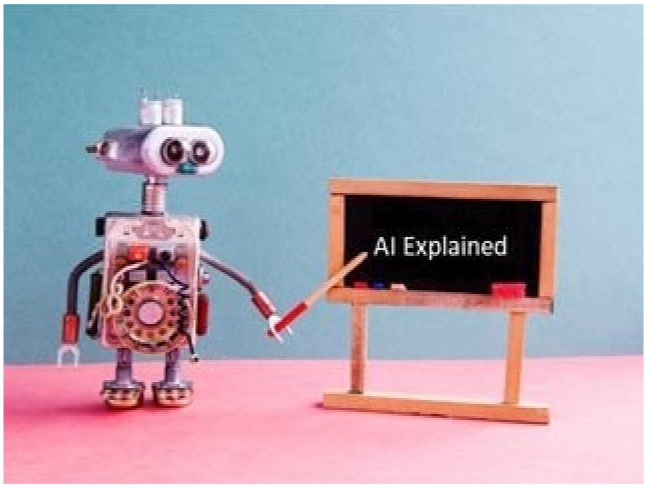

PHOTO: Shutterstock

Artificial intelligence appears to be creeping into every corner of our lives. And it's making some pretty big decisions for us, which begs the question, how is it making these decisions? An answer that organizations find is many times very complex and not well understood. Whether for compliance reasons or to eliminate bias, there is a need to make the decision making capabilities understand. This is where explainable AI {XAI) or Transparent AI come in.

Let's begin by looking at this short list of decisions that AI is making for us in the here and now.

- Whether or not a tumor has become cancerous.
- Whether or not an insurance claim should be processed or denied.
- Whether or not a traveler is approved to go through airport security.
- Whether or not a loan should be made.
- Whether or not a missile launch is authorized.
- Whether or not a self-driving vehicle brakes.

These are complex matters that are well suited to AI’s strength — its ability to process infinitely greater data than a human can, said Mike Abramsky, CEO of RedTeam Global. But the decisions AI can make are also reflective of the technology’s weakness, the so-called "Black Box" problem, Abramsky said. Because deep learning is non transparent, the system simply can’t explain why it got to the decision. No matter how much you respect AI’s advance, though, most of us would also like to know *how* AI came to the conclusions that it did, if only out of curiosity’s sake. So do proponents of a movement called explainable AI, and their reasons for wanting to know go far beyond mere curiosity.

“With AI-powered systems increasingly making decisions such as credit card approval for an application, a self-driving car applying the brakes after getting closer to an obstacle, and parole recommendation for incarcerated felons, it has become vital for humans to understand the decision-making mechanism of the underlying AI to ascertain that the AI makes accurate and fair decisions,” said Abhijit Thatte, VP of Artificial Intelligence at Aricent.

Related Article: 11 Industries Being Disrupted By AI

Understanding What Explainable AI Is

Explainable AI, in short, is a concept in which AI and how it comes to its decisions are made transparent to users. However that is only one definition. Thatte, for example, notes that the simplicity of an explanation is relative. “An electrical engineer may find an explanation of how an electromagnetic field in a household fan works easy to understand,” he said. “However, people without a background in electrical engineering or physics may find the same explanation complicated.” That is why he defines explainable AI as an AI whose decision-making mechanism for a specific problem can be understood by humans who have expertise in making decisions for that specific problem.

Thatte’s tweak, though, only scratches the surface of the varying opinions about what XAI should mean. “The concept is inherently subjective, just like it would be if you asked a person to ‘explain’ his or her decisions,” said Timo Elliott, VP and global innovation evangelist for SAP.

Indeed there are even different opinions on how ‘explainability’ should be defined, wrote Rudina Seseri, founder and managing partner at Glasswing Ventures, in a TechCrunch article earlier this year. “What do we want to know? The algorithms or statistical models used? How learning has changed parameters throughout time? What a model looked like for a certain prediction? A cause-consequence relationship with human-intelligible concepts?”

Despite these differences in what XAI means, there is near universal agreement on *why* it is necessary. “As AI systems are given more responsibilities and more complex tasks, we want to understand when we can trust them,” said Mark Stefik, research fellow at PARC, a Xerox company. “Machine learning systems are not perfect. They fail, sometimes surprisingly and catastrophically, and people want to know why. With XAIs, we can better determine the limits of the machines.”

Related Article: 6 Ways Artificial Intelligence Will Impact the Future Workplace

The Problems With XAI

The current approaches to XAI either hint at the reasoning, by having the AI system highlight parts of the input data (such as an image, keywords in an email or other key data points) that were most significant to the decision, Abramsky explained. Other researchers have recently developed systems where the AI can "point" at evidence it used to answer a question or make a recommendation — explaining itself in plain language, he continued.

But there are drawbacks to these methodologies, according to Abramsky. “Sometimes the explanations are themselves confusing or wrong,” he said. “Plus these approaches oversimplify the explanation for a recommendation which has been reached in a far more complex manner.” In addition, in many scenarios there isn’t time to sift through or analyze the reasoning of a machine learning recommendation, such as when the stock market reacts to an event or when a car is speeding along a road. Also worth noting, said Michael Youngblood, a principal engineer and scientist at PARC, is that explainable AI has been used for years in AI that are based on transparent methods. These include Expert Systems, Production Rule Systems, Symbolic Reasoning Systems-anything that is considered GOFAI (Good Old-Fashioned AI) methods, he said.

Also, Thatte noted that some AI models are more transparent than others, making a complex task even harder. AI models created from machine learning algorithms such as linear regression, logistic regression, Naive Bayes classifier, and decision trees are transparent, he said. “A human can easily interpret the decision-making rules they extract from historical data by studying the values of their parameters.” However, AI models created from machine learning algorithms such as Support Vector Machines, Random Forests, Gradient Boosted Trees, k-Nearest Neighbors and deep learning algorithms such as Artificial Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks are challenging to interpret even for an AI researcher, he said.

AI based on automated forms of traditional statistical approaches are more mature and explainable than the recent advances in neural-network approaches, SAP’s Elliott said. “The majority of AI/machine learning actually used today uses automated forms of traditional statistical approaches like for logistics, fraud and marketing.” These models are easily "explainable" because they show the weights used for different characteristics and people readily accept these models because they seem more transparent, he said.

On the other hand, neural-net-based approaches are used for things like image recognition and self-driving cars. “These models are more opaque, and when things go wrong, like if a self-driving car was involved in an accident, it can be harder to understand why,” according to Elliott.

Related Article: Why the Benefits of Artificial Intelligence Outweigh the Risks

Who is Driving Explainable AI

There are several efforts to create explainable AI, Thatte said. The two most notable are DARPA-XAI and LIME. US Department of Defense's Defense Advanced Research Projects Agency (DARPA) launched the Explainable Artificial Intelligence (XAI) project to develop a software library toolkit for explainable AI, he said. In May 2018, researchers demonstrated initial implementation of their explainable AI systems. A complete evaluation of those systems is expected to conclude in November of this year.

Another tool is LIME (Local Interpretable Model-Agnostic Explanations), which was created by Marco Ribeiro and some colleagues. LIME can highlight the features in an image or a text responsible for an ML model's predictions for that image or the text, Thatte said.

Explainable AI is being used now in certain limited circumstances. Stefik said that there are emerging prototypes from applied research. They can report what is in a photo or describe why an autonomous system did something rather than something else, for example. “However, XAI systems that are robust, natural, and flexible-and know their limits-are mainly in early stages of research,” he said.

Will Explainable AI Succeed?

It is possible there is no perfect solution and may end up taking these decisions from AI systems on faith, RedTeam Global’s Abramsky said.“Today, we drive cars at high speed, take medicines that have been engineered to cure us, fly through the air on planes-all on faith in the technology that was used to develop them.”

Sohrob Kazerounian, senior data scientist at Vectra, for one, has his doubts. “My general feeling is that fully explainable AI systems are basically a pipe dream,” he said. “As models become increasingly complex, it will become correspondingly difficult to determine simple, interpretable rules that describe why some AI system classified/clustered/acted in the way that it did, with respect to any given input.” If these rules were simple and easily understood, he added, then it likely wouldn’t have required a complex learning system in the first place.

Andrew Burt, legal engineer and chief privacy officer at Immuta, for his part, said that XAI is a critical field of thought that will be central to the future of machine learning. We need new ways of understanding how and why technology like AI operates the way it does, he said.

■