THE WALL STREET JOURNAL.

https://blogs.wsj.com/cio/2017/08/10/inside-darpas-push-to-make-artificial-intelligence-explain-itself/

CIO JOURNAL.

Inside Darpa's Push to Make Artificial Intelligence Explain Itself

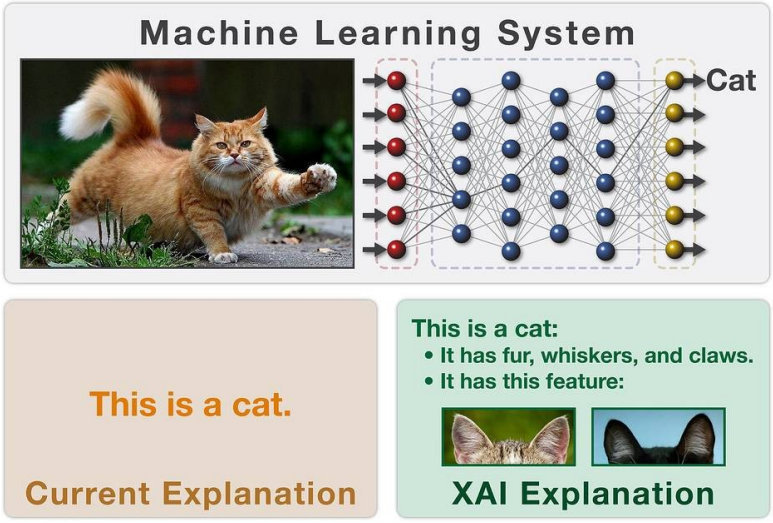
Researchers look to create systems that translate decisions made by algorithms into language humans can understand

By Sara Castellanos and Steven Norton

Aug 10, 2017 12:59 pm ET

ARLINGTON, Va. — The research arm of the U.S. Department of Defense is marshalling an international effort to overcome what many say is the biggest obstacle to widespread adoption of artificial intelligence: teaching algorithms to explain their decision-making to humans.

A group of 100 researchers at more than 30 universities and private institutions is working to build "explainable AI" systems that can translate complex algorithmic-made decisions into language humans can understand. Coordinating the effort is the Defense Advanced Research Projects Agency, known as a catalyst for technology breakthroughs including the internet.



This is a cat, but how do you know? Darpa's explainable AI initiative seeks to develop systems that can translate complex algorithmic decision-making into language humans can understand. Easy-to-grasp explanations could help humans develop more trust in artificial intelligence.
PHOTO: DARPA

Explaining how complex algorithms reach conclusions is critical as artificial intelligence becomes more deeply embedded in everyday life, says David Gunning, the program manager at Darpa overseeing the project. Useful explanations can give humans insight into algorithms' decisionmaking process, building a greater level of trust between humans and machines.

"If it's finding patients that need special attention in the hospital, or wanting to know why your car stopped in the middle of the road, or why your drone turned around and didn't do its mission ... then you really need an explanation," Mr. Gunning said.

Read the CIO Explainer: What Is Artificial Intelligence?

The project also coincides with the European Union's General

Data Protection Regulation, expected to take effect in May 2018. Under the rule, people who are significantly affected by an algorithmic decision have the right to ask for an explanation.

Teams involved in the $75 million Darpa effort, which began work on their projects in May, have spent the first phase working on focus areas of their choosing. They will meet to present their initial findings to Darpa in November.

Darpa's David Gunning PHOTO: DEFENSE ADVANCED RESEARCH PROJECT AGENCY

In phase two, expected to last two-and-a-half years, each institution will tackle one of two "challenge problems" recommended by the Naval Research Lab. The first centers around using AI to classify events in multimedia, while the second involves training a simulated autonomous system to perform a series of missions.

The final product will be a portfolio of new machine learning techniques and user interfaces that government or commercial

groups could use to build their own explainable AI systems.

The Darpa project's mission involves getting to the heart of deep learning, a powerful but relatively little-understood branch of artificial intelligence. Deep learning tools include neural networks, software whose structure roughly tries to mimic the human brain's operations. A neural network, which is composed of layers of interconnected artificial "neurons," learns by being fed large amounts of "training data," which it uses to identify patterns. For instance, by looking at thousands or millions of images, the algorithm learns, by tweaking the connection strengths between its neurons, the features that make a dog a dog and a cat a cat. If it has learned those patterns well, it should be able to look at a new, or "test," image and correctly identify it. If it stumbles, its engineers can give it more data or modify the structure of the neural network.

While these systems can draw conclusions with unprecedented accuracy and speed, it's not always clear how the dense web of computations reach a specific decision.

With such programs, "you basically have to trust a black box," said Dirk Englund, professor of electrical engineering and computer science at the Massachusetts Institute of Technology.

Researchers at Charles River Analytics Inc., alongside Brown University and the University of Massachusetts, are building systems that can explain how AI tools designed to classify people's activities in crowds, for example, flag suspicious people. Rather than just identify a suspect, the system would serve images showing the features that caused the algorithm to classify a person as suspicious, which could be useful if a human saw no obvious reason to flag the person.

"That (neural network) might be right almost all the time, but when it does something surprising or when it looks wrong, I can't ask why it did that," said James Tittle, principal scientist at Charles River Analytics.

Darpa's tech legacy is due in part to its ability to wrangle talent from around the world to tackle unsolved problems. The internet, no less, came out of an original Darpa demonstration in 1969 and Mr. Gunning led a project inside Darpa that eventually became the iPhone assistant Siri. Its current AI research could ultimately be spun off into explainable AI software programs that could be used by large corporations.

If humans can't develop a way to understand how machines reason, deep learning deployments in both military and corporate environments could be slowed or even halted, Mr. Gunning said.

"What I think could really happen is they just don't get implemented, or people won't trust them enough to use them."