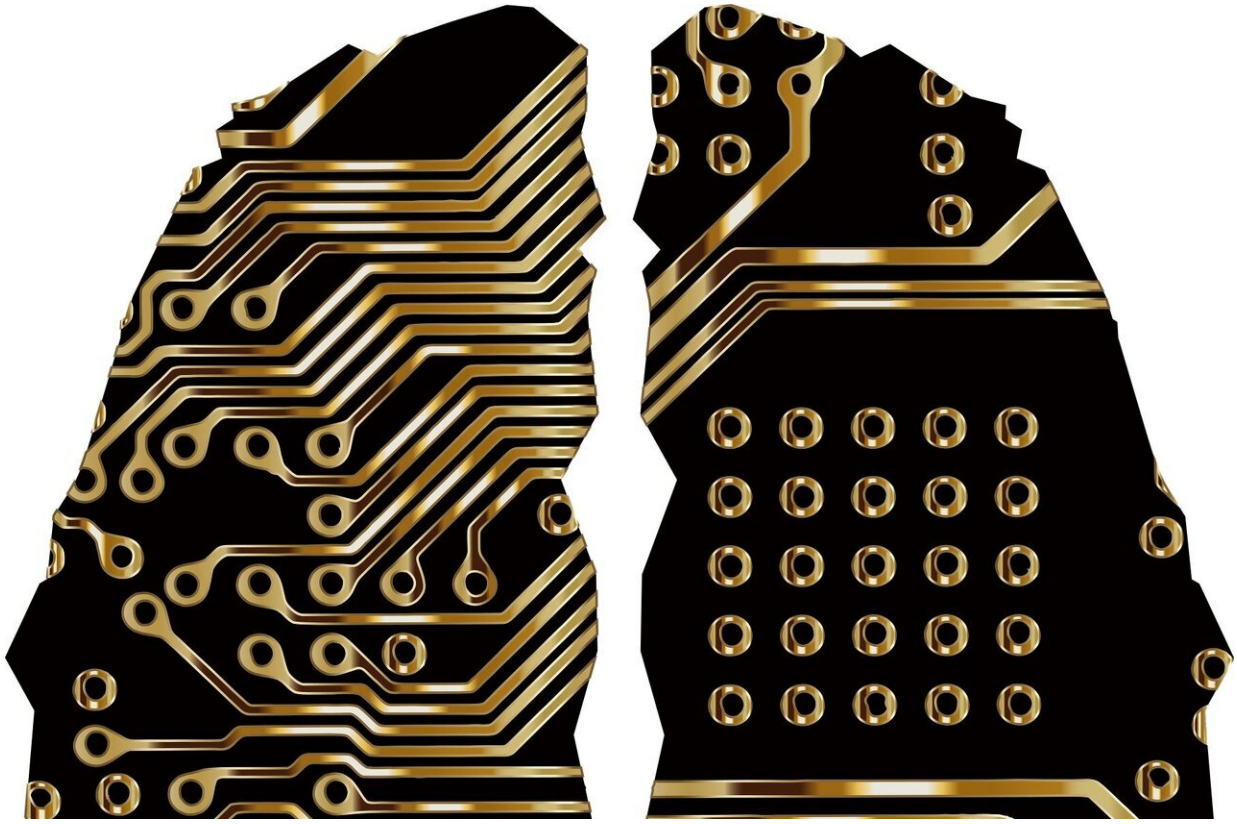
How to make AI trustworthy

August 27 2020



Credit: Pixabay/CC0 Public Domain

One of the biggest impediments to adoption of new technologies is trust in AI.

Now, a new tool developed by USC Viterbi Engineering researchers generates automatic indicators if data and predictions generated by AI

algorithms are trustworthy. Their research paper, "There Is Hope After All: Quantifying Opinion and Trustworthiness in Neural Networks" by Mingxi Cheng, Shahin Nazarian and Paul Bogdan of the USC Cyber Physical Systems Group, was featured in *Frontiers in Artificial Intelligence*.

Neural networks are a type of artificial intelligence that are modeled after the brain and generate predictions. But can the predictions these neural networks generate be trusted? One of the key barriers to adoption of self-driving cars is that the vehicles need to act as independent decision-makers on auto-pilot and quickly decipher and recognize objects on the road—whether an object is a speed bump, an inanimate object, a pet or a child—and make decisions on how to act if another vehicle is swerving towards it. Should the car hit the oncoming vehicle or swerve and hit what the vehicle perceives to be an inanimate object or a child? Can we trust the computer software within the vehicles to make sound decisions within fractions of a second—especially when conflicting information is coming from different sensing modalities such as computer vision from cameras or data from lidar? Knowing which systems to trust and which sensing system is most accurate would be helpful to determine what decisions the autopilot should make.

Lead author Mingxi Cheng was driven to work on this project by this thought: "Even humans can be indecisive in certain decision-making scenarios. In cases involving conflicting information, why can't machines tell us when they don't know?"

A tool the authors created named DeepTrust can quantify the amount of uncertainty," says Paul Bogdan, an associate professor in the Ming Hsieh Department of Electrical and Computer Engineering and corresponding author, and thus, if human intervention is necessary.

Developing this tool took the USC research team almost two years

employing what is known as subjective logic to assess the architecture of the neural networks. On one of their test cases, the polls from the 2016 Presidential election, DeepTrust found that the prediction pointing towards Clinton winning had a greater margin for error.

The other significance of this study is that it provides insights on how to test reliability of AI algorithms that are normally trained on thousands to millions of data points. It would be incredibly time-consuming to check if each one of these data points that inform AI predictions were labeled accurately. Rather, more critical, say the researchers, is that the architecture of these neural network systems has greater accuracy. Bogdan notes that if computer scientists want to maximize accuracy and trust simultaneously, this work could also serve as guidepost as to how much 'noise' can be in testing samples.

The researchers believe this model is the first of its kind. Says Bogdan, "To our knowledge, there is no trust quantification model or tool for deep learning, artificial intelligence and machine learning. This is the first approach and opens new research directions." He adds that this tool has the potential to make "artificial intelligence aware and adaptive."

More information: Mingxi Cheng et al, There Is Hope After All: Quantifying Opinion and Trustworthiness in Neural Networks, *Frontiers in Artificial Intelligence* (2020). DOI: 10.3389/frai.2020.00054

Provided by University of Southern California