Explainable AI

Kelly Carter, PhD. Data Scientist COL(R), CACI
Categories: AI "artificial intelligence", Deep Learning, Machine Learning, Explainability, Dashboards

The DoD must move quickly from black-box Artificial Intelligence (AI) solutions to explainable AI. This presentation will delineate the pitfalls of black-box AI solutions and demonstrate methods to provide explainability in AI and deep learning.

AI that uses deep learning identifies patterns and correlates answers in a way that is not explicitly programed, creating both higher risks and rewards. In 2016, an exposé by ProPublica was done on the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm. This algorithm is used by criminal justice officials to determine recidivism rates. This is one of several algorithms used by California, New York, Wisconsin, and parts of Florida. The COMPAS algorithm was found to have significant racial bias, which was not correlated to actual recidivism rates. Using data from Broward County, Florida ProPublica asserted that COMPAS predictions of violent crimes were correct 20 percent of the time, and that blacks were labeled at higher risk almost twice as much as whites. This study was widely criticized, including by the think tank Community Resources for Justice for misinterpretations of the data and the subject matter. Regardless, the problem of bias and trust in AI outcomes still exists.

The deeper the neural networks, the more layers of analysis, means less insight into how the answer is obtained, adding deep learning compounds the problem. One solution to explainability is a heat map. A heat map shows you where the algorithm focused.

In 2019, it was revealed that an algorithm developed by a Stanford graduate student diagnosed TB correctly 75% of the time, compared to doctors in South Africa, at 62%. By using a heat map, it was revealed that the algorithm included margin data, i.e. non-image information. For example, by considering markings on the x-ray from radiologist, where the x-ray was taken, and what type of machine took the x-ray, the algorithm boosted the score for positive disease find. The x-rays that the algorithm was trained on had markings from radiologist that indicated a positive find for disease – a simple tic mark. Machines located in hospitals were more apt to find disease than mobile doctor's office machines, so images from these machines were scored more likely to be diseased. This is a significant bias in the algorithm and means the actual intended performance of the algorithm for image only data is probably much lower.

Explainable AI can be built. Explainability can be created via dashboards, reports, and constant model updates. In addition to heat maps, methods for AI explainability include: Maximum Mean Discrepancy (MDM) - shows differences in dataset distributions, Partial Dependency Plot for Individual Conditional Explanation (PDP-ICE) - shows the effect a variable on a model, Accumulated Local Effects (ALE) - describes how local features influence the prediction, Interactions-based Method for Explanation (IME) – calculates the contribution of factors. These methods can be applied to black-box AI also. These are just a few examples of ways to build in explainability. Explainability must be expected, demanded, and build-in.