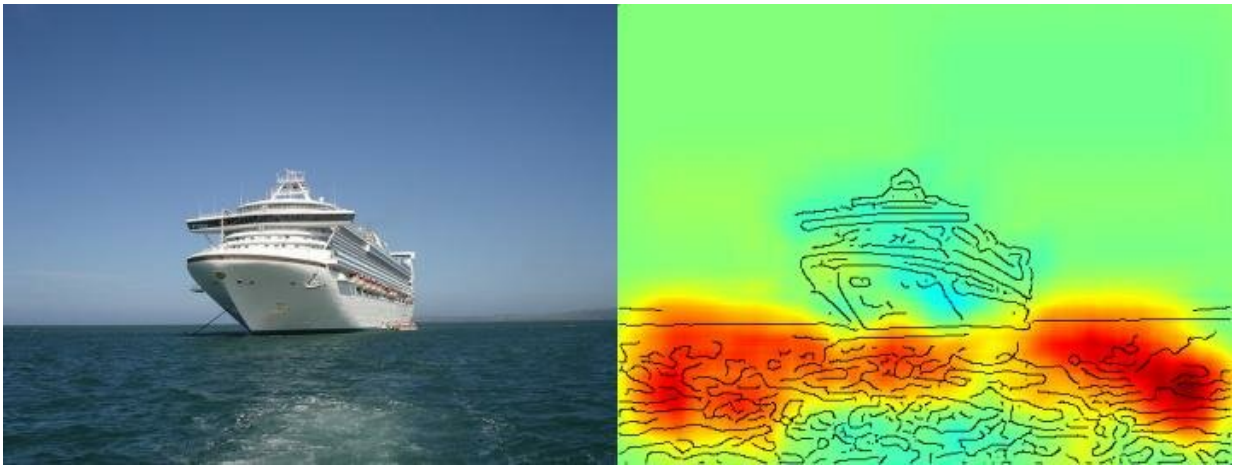
Clarifying how artificial intelligence systems make choices

March 13 2019



The heatmap shows quite clearly that the algorithm makes its ship/not ship decision on the basis of pixels representing water and not on the basis of pixels representing the ship. Credit: *Nature Communications* , CC BY Lizenz

Artificial intelligence (AI) and machine learning architectures such as deep learning have become integral parts of our daily lives—they enable digital speech assistants or translation services, improve medical diagnostics and are an indispensable part of future technologies such as autonomous driving. Based on an ever increasing amount of data and powerful novel computer architectures, learning algorithms seemingly approach human capabilities, sometimes even surpassing them. So far, however, it often remains unknown to users how exactly AI systems

reach their conclusions. Therefore, it may often remain unclear whether the AI's decision-making behavior is truly intelligent or whether the procedures are just averagely successful.

Researchers from TU Berlin, Fraunhofer Heinrich Hertz Institute HHI and Singapore University of Technology and Design (SUTD) have tackled this question and have provided a glimpse into the diverse "intelligence" spectrum observed in current AI systems, specifically analyzing these AI systems with a novel technology that allows automatized analysis and quantification.

The most important prerequisite for this novel technology is a method developed earlier by TU Berlin and Fraunhofer HHI, the so-called Layer-wise Relevance Propagation (LRP) algorithm that allows visualizing according to which input variables AI systems make their decisions. Extending LRP, the novel Spectral relevance analysis (SpRAy) can identify and quantify a wide spectrum of learned decision making behavior. In this manner it has now become possible to detect undesirable decision making even in very large data sets.

This so-called 'explainable AI' has been one of the most important steps towards a practical application of AI, according to Dr. Klaus-Robert Müller, professor for machine learning at TU Berlin. "Specifically in medical diagnosis or in safety-critical systems, no AI systems that employ flaky or even cheating problem solving strategies should be used."

By using their newly developed algorithms, researchers are finally able to put any existing AI system to a test and also derive quantitative information about them: a whole spectrum starting from naive problem solving behavior, to cheating strategies up to highly elaborate "intelligent" strategic solutions is observed.

Dr. Wojciech Samek, group leader at Fraunhofer HHI said: "We were very surprised by the wide range of learned problem-solving strategies. Even modern AI systems have not always found a solution that appears meaningful from a human perspective, but sometimes used so-called Clever Hans Strategies."

Clever Hans was a horse that could supposedly count and was considered a scientific sensation during the 1900s. As it was discovered later, Hans did not master math, but in about 90 percent of the cases, he was able to derive the correct answer from the questioner's reaction.

The team around Klaus-Robert Müller and Wojciech Samek also discovered similar "Clever Hans" strategies in various AI systems. For example, an AI system that won several international image classification competitions a few years ago pursued a strategy that can be considered naïve from a human's point of view. It classified images mainly on the basis of context. Images were assigned to the category "ship" when there was a lot of water in the picture. Other images were classified as "train" if rails were present. Still other pictures were assigned the correct category by their copyright watermark. The real task, namely to detect the concepts of ships or trains, was therefore not solved by this AI system—even if it indeed classified the majority of images correctly.

The researchers were also able to find these types of faulty problem-solving strategies in some of the state-of-the-art AI algorithms, the so-called deep neural networks—algorithms that had been considered immune against such lapses. These networks based their classification decisions in part on artifacts that were created during the preparation of the images and have nothing to do with the actual image content.

"Such AI systems are not useful in practice. Their use in medical diagnostics or in safety-critical areas would even entail enormous

dangers," said Klaus-Robert Müller. "It is quite conceivable that about half of the AI systems currently in use implicitly or explicitly rely on such Clever Hans strategies. It's time to systematically check that so that secure AI systems can be developed."

With their new technology, the researchers also identified AI systems that have unexpectedly learned "smart" strategies. Examples include systems that have learned to play the Atari games Breakout and Pinball. "Here, the AI clearly understood the concept of the game, and found an intelligent way to collect a lot of points in a targeted and low-risk manner. The system sometimes even intervenes in ways that a real player would not," said Wojciech Samek.

"Beyond understanding AI strategies, our work establishes the usability of explainable AI for iterative dataset design, namely for removing artefacts in a dataset which would cause an AI to learn flawed strategies, as well as helping to decide which unlabeled examples need to be annotated and added so that failures of an AI system can be reduced," said SUTD Assistant Professor Alexander Binder.

"Our automated technology is open source and available to all scientists. We see our work as an important first step in making AI systems more robust, explainable and secure in the future, and more will have to follow. This is an essential prerequisite for general use of AI," said Klaus-Robert Müller.

More information: Sebastian Lapuschkin et al, Unmasking Clever Hans predictors and assessing what machines really learn, *Nature Communications* (2019). DOI: 10.1038/s41467-019-08987-4

Provided by Singapore University of Technology and Design