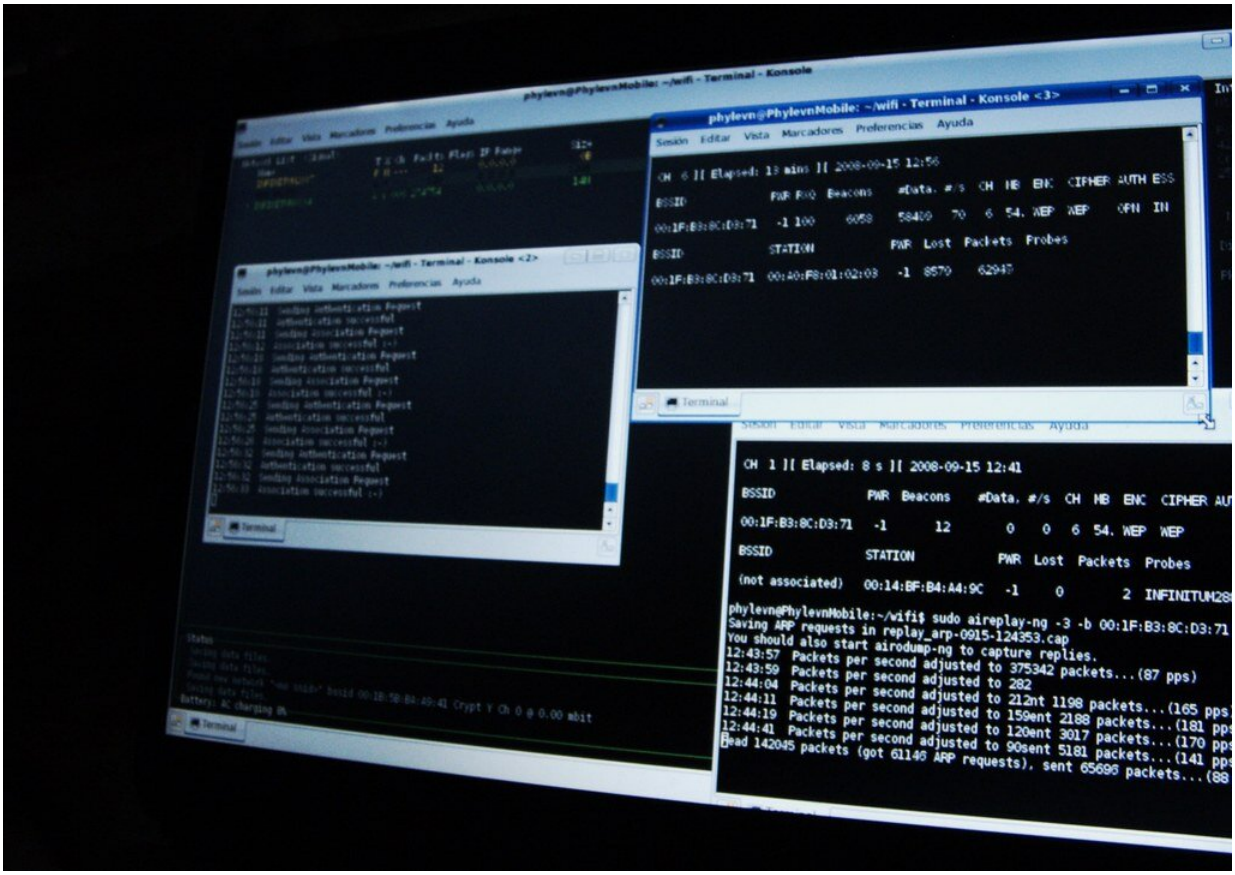
Opening the 'black box' of artificial intelligence

December 1 2020, by Tom Cassauwers, From Horizon Magazine



When decisions are made by artificial intelligence, it can be difficult for the end user to understand the reasoning behind them. Credit: phylevn/Flickr, licenced under CC BY 2.0

Artificial intelligence is growing ever more powerful and entering

people's daily lives, yet often we don't know what goes on inside these systems. Their non-transparency could fuel practical problems, or even racism, which is why researchers increasingly want to open this 'black box' and make AI explainable.

In February of 2013, Eric Loomis was driving around in the small town of La Crosse in Wisconsin, US, when he was stopped by the police. The car he was driving turned out to have been involved in a shooting, and he was arrested. Eventually a court sentenced him to six years in prison.

This might have been an uneventful case, had it not been for a piece of technology that had aided the judge in making the decision. They used COMPAS, an algorithm that determines the risk of a defendant becoming a recidivist. The court inputs a range of data, like the defendant's demographic information, into the system, which yields a score of how likely they are to again commit a crime.

How the algorithm predicts this, however, remains non-transparent. The system, in other words, is a black box—a practice against which Loomis made a 2017 complaint in the US Supreme Court. He claimed COMPAS used gender and racial data to make its decisions, and ranked Afro-Americans as higher recidivism risks. The court eventually rejected his case, claiming the sentence would have been the same even without the algorithm. Yet there have also been a number of revelations which suggest COMPAS doesn't accurately predict recidivism.

Adoption

While algorithmic sentencing systems are already in use in the US, in Europe their adoption has generally been limited. A Dutch AI sentencing system, that judged on private cases like late payments to companies, was for example shut down in 2018 after critical media coverage. Yet AI has entered into other fields across Europe. It is being rolled out to help

European doctors diagnose COVID-19. And start-ups like the British M:QUBE, which uses AI to analyse mortgage applications, are popping up fast.

These systems run historical data through an algorithm, which then comes up with a prediction or course of action. Yet often we don't know how such a system reaches its conclusion. It might work correctly, or it might have a technical error inside of it. It might even reproduce some form of bias, like racism, without the designers even realising it.

This is why researchers want to open this black box, and make AI systems transparent, or 'explainable," a movement that is now picking up steam. The EU White Paper on Artificial Intelligence released earlier this year called for explainable AI, major companies like Google and IBM are funding research into it and GDPR even includes a right to explainability for consumers.

"We are now able to produce AI models that are very efficient in making decisions," said Fosca Giannotti, senior researcher at the Information Science and Technology Institute of the National Research Council in Pisa, Italy. "But often these models are impossible to understand for the end-user, which is why explainable AI is becoming so popular."

Diagnosis

Giannotti leads a research project on explainable AI, called XAI, which wants to make AI systems reveal their internal logic. The project works on automated decision support systems like technology that helps a doctor make a diagnosis or algorithms that recommend to banks whether or not to give someone a loan. They hope to develop the technical methods or even new algorithms that can help make AI explainable.

"Humans still make the final decisions in these systems," said Giannotti.

"But every human that uses these systems should have a clear understanding of the logic behind the suggestion. '

Today, hospitals and doctors increasingly experiment with AI systems to support their decisions, but are often unaware of how the decision was made. AI in this case analyses large amounts of medical data, and yields a percentage of likelihood a patient has a certain disease.

For example, a system might be trained on large amounts of photos of human skin, which in some cases represent symptoms of skin cancer. Based on that data, it predicts whether someone is likely to have skin cancer from new pictures of a skin anomaly. These systems are not general practice yet, but hospitals are increasingly testing them, and integrating them in their daily work.

These systems often use a popular AI method called deep learning, that takes large amounts of small sub-decisions. These are grouped into a network with layers that can range from a few dozen up to hundreds deep, making it particularly hard to see why the system suggested someone has skin cancer, for example, or to identify faulty reasoning.

"Sometimes even the computer scientist who designed the network cannot really understand the logic," said Giannotti.

Natural language

For Senén Barro, professor of computer science and artificial intelligence at the University of Santiago de Compostela in Spain, AI should not only be able to justify its decisions but do so using human language.

"Explainable AI should be able to communicate the outcome naturally to humans, but also the reasoning process that justifies the result," said

Prof. Barro.

He is scientific coordinator of a project called NL4XAI which is training researchers on how to make AI systems explainable, by exploring different sub-areas such as specific techniques to accomplish explainability.

He says that the end result could look similar to a chatbot. "Natural language technology can build conversational agents that convey these interactive explanations to humans," he said.

"Explainable AI should be able to communicate the outcome naturally to humans, but also the reasoning process that justifies the result."

Prof. Senén Barro, University of Santiago de Compostela Spain

Another method to give explanations is for the system to provide a counterfactual. "It might mean that the system gives an example of what someone would need to change to alter the solution," said Giannotti. In the case of a loan-judging algorithm, a counterfactual might show to someone whose loan was denied what the nearest case would be where they would be approved. It might say that someone's salary is too low, but if they earned €1,000 more on a yearly basis, they would be eligible.

White box

Giannotti says there are two main approaches to explainability. One is to start from black box algorithms, which are not capable of explaining their results themselves, and find ways to uncover their inner logic. Researchers can attach another algorithm to this black box system—an 'explanator' – which asks a range of questions of the black box and compares the results with the input it offered. From this process the explanator can reconstruct how the black box system works.

"But another way is just to throw away the black box, and use white box algorithms, ' said Giannotti. These are machine learning systems that are explainable by design, yet often are less powerful than their black box counterparts.

"We cannot yet say which approach is better," cautioned Giannotti. "The choice depends on the data we are working on." When analysing very big amounts of data, like a database filled with high-resolution images, a black box system is often needed because they are more powerful. But for lighter tasks, a white box algorithm might work better.

Finding the right approach to achieving explainability is still a big problem though. Researchers need to find technical measures to see whether an explanation actually explains a black-box system well. "The biggest challenge is on defining new evaluation protocols to validate the goodness and effectiveness of the generated explanation," said Prof. Barro of NL4XAI.

On top of that, the exact definition of explainability is somewhat unclear, and depends on the situation in which it is applied. An AI researcher who writes an algorithm will need a different kind of explanation compared to a doctor who uses a system to make medical diagnoses.

"Human evaluation (of the system's output) is inherently subjective since it depends on the background of the person who interacts with the intelligent machine," said Dr. Jose María Alonso, deputy coordinator of NL4XAI and also a researcher at the University of Santiago de Compostela.

Yet the drive for explainable AI is moving along step by step, which would improve cooperation between humans and machines. "Humans won't be replaced by AI," said Giannotti. "They will be amplified by

computers. But explanation is an important precondition for this cooperation."

Provided by Horizon: The EU Research & Innovation Magazine