Explainable AI: What, How and Why

Youssra Outelli
500756021
Hogeschool van Amsterdam
Intelligent Systems

Table of contents

Introduction

In this report I'll try to answer the question: How does Artificial Intelligence comes to its conclusion? In other words, explainable AI. The report starts by explaining what explainable AI is. Then, how it works? Followed by answering the question why do you want to know how AI comes to its conclusion? And last but not least the progress of developing explainable AI and the current stance.

What?

In this chapter I'll be explaining what explainable AI is.

According to Wikipedia, "explainable AI refers to methods and techniques in the application of artificial intelligence technology such that the results of the solution can be understood by human experts." (wikipedia, 2020).

But what does this mean exactly? AI systems take an input and based on the machine learning model it produces an output. The steps that an AI system takes to get to the conclusion can also be referred to as the black box of an AI system. You can't see what happens in that box and why it works that way. And that's where explainable AI comes into the picture.

Consider Face ID on an iPhone. You scan your face two times from different angles and you're able to unlock your phone. The input of this system are the two scans you've made and the output is that you're able to unlock your phone. But we don't know HOW the system concluded that it's you trying to unlock your phone and not something else. What are the steps the system took? How was it so sure that it was you that it unlocked your phone? There's were explainable AI comes into the picture. By analyzing these choices, you can better understand how the system works and what to do to make it better.

How?

In this chapter I'll talk about how explainable AI works including examples.

Now that I've explained what explainable AI is, it's time to look at how it works. There are multiple theoretical ways to get access to the black box of an AI system. One way is to use machine learning algorithms that are inherently explainable (Schmelzer, 2019). There are numerous machine learning algorithms that show the steps it took by using certain graphs or decision trees. By implementing these algorithms into an AI system this can help give more visibility in how the system came to its conclusion.

Another way to get an inner look at the black box of an AI system is by feature importance or feature attributions. Google makes use of this concept. According to Google feature importance enables you to see which features contributed the most to model training and individual predictions (Google, 2020). These features are being extracted from the test dataset that is provided. By using feature importance, you get a look into the black box because now you know what feature played a big role in getting to that certain conclusion.

Why?

In this chapter I'll be talking about why you would want to know the process of AI.

Now that I've talked about what explainable AI is and how it works, it's time to look at why we need it. As mentioned before the solution an AI gives is calculated in a black box. Nobody knows what's happening and how the AI made its decision.

Now the question is why do we need to know what's happening in the black box? By knowing what's happening in the black box we can ensure trust, clarity and understanding of these applications for everybody involved. When you don't know how the AI makes its decision, there is a possibility the AI would make a decision that would go at the expense of human lives and/or safety. With AI solutions being able to explain their decision making this can be prevented and adjusted when necessary.

Progress

In this last chapter I'll talk about how far the concept of explainable AI currently is and what the plans are for the future.

Google has made tools and frameworks in which you're able to make AI-models that have explainable features. Some of these features are receiving a score explaining how each factor contributed to the final result of the model predictions (Google, 2020). Another feature is their What-If tool. This tool is an interactive visual interface in which you can compare models, visualize inference results and feature attributions and much more (Google PAIR, 2020).

The goal for explainable AI is to implement it in every AI solution. One solution where XAI is set to play a big role in is the world of finance. Especially now that many banks worldwide have a big presence in doing your banking online (Hagras, 2020). But the finance industry isn't the only industry in which XAI will be present in the future. The health industry is also an industry in which XAI can get really big.

Sources

Google. (2020). *Explainable AI - features*. Opgehaald van Google Cloud: https://cloud.google.com/explainable-ai

Google. (2020). *Introduction to AI Explanations for AI Platform*. Opgehaald van Google Cloud: https://cloud.google.com/automl-tables/docs/features#ai-explanations/

Google PAIR. (2020). *What If...* Opgehaald van What-If Tool: https://pair-code.github.io/what-if-tool/index.html#intro

Schmelzer, R. (2019, Juli 23). *Understanding Explainable AI*. Opgehaald van Forbes: https://www.forbes.com/sites/cognitiveworld/2019/07/23/understanding-explainable-ai/#4861ce7c9ef5

wikipedia. (2020). *Explainable artificial intelligence*. Opgehaald van Wikipedia: https://en.wikipedia.org/wiki/Explainable_artificial_intelligence