HOW TO **EXPLAIN** THE DECISIONS MADE BY AN AI SYSTEM?

How can you know if the decision made by an AI system is correct, honest and no error occured in the system? In this brAInfood, we focus on the 'explainability' of AI systems. What does explainability mean? How can AI systems be explained? What obligations do you have as an AI developer towards your users when they ask about the explainability of your AI system? Which domains already focus on explainable AI systems? A lot of information, therefore, this brAInfood contains two pages. Make sure to read both pages.

Knowledge Centre Data & Society (2020). How to explain the decision made by an AI system? brAInfood of the Knowledge Centre Data & Society. Brussels: Knowledge Centre Data & Society.

brAInfood of the Knowledge Centre
Data & Society

What is explainability and why is it important?

Wheter we know it or not, we all come across AI systems at some point in our daily lives. Sometimes an AI system is used to give you a fun, innocent recommendation, in Spotify for example. But sometimes an AI system is used to make decisions that can strongly influence your life, for example wheter you are eligible for a loan or not. If an employee of the bank makes this decision for you, you expect an explanation **why** you were rejected and what you can change to get the loan. If an AI system makes such a decision for you, then how do you (or the bank employee who tells you about it) know if the decision is right, fair and not based on a bugged system? These are questions that '**Explainable AI**' tries to answer.

The explainability of an AI-system says something about (1) **in what capacity** a human can understand how the AI system globally works and/or (2) **how the system came** to an individual prediction. From an explainability perspective the different types of AI systems can be divided into **three categories**.

White box, black box, grey box

A **white box** AI system is a **simple system**, of which the **global working** of the system is by itself **understandable** for humans. This can for example consist of a basic decision tree, or a set of rules which apply to the AI-system. To make a white box system explainable, these rules only need to be translated into **natural language**. From the understanding of the global working of the system, it's relatively easy to determine the reasons for the individual recommendations or decisions of the system. However, the simplicity of a white box system also makes it less accurate and **less able to solve complex problems**. For that reason, white box systems are used much less frequently nowadays.

On the other side of the spectrum are the **black box** AI systems. Black box AI-systems are highly **complex systems** of which the **global working is often hidden**, for example because they can learn patterns themselves. This makes a black box system much more able to accurately solve **complex questions**, but also much less understandable for people.

To make a black box system explainable, it is usually necessary to add an **extra module** to the system that can simulate the black box, for example using a white box model. This will give a **simplified impression of the global working of the system**. Another option is to distract the **set of rules** on which individual conclusions of the system are based that are understandable for humans using a white box module. In some cases it can be useful to ask the black box system in a separate module what (features of the) **input data** were **most important** in determining the individual conclusion.

Finally, a **grey box** is situated in between the white box and the black box on the spectrum. Grey box systems try to combine the **high accuracy of a black box** system and the **simplicity of a white box** system. For example by considering explainability throughout the full development process and building it into the black box model from the beginning (**explainability-by-design**). In this way, there is no separate module needed anymore to simulate the global working or explain individual conclusions.

How can you find out what decisions have been made?

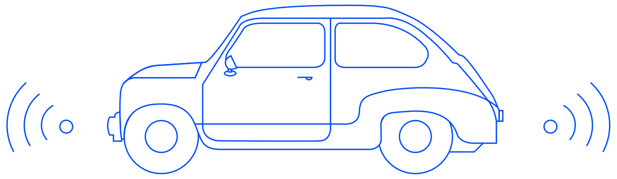
As a user of a service that contains an AI component, it is not always easy to find out why the service presents you with certain recommendations and to check / verify the decisions taken. The use of a white box or grey box can provide an explanation here. Access to these boxes is often not possible because companies do not want to provide insight into the internal workings of their algorithm as they contain, for example, trade secrets or intellectual property of the organization. Even when **access** is granted, the statements are often **worded in a technical way**, linked to specific "AI terminology". As a result, an average user is **not always informed in an appropriate, adjusted way**. In order to provide a (legal) solution for this challenge this element has explicitly been taken into account when drawing up the **GDPR regulations**.

In summary, the GDPR states that an organization that collects and processes personal data **must inform the user** of these practices. The user has the **right to access** his or her personal data and, in the case of automated decision-making, he/she has the right to access meaningful **information about the logic** involved, as well as the **importance** and **intended consequences of the processing**. This information must be communicated "in a concise, transparent, intelligible and easily accessible form, using **clear and plain language**".

Thus, every user has the right to submit a request to access his/her personal data. This data must be provided in an accessible manner, together with information about the processing and analysis that the organization performed on/with this data. An organization that uses AI systems to provide a service must provide the necessary explanations, adapted to the individual user(groups) or stakeholders.

A **user survey** in which different user(groups) and stakeholders are asked about the **expectations** regarding, among other things, "**meaningful information**", "**accessible form**" and "**clear and plain language**" can give a concrete interpretation to these concepts. In this way, organizations are not only an example for the implementation of the GDPR regulations, but they can also increase user confidence and transparency towards them.

In **which areas** is there a strong focus on explainable AI systems?



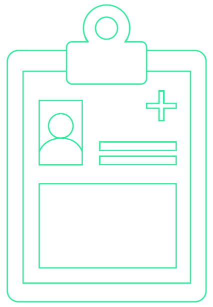
Autonomous vehicles

It has long been predicted that AI will become the 'engine' of future mobility. AI enables the control of autonomous vehicles which, in turn, should ensure a better and more efficient traffic flow and, above all, make traffic much safer. Ensuring safety is also one of the main reasons why autonomous vehicles have not yet made a major breakthrough.

The AI system in autonomous vehicles makes decisions based on the **enormous flow of data about road and driving conditions** that it interprets via the vehicle's **sensors and cameras**. The data is processed by AI algorithms that estimate a situation. Based on this assessment, the AI system makes a decision.

This decision will have to be completely transparent. AI algorithms that you can train through deep learning, for example on classification and pattern recognition, will do what you expect in many scenarios, but not in all cases. When using Deep Learning, you will therefore need to use a gray box model or white box modulel, in order to gain **insight into how decisions are made**.

In a **safety-sensitive matter** such as autonomous vehicles, the outcome must therefore be fully explainable. One must be aware of the moral priorities that have been programmed. A car manufacturer must be able to trace how the model that makes decisions goes from point A to point B. Otherwise, the vehicle will not be licensed.

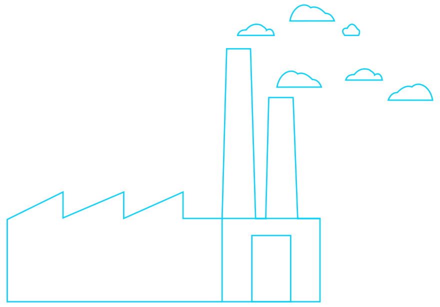


Health care

IBM announced in 2013 an application that would be revolutionary in the fight against cancer. In partnership with the University of Texas MD's Anderson Cancer Center, IBM, through its Watson for Oncology application, would enable doctors to discover valuable **insights from the cancer centre's rich patient and research databases**.

The system processes data from medical literature, treatment guidelines, medical records, laboratory reports, etc. to formulate cancer treatment recommendations. The algorithm then suggests **treatment options** the physician can use to treat the patient. The algorithm's programming includes many different data sources that can be weighed differently.

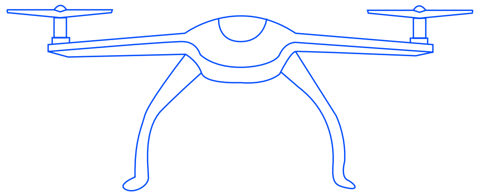
However, in 2017, it was revealed that the system gave erroneous and downright dangerous advice for cancer treatment. The black-box nature of the AI system caused problems. Although the platform indicated which literature it used to draw its conclusion, it was **unclear and inexplicable** how it used the **scoring criteria** to rate some studies against others. As a result of this, the project was put on hold for a certain period of time



Manufacturing

Currently, there are already robots in factories that assist employees in the execution of their job, for example by taking over a part of the welding process. Some robots are even able to work hand in hand with employees. When robots support and assist workers, they are often called '**cobots**'.

Because of the strong cooperation between robots and employees, which will most likely increase in the future, it is necessary to have **good communication**, **trust**, **clarity** and **understanding** between the machine and the human. Explainable AI plays an important role in creating trust, among other things, because the system can explain the underlying model to the employee and explain why it makes certain decisions.



Defence

Smart AI drones can be used as weapons in wars in order to avoid a physical (face-to-face) confrontation with the target. This system can already be applied, but a person still has to make/approve the final decision. If this system would become fully autonomous, the **results** of the system must be **very accurate**. Nobody wants to hit a wrong target. It must be very clear how the system works exactly and how decisions are made. The U.S. Defense Advanced Research Projects Agency is investigating the use of Explainable AI further.