José Esteban Lauzan, Head of Innovation at Iberia and founding member of the Scientific Community and Distinguished Expert, Atos
Amélie Groud, Senior Data Scientist and member of the Scientific Community, Atos

AI Explainability: making the complex comprehensible

AI Explainability is the name given to the approaches, techniques and efforts that aim to make Artificial Intelligence (AI) algorithms explainable to humans.

In the case of some AI algorithms, especially machine learning (ML) ones, the result of an AI solution cannot be understood by a human expert in a particular subject matter and the designer of the solution cannot explain why the AI arrived at that specific result. Lack of explainability raises concerns around safety, ethics, fairness, reliability and ultimately trust in the proposed solution.

AI Explainability is complicated. ML algorithms aim to detect patterns and hence insight from input data, but this process cannot be comprehended by simply listing rules or instructions in human-readable format. The machine learning process also cannot be understood by comparing it to a human learning process. An ML model can integrate thousands of dimensions in its learning process whereas a human being can barely work with more than a handful simultaneously. ML algorithms usually require a large amount of input data whereas humans only need a few examples to start making accurate decisions.

Responsibility for automated decision-making

To make things more complicated, it turns out that we apply different standards to humans and familiar algorithms (such as rule-based ones) than we apply to more innovative algorithms such as ML ones.

Algorithms are not (currently) responsible for their decisions, which means that determining liability in automated decision-making is still an open legal question.

Can technology help with rational decisions?

Humans are known to display bias in judgement and decision-making. Studies have shown that in hiring processes, for example, if photos and certain demographic information are removed from application forms, people often arrive at different selection decisions. Provided they are well designed with ethical considerations built in from the outset, digital applications could in principle vet applications with greater impartiality.

Nevertheless, humans consider themselves explainable because most of us can articulate why we took a particular decision. And there is an incentive to be able to explain – because humans can be held legally responsible for the consequences of their decisions.

By contrast, algorithms are not (yet) responsible for their decisions, which means that determining liability in automated decision-making is still an open legal question. Because of this gap in liability and because many AI algorithms are new to humans and business applications, there is a natural lack of trust in them and a strong desire for AI Explainability.

Making a tangible difference to citizens and society

Achieving AI Explainability requires understanding and insights aligned to both the socio-economic and scientific-technical dimensions.

Societies will probably progressively trust AI algorithms as their use becomes more widespread and as legal frameworks refine the allocation of liabilities. Of course, cultural differences greatly affect how countries and regulatory regions approach AI. In countries such as China, regulation is lax and the political system seemingly places little importance on the freedom of individuals; for example, China is implementing a social credit system, based on algorithms, which aims to provide a standardized assessment of the trustworthiness of its citizens. This context makes Ethical AI and Explainable AI, as we see it in Europe, less applicable. In the US, while the rights of individuals are more important, regulation is also lax (especially for business purposes), so the workability and benefits of AI solutions represent greater value than their explainability. In other regulatory areas like the EU, emphasis is placed both on individual rights and regulation (such as the General Data Protection Regulation), with the result that explainability is often more important than workability – especially in heavily regulated sectors, such as energy and finance.

From a scientific and technical perspective, methods and techniques are being researched and developed with the objective of increasing the interpretability of algorithms. Some of these methods are model-agnostic and can provide meaningful insights from any trained ML models (e.g. Shapley values – a method for interpreting such models which originates in cooperative game theory and assigns payouts to players according to their contribution to the total payout). Other techniques are specific to a given family of algorithms.

Depending on the use-case, client, sector, market, regulation, political environment, culture, etc., using a specific AI-ML model may or may not make sense, be legal, or ethical. We should identify the nuances and highlight the need for case-by-case analysis and decision-making. In the case of healthcare, finance or energy, most scenarios are heavily regulated (which means more explainability is needed). In other scenarios we might favor benefits over explainability – for example, using an AI-ML model that monitors the manufacturing of non-critical parts and flags up when problems are likely to appear in the process.

What all these scenarios have in common is a need for focused consideration of the level of transparency that is required and how it can be achieved.