Responsible Machine Learning

How to decide between explicability or transparency?

Marcin Detyniecki
Head of Research and Development
& Group Chief Data Scientist

The “interpretable” quality of Machine Learning models is one of the most important concepts when it comes to the definition of a responsible AI. It is also one of the necessary conditions in order to gain the consumers’ trust, which is a key element for the deployment of AI solutions at scale. When a model makes a choice that has an impact on a person or a community, it legitimately cannot result from an opaque decision-making mechanism. It must be interpretable. But this interpretability of artificial intelligence actually harbors two distinct notions: transparency and explicability. We will try to define these notions, analyze the way one interacts with the other and understand their limits.

Let's imagine M. Dupont, a customer who applies for a credit. Unfortunately for him, his banker refuses the request. It is natural for M. Dupont to be able to understand why this decision was made, what elements were taken into account, and eventually, what actions he could take to improve his profile and stand a better chance of getting a favorable response on his next application. These questions are just as legitimate – if not more so – if the decision was taken by an artificial intelligence. But how does one account for a decision made by a Machine Learning model? Two notions can help answer this question.

Transparency, first of all, which consists in being able to describe each element and each step of the decision process. In the case of Artificial Intelligence, we will consider elements such as the type of Machine Learning model that is used, the way it was trained, or the type of data that was used during its learning phase. Total transparency can go as far as the disclosure of the type of algorithm used, its specific parameters, its architecture, its production data, and the potential use of advanced techniques, such as deep learning… which would amount to sharing a considerable amount of incomprehensible data without automated processing. An alternative solution, which would consist in sharing the architecture, could be useful to a Machine Learning expert, but not to a neophyte.

Explicability is a different matter. It aims to provide a rationale that is understandable and usable by a given user in a way that is adapted to his intended use. It is most often a simplification of what exactly happens, with the goal of providing an actionable. In the case of Artificial Intelligence, it may for example enable a customer to understand the reasons why a decision has been made, then use this information to take action. If a Machine Learning model has predicted that a claim is potentially fraudulent, explicability is the ability to explain the factors that motivated that prediction and to allow formal verification – that is, to make those factors intelligible, understandable and actionable.

It is clear how these two notions are complementary, but also how they carry different attributes: where transparency is synonymous with completeness but also technicality, explicability focuses on the best way to be understandable and useful to the user.

Context-dependent usefulness

It would be inaccurate to oppose transparency to explicability; however, their degree of usefulness varies depending on context, situations and purpose (for instance, it depends on the person to whom we wish to be

transparent and/or provide explanations to). In most cases, explicability takes precedence. This is the case in the example we used earlier: Mr Dupont does not have the tools to analyze the parameterization of a Machine Learning model. It is much more useful for him to have clear explanations that will allow him to understand how and why a decision concerning him was made.

But in other circumstances, transparency might be more useful. This can be the case, for instance, when it comes to regulators who may have the skills and capacity to process large amounts of technical information and to review and audit complex algorithms. In some cases, transparency can also be of interest from a consumer perspective. It can be as simple as informing the user that the company is using an algorithm as part of its processes. It can also involve answering a customer's questions about the way their data is used by the company. Will they be leveraged down the line in a context that goes beyond the strict use for which they shared it?

Interpretability challenges

While the doctrine that artificial intelligence must be explainable and transparent in order to meet the criteria for responsible AI cannot be questioned, it does pose a number of challenges. How far should the requirements go, and what limits should be set? If transparency is a virtuous concept in itself, when pushed to the extreme, it raises the question of competition and business secrecy. It is also a matter of security, since it can facilitate attacks. It is important to find the right balance between opacity that is harmful to the consumer and the revelation of the mechanisms of the models that enable the company to generate value and be competitive.

The search for explicability also faces obstacles and limits. One of those can be related to the notions of veracity and faithfulness in the explanation of the world. This problem is all the more difficult to resolve since the most efficient Machine Learning models include several billion parameters. Yet it is precisely in the multiplicity of these parameters that lies the interest of these tools, since it allows them to reflect complexity and to solve seemingly insoluble problems – while being able to integrate a large number of particular cases. Simplifying them to make them intelligible entails the risk of distancing them from reality. This is the problem of the trade-off between explicability and precision; a limit that is well known to the Machine Learning research community.

A matter of audience

Another important challenge of explicability is linked to the notion of audience. Ideally, the explanation is not the same depending on the person receiving it and his or her specific needs. The information to be extracted from the model and the explanations to be provided should be different depending on whether one is talking to an expert, a regulator or a client. How to address these different audiences and meet their expectations? How to best adapt the elements provided according to their needs and degree of understanding? This is one of the major challenges to meet when developing a Responsible AI.

Let's take two examples of very different applications that we are currently developing at AXA. The first one, the Claims Analytics Library, is a Machine Learning solution for detecting cases of potential fraud. The second one covers a research program conducted jointly by AXA and the OECD, aimed at explaining and predicting the mechanisms that cause economic crises. It is clear that needs vary in terms of explanation: in the first example, explanations are addressed to agents in order for them to be able to verify, case by case, that a certain claim is indeed fraudulent in nature. The second one is about extracting global knowledge that will enable experienced economists to propose financial measures in order to avoid crises.

Contacts

Marcin Detyniecki

Head of Research and Development & Group Chief Data Scientist

marcin.detyniecki@axa.com

Xavier Renard

Research & Advanced ML Executive Manager

xavier.renard@axa.com

And tomorrow?

"Why was this decision taken?" Tomorrow, it should be possible to provide a clear and well-argued answer to this question, and it will be graded according to the degree of knowledge of the interlocutor. To achieve this, we propose to bypass the famous accuracy-interpretability trade-off. Our idea is to keep a precise but complex AI model, and to add a second model specifically designed to explain the decisions made by the first one. This approach, known as "post-hoc", presents promising results, even though the veracity issue remains on the explanation side – which is less problematic than a bad decision made by an overly simplistic AI. Indeed, the raison d'être of the explanation is to simplify the world to make it intelligible; but how can we be sure that the surrogate – and thus the associated explanation – are faithful to the initial complex model? This is the central question our research teams are working on today.

3 questions to...

David Martens

Professor at the university of Antwerpen and explainable AI specialist

You work on applied interpretability with AXA Belgium through the AXA Research Fund. What do you focus on in this project?

We develop and apply novel algorithms, that explain why advanced prediction models issue a certain decision. We currently focus more specifically on image data: we apply counterfactual methods to explain why an image is classified in a specific class. Suppose you have the image a car and want to predict something about it, such as whether it was involved in an accident. You would then likely need an explanation as to why the model provided a specific prediction. A counterfactual explanation would reveal what part of the image has led to the prediction. The use of counterfactual methods lead to improved trust, but also enable us to make significant progress in explaining misclassification – ie, understanding why a model makes mistakes. Explanations therefore serve another cause: to improve model accuracy, especially with deep learning models. In the long run, this research can help to better assess risks and damages and have a positive impact to customer satisfaction.

What are some of the other concrete applications to interpretability you have been working on?

Interpretability is very useful for risk management. We have been working on fraud risks with the tax administration. One of the main challenges when using complicated models is explaining to agents why they are being sent to a specific company or individual to audit them. Otherwise, auditors might be reluctant to follow the instructions of a black box. Providing an explanation as to why a company might be fraudulent makes the process much easier and increases trust in the system. In another use-case, we worked with the European Central Bank in order to predict how the financial markets would react to a certain announcement. Explaining individual predictions is key in this case since it enables fine-tuning the message to avoid unintended market reactions.

Is the use of interpretability limited to finance?

Surely not. I have worked on an application where the purpose of the machine learning model was to verify the content of websites. Companies usually don't want to display ads on websites with adult content or hate speech. There are so many different websites that there is the need for a system to predict whether any given site is likely to contain such content. The explanations are useful for advertisement companies, but also for the website owners: naturally, they want to know why they were flagged, and, if they disagree with the prediction, what they need to change for the algorithm to change its prediction. Basically any application area where automated decision making has an impact on individuals, explanations are key. For example: explain why a person was rejected for a job interview, explain why an offer was being made, explain why a medical diagnosis was made, and so on. I strongly believe that explicability will become standard practice in the deployment of future machine learning models.