Improving the Explainability and Accountability of Algorithms

Winston Maxwell, Ryadh Benlahrech

IMPROVING THE EXPLAINABILITY AND ACCOUNTABILITY OF ALGORITHMS

Thanks to advances in deep learning, decision support algorithms are emerging in almost all sectors, but many of these artificial intelligence (AI) tools remain "black boxes". A multidisciplinary group of researchers has been working on this problem of opacity in order to remedy it and thus promote public confidence in algorithms.

"Everybody agrees algorithms should be explainable, especially in safety-critical areas such as health and air transport. There's a real consensus," Winston Maxwell says. For instance, the European Commission recently published a white paper on its AI strategy, which emphasizes the explainability, transparency and accountability of algorithmic decisions. In France, the Villani report on AI, published in 2018, also insisted on the need to make algorithms more transparent and understandable. However, this policy objective runs up against numerous implementation problems related to the concept of explainability, such as ethics, the definition of the right level of explanation, the technical characteristics of the AI methods used, the preservation of trade secrets, the additional costs generated by explanations, and the lack of clear legal definitions regarding what explanations mean. Existing legislation already requires explainability, particularly for algorithms used by administrative authorities in France, but the law generally leaves a great deal of leeway, thus adding further difficulties for developers, users and regulators of these tools.

"Our research work has been carried out using a multidisciplinary approach, bringing together data science, applied mathematics, computer science, economics, statistics, law and sociology, so as to provide in-depth thinking with regard to definitions, techniques and the need for explainability, which are integrated into the broader notions of transparency and accountability," says David Bounie, a co-author of the report. In short, explainability may be used, for example, to help users understand how a search engine works, to help investigators find out why an autonomous vehicle crashed, or to detect possible discrimination in loan approval processes.

EXPLAINABILITY DEPENDS ON FOUR CONTEXTUAL FACTORS

To attain their goal of demystifying explainability, the researchers developed an original methodology whose starting point is contextual. They identified four important contextual factors.

– The audience, i.e. the persons targeted by the explanation. The level of the explanation will differ depending on whether it is addressed to a user or a regulator, for example.

– The level of impact of the algorithm. Explaining why an autonomous car crashed is of greater importance to society than explaining the mechanisms behind an advertising or video recommendation algorithm.

– The legal and regulatory framework, which varies according to different geographical areas, for example in Europe with General Data Protection Regulation (GDPR).

– The operational environment surrounding explainability, such as its mandatory status for certain critical applications, the need for certification prior to deployment, or ease of use by users.

"Taking into account the four contextual factors of explainability is essential. For business users and developers, explainability is primarily driven by operational requirements, and those are quite different from the legal requirements," says Winston Maxwell.

Based on the new working paper *Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach,* by Valérie Beaudouin, Isabelle Bloch, David Bounie, Stephan Clemençon, Florence d'Alché-Buc, James Eagan, Winston Maxwell, Pavlo Mozharovskyi and Jayneel Parekh, and on an interview with Winston Maxwell.

Winston Maxwell is director of law and digital technology studies in the economics and social science department of Telecom Paris. Previously, he was a partner at the law firm Hogan Lovells, specializing in data law. He is a graduate of Cornell Law School and obtained a PhD in economics from Telecom Paris (*Smarter Internet Regulation Through Cost-Benefit Analysis,* published by Presses des Mines in 2017). His research work is mainly focused on the regulation of artificial intelligence.

Methodology

In producing their "position paper" on the problems of algorithm explainability and accountability, the researchers discussed the subject with the various stakeholders (political, academic, business) using a multidisciplinary approach (mathematics, computer science, social sciences, etc.). The aim is to determine the state of the art and to identify scientific and policy recommendations for improving the explainability of algorithms.

EXPLAINABILITY'S COSTS AND BENEFITS TO SOCIETY

After the first stage related to the different contexts of explainability, the researchers studied technical explainability solutions for different AI algorithms. In broad terms, they drew up an inventory of the different methods used to render different machine-learning models more transparent. These methods include hybrid AI approaches, which combine the best of several AI techniques including symbolic, knowledge-based, AI. Hybrid approaches are particularly promising, according to Isabelle Bloch, one of the study's co-authors: "These approaches could reduce the gap between algorithm performance and explainability. Ultimately, explainability will be an integral part of performance characteristics."

The researchers have made another major innovation by incorporating a cost-benefit analysis into explainability. In other words, the level of explainability will be driven in part by comparing the costs and benefits of explainability for society. As mentioned above, the explanation of an autonomous car accident and the explanation of a search engine result do not have the same impact on society; the benefits of explanation will be different in each case. The researchers identified several categories of costs related to explainability, in particular costs relating to the storage of data logs in dedicated registers, which will prove essential to permit ex post explainability of decisions. The researchers point out, however, that the GDPR limits the storage of personal data. "The issue of data storage and explainability will necessarily involve political choices, because the GDPR discourages data retention, in particular with regard to biometric and facial recognition data. Thinking on this subject is still in its infancy and regulators will certainly have make their decisions based on the particular application and its impact on society", Winston Maxwell points out.

LOCAL AND GLOBAL EXPLAINABILITY

In addition to contextual factors and cost-benefit analyses, the researchers also observed that the right level of explainability must take into account both global and local considerations. Global explainability involves describing the algorithm as a whole and how it should be used (or not, as the case may be). "It is like a user's manual and a warning notice, which includes the type of data used to train the algorithm, and the situations in which the algorithm should or should not be used. The European Commission has adopted this approach in its white paper on AI strategy," notes Winston Maxwell. As for local explicability, this involves explaining particular algorithmic decisions, such as a loan refusal. "Both these dimensions of explainability are necessary, even if they are aimed at completely different things and depend on the four contextual factors we identified," Winston Maxwell emphasizes. There is no doubt that, in the coming months, the issue of the explainability of AI, and in particular of deep learning algorithms, will become more prominent as the policy debate on AI evolves in Brussels. ●

Key points

- Explainability of an algorithm depends on four contextual factors: who the explanation is addressed to, the impact of the algorithmic application, the legal and regulatory environment, and the operational framework. In addition, the global (general operation of the algorithm) and local (specific decision-making) level of explainability must be taken into account.
- The explainability of an algorithm must be considered in light of the costs and benefits for society. In particular, the storage of data logs for algorithmic decisions will require a political choice, as it is costly and is not always compatible with the GDPR.
- Explainability is generally at odds with the performance of algorithms, because machine learning models are often built with only one performance objective in mind. However, with the development of hybrid AI techniques, explainability will become an integral part of performance parameters.