Title: Right to explanation

URL: https://en.wikipedia.org/wiki/Right_to_explanation

PageID: 54625345

Categories: Category:Accountability, Category:Algorithms, Category:Human rights, Category:Machine learning, Category:Regulation of artificial intelligence

-----

In the regulation of algorithms , particularly artificial intelligence and its subfield of machine learning , a right to explanation (or right to an explanation) is a right to be given an explanation for an output of the algorithm. Such rights primarily refer to individual rights to be given an explanation for decisions that significantly affect an individual, particularly legally or financially. For example, a person who applies for a loan and is denied may ask for an explanation, which could be " Credit bureau X reports that you declared bankruptcy last year; this is the main factor in considering you too likely to default, and thus we will not give you the loan you applied for."

Some such legal rights already exist, while the scope of a general "right to explanation" is a matter of ongoing debate. There have been arguments made that a "social right to explanation" is a crucial foundation for an information society, particularly as the institutions of that society will need to use digital technologies, artificial intelligence, machine learning. [1] In other words, that the related automated decision making systems that use explainability would be more trustworthy and transparent. Without this right, which could be constituted both legally and through professional standards , the public will be left without much recourse to challenge the decisions of automated systems.

Examples

Credit scoring in the United States

Under the Equal Credit Opportunity Act (Regulation B of the Code of Federal Regulations),

Title 12, Chapter X, Part 1002, §1002.9 , creditors are required to notify applicants who are denied credit with specific reasons for the detail. As detailed in §1002.9(b)(2): [2]

(2) Statement of specific reasons. The statement of reasons for adverse action required by paragraph (a)(2)(i) of this section must be specific and indicate the principal reason(s) for the adverse action. Statements that the adverse action was based on the creditor's internal standards or policies or that the applicant, joint applicant, or similar party failed to achieve a qualifying score on the creditor's credit scoring system are insufficient.

The official interpretation of this section details what types of statements are acceptable. Creditors comply with this regulation by providing a list of reasons (generally at most 4, per interpretation of regulations), consisting of a numeric reason code (as identifier) and an associated explanation, identifying the main factors affecting a credit score. [3] An example might be: [4]

European Union

The European Union General Data Protection Regulation (enacted 2016, taking effect 2018) extends the automated decision-making rights in the 1995 Data Protection Directive to provide a legally disputed form of a right to an explanation, stated as such in Recital 71 : "[the data subject should have] the right ... to obtain an explanation of the decision reached". In full:

The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

...

In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.

However, the extent to which the regulations themselves provide a "right to explanation" is heavily debated. [5] [6] [7] There are two main strands of criticism. There are significant legal issues with the right as found in Article 22 — as recitals are not binding, and the right to an explanation is not mentioned in the binding articles of the text, having been removed during the legislative process. [6] In addition, there are significant restrictions on the types of automated decisions that are covered — which must be both "solely" based on automated processing, and have legal or similarly significant effects — which significantly limits the range of automated systems and decisions to which the right would apply. [6] In particular, the right is unlikely to apply in many of the cases of algorithmic controversy that have been picked up in the media. [8]

A second potential source of such a right has been pointed to in Article 15, the "right of access by the data subject". This restates a similar provision from the 1995 Data Protection Directive, allowing the data subject access to "meaningful information about the logic involved" in the same significant, solely automated decision-making, found in Article 22. Yet this too suffers from alleged challenges that relate to the timing of when this right can be drawn upon, as well as practical challenges that mean it may not be binding in many cases of public concern. [6]

Other EU legislative instruments contain explanation rights. The European Union's Artificial Intelligence Act provides in Article 86 a "[r]ight to explanation of individual decision-making" of certain high risk systems which produce significant, adverse effects to an individual's health, safety or fundamental rights. [9] The right provides for "clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken", although only applies to the extent other law does not provide such a right. The Digital Services Act in Article 27, and the Platform to Business Regulation in Article 5, [10] both contain rights to have the main parameters of certain recommender systems to be made clear, although these provisions have been criticised as not matching the way that such systems work. [11] The Platform Work Directive , which provides for regulation of automation in gig economy work as an extension of data protection law, further contains explanation provisions in Article 11, [12] using the specific language of "explanation" in a binding article rather than a recital as is the case in the GDPR. Scholars note that remains uncertainty as to whether these provisions imply sufficiently tailored explanation in practice which will need to be resolved by courts. [13]

France

In France the 2016 Loi pour une République numérique (Digital Republic Act or loi numérique) amends the country's administrative code to introduce a new provision for the explanation of decisions made by public sector bodies about individuals. [14] It notes that where there is "a decision taken on the basis of an algorithmic treatment", the rules that define that treatment and its “principal characteristics” must be communicated to the citizen upon request, where there is not an exclusion (e.g. for national security or defence). These should include the following:

the degree and the mode of contribution of the algorithmic processing to the decision- making;

the data processed and its source;

the treatment parameters, and where appropriate, their weighting, applied to the situation of the person concerned;

the operations carried out by the treatment.

Scholars have noted that this right, while limited to administrative decisions, goes beyond the GDPR right to explicitly apply to decision support rather than decisions "solely" based on automated processing, as well as provides a framework for explaining specific decisions. [14] Indeed, the GDPR automated decision-making rights in the European Union, one of the places a "right to an explanation" has been sought within, find their origins in French law in the late 1970s. [15]

Criticism

Some argue that a "right to explanation" is at best unnecessary, at worst harmful, and threatens to stifle innovation. Specific criticisms include: favoring human decisions over machine decisions, being redundant with existing laws, and focusing on process over outcome. [16]

Authors of study “Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For” Lilian Edwards and Michael Veale argue that a right to explanation is not the solution to harms caused to stakeholders by algorithmic decisions. They also state that the right of explanation in the GDPR is narrowly defined, and is not compatible with how modern machine learning technologies are being developed. With these limitations, defining transparency within the context of algorithmic accountability remains a problem. For example, providing the source code of algorithms may not be sufficient and may create other problems in terms of privacy disclosures and the gaming of technical systems. To mitigate this issue, Edwards and Veale argue that an auditing system could be more effective, to allow auditors to look at the inputs and outputs of a decision process from an external shell, in other words, “explaining black boxes without opening them.” [8]

Similarly, Oxford scholars Bryce Goodman and Seth Flaxman assert that the GDPR creates a ‘right to explanation’, but does not elaborate much beyond that point, stating the limitations in the current GDPR. In regards to this debate, scholars Andrew D Selbst and Julia Powles state that the debate should redirect to discussing whether one uses the phrase ‘right to explanation’ or not, more attention must be paid to the GDPR's express requirements and how they relate to its background goals, and more thought must be given to determining what the legislative text actually means. [17]

More fundamentally, many algorithms used in machine learning are not easily explainable. For example, the output of a deep neural network depends on many layers of computations, connected in a complex way, and no one input or computation may be a dominant factor. The field of Explainable AI seeks to provide better explanations from existing algorithms, and algorithms that are more easily explainable, but it is a young and active field. [18] [19]

Others argue that the difficulties with explainability are due to its overly narrow focus on technical solutions rather than connecting the issue to the wider questions raised by a "social right to explanation." [1]

Suggestions

Edwards and Veale see the right to explanation as providing some grounds for explanations about specific decisions. They discuss two types of algorithmic explanations, model centric explanations and subject-centric explanations (SCEs), which are broadly aligned with explanations about systems or decisions. [8]

SCEs are seen as the best way to provide for some remedy, although with some severe constraints if the data is just too complex. Their proposal is to break down the full model and focus on particular issues through pedagogical explanations to a particular query, “which could be real or could be fictitious or exploratory”. These explanations will necessarily involve trade offs with accuracy to reduce complexity.

With growing interest in explanation of technical decision-making systems in the field of human-computer interaction design, researchers and designers put in efforts to open the black box in terms of mathematically interpretable models as removed from cognitive science and the actual needs of people. Alternative approaches would be to allow users to explore the system's behavior freely through interactive explanations.

One of Edwards and Veale's proposals is to partially remove transparency as a necessary key step towards accountability and redress. They argue that people trying to tackle data protection issues have a desire for an action, not for an explanation. The actual value of an explanation will not be to relieve or redress the emotional or economic damage suffered, but to understand why something happened and helping ensure a mistake doesn't happen again. [8]

On a broader scale, In the study Explainable machine learning in deployment, authors recommend building an explainable framework clearly establishing the desiderata by identifying stakeholder, engaging with stakeholders, and understanding the purpose of the explanation. Alongside, concerns of explainability such as issues on causality, privacy, and performance improvement must be considered into the system. [20]

See also

Algorithmic transparency

Automated decision-making

Explainable artificial intelligence

Regulation of algorithms

References