cerre Centre on Regulation in Europe

ISSUE PAPER

February 2020

Alexandre de Streel
Adrien Bibal
Benoit Frenay
Michael Lognoul

EXPLAINING THE BLACK BOX

WHEN LAW CONTROLS AI

EXPLAINING THE BLACK BOX: WHEN LAW CONTROLS AI

Alexandre de Streel, Adrien Bibal, Benoit Frenay, Michael Lognoul

February 2020

info@cerre.eu
www.cerre.eu

The event, for which this Issue Paper has been prepared, has received the support and/or input of the following CERRE members: Facebook, Microsoft, Ofcom and Vodafone. As provided for in CERRE's by-laws, this Issue Paper has been prepared in strict academic independence. At all times during the development process, the author, the Joint Academic Directors and the Director General remain the sole decision-makers concerning all content in the Paper.

This Issue Paper is partly based on: A. Bibal, M. Lognoul, A. de Streel and B. Frenay, "Legal Requirements on Explainability in Machine Learning", Artificial Intelligence and Law, 2020, forthcoming.

EXPLAINING THE BLACK BOX
WHEN LAW CONTROLS AI

I. INTRODUCTION

The explainability of Artificial Intelligence (AI) algorithms, in particular Machine-Learning (ML) algorithms, has become a major concern for society.[1] Policy makers across the globe are starting to reply to such concern. In Europe, a High-level Expert Group on AI has proposed seven requirements for a trustworthy AI, which are: human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity/non-discrimination/fairness, societal and environmental wellbeing, and accountability.[2] On that basis, the Commission proposed six types of requirement for high risk AI applications in its White Paper on AI: ensuring quality of training data; keeping data and records of the programming of AI systems; information to be proactively provided to various stakeholders (transparency and explainability); ensuring robustness and accuracy; having human oversight; and other specific requirements for certain particular AI applications, such as those used for purposes of remote biometric identification.[3] Thus in both documents, transparency and explainability are considered key. This is why several new obligations, specific to automated systems (and thus, to AI), in particular in data protection rules and consumer protection rules, have been adopted in Europe to enhance the explainability of algorithmic decisions.

This Issue Paper deals with those AI explainability obligations as follows: after this introduction, Section 2 deals with the different meanings of explainability, in particular by confronting the legal and the computer science meanings; Section 3 focuses on the European AI-specific obligations imposing explainability to operators of such systems; Section 4 studies the rationale of those rules and briefly touches upon how those obligations may be implemented by different ML techniques; finally, Section 5 mentions some issues for further discussion.

[1] F. Pasquale, *Black Box Society. The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015.
[2] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines of 8 April 2019 for Trustworthy AI, p. 14.
[3] Communication White Paper of 19 February 2020 on Artificial Intelligence - A European approach to excellence and trust, COM(2020) 65, p. 18.

II. THE DIFFERENT MEANINGS OF EXPLAINABILITY

In law and ethics, there is no precise definition of AI explainability. The High-Level Expert Group on AI set up by the European Commission defines the principle of explainability as follows:

> "Explicability is crucial for building and maintaining users' trust in AI systems. This means *that processes need to be transparent*, the *capabilities and purpose of AI systems openly communicated*, and *decisions – to the extent possible – explainable* to those directly and indirectly affected. Without such information, a decision cannot be duly contested. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'black box' algorithms and require special attention. In those circumstances, *other explicability measures (e.g. traceability, auditability and transparent communication on system* capabilities) may be required, provided that the system as a whole respects fundamental rights. The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate"[4]

This conception illustrates the different meanings of explainability which may relate to the process, the capabilities or the decision and shows that, for some black box algorithms, auditability of the algorithm may serve as an alternative to explanation to users.

The High-Level Expert Group on AI also notes that:

> "Explainability concerns the ability to explain both the *technical processes* of an AI system and the related human *decisions* (e.g. application areas of a system). Technical explainability requires that the decisions made by an AI system can be understood and traced by human beings. Moreover, *trade-offs* might have to be made between enhancing a system's explainability (which may reduce its accuracy) or increasing its accuracy (at the cost of explainability). Whenever an AI system has a *significant impact on people's* lives, it should be possible to demand a suitable explanation of the AI system's decision-making process. Such *explanation should be timely and adapted* to the expertise of the stakeholder concerned (e.g. layperson, regulator or researcher). In addition, explanations of the degree to which an AI system influences and shapes the organisational decision-making process, design choices of the system, and the rationale for deploying it, should be available (hence ensuring business model transparency)."[5]

This statement underlines a number of important issues: first, explainability may be imposed at different levels such as the overall process or the final decision; second, at the current stage of the technology, there is often a trade-off between accuracy and explainability; third, explainability obligations should be risk-based and depend on the impacts that the algorithmic decision has on users' life; fourth, explainability should be adapted to the technical understanding of the users.

The UK Data Protection Authority, The Information Commissioner's Office (ICO), recently opened a public consultation on draft guidance on explaining decisions made with AI. This draft identifies six main types of explanation:

[4] Ethics Guidelines for Trustworthy AI on Trustworthy AI, p. 13 (our underlining).
[5] *Ibidem*, p. 18 (our underlining).

- *Rationale explanation*: the reasons that led to a decision, delivered in an accessible and non-technical way;

- *Responsibility explanation*: who is involved in the development, management and implementation of an AI system, and who to contact for a human review of a decision;

- *Data explanation*: what data has been used in a particular decision and how; what data has been used to train and test the AI model and how;

- *Fairness explanation*: steps taken across the design and implementation of an AI system to ensure that the decisions it supports are generally unbiased and fair, and whether or not an individual has been treated equitably;

- *Safety and performance explanation*: steps taken across the design and implementation of an AI system to maximise the accuracy, reliability, security and robustness of its decisions and behaviours;

- *Impact explanation*: the impact that the use of an AI system and its decisions has or may have on an individual, and on wider society.[6]

In the computer sciences literature, two types of models are generally described:[7]

- *Interpretable models* are models that are understandable either because their mathematical expressions are easy to understand (as it is the case with linear models) or can be represented in an easily understandable manner (as it is the case with decision trees);

- *Black-box models* are models that are not easy to understand because their mathematical expression is neither straightforward nor easily representable in an understandable manner. For those models, understanding can be improved through explanations by using methods which are external to the models such as visualisation or approximation with interpretable models[8].

III. EUROPEAN LEGAL OBLIGATIONS ON EXPLAINABILITY

Europe (through the European Union and the Council of Europe) already has several rules imposing some forms of explainability when decisions are made with the help of an automated system – and thus, with the help of AI tools:

- Some of those rules are specific to algorithmic decisions and have been adopted recently. They are mainly contained in personal data protection law, consumer protection law or B2B fairness rules;

[6] ICO Draft Guidance of 2 December 2019 on Explaining decisions made with AI, Part 1, p.19.
[7] See A. Bibal and B. Frenay, "Interpretability of machine learning models and representations: an introduction", in *Proceedings of ESANN*, 77:82, 2016.

[8] B. Mittelstadt, C. Russell, and S. Wachter, Explaining explanations in AI, in *Proceedings of the conference on fairness, accountability, and transparency (FAT*), 2019, 279-288.

- Other transparency rules, which are often older, are not specific to automated systems but carry some obligations for algorithmic decisions just as they do for other decisions. These are mainly contained in consumer protection rules for private decisions and in constitutional and administrative law for public decisions.

Another important distinction is between the rules which are horizontal and apply to all sectors of the economy, and the rules which are vertical and apply only to specific sectors of the economy. AI-specific rules are starting to emerge in some sectors where AI is increasingly used and raises specific or important risks for its users.

Table 1: Typology of the explainability rules applicable to AI

	Horizontal legislation	Sector-specific legislation
AI-specific obligations	- Personal data protection: GDPR, Convention 108+ - Consumer acquis: Consumer Rights Directive - P2B Regulation	- Finance - Health - Automotive …
General obligations	Consumer acquis	

3.1. General Data Protection Regulation

The main explainability obligations derive from data protection law.[9] They apply when the decisions (i) involve the processing of personal data, (ii) are based solely on an automated processing of data and (iii) produce legal or significant effects on the recipient of the decision, whatever the field of activity in which those decisions occur. In this case, the processor of personal data has to give certain information to recipients of decisions. Part of this information relates to explainability and is defined as:

> "the existence of automated decision-making, including profiling (…) and, at least in those cases, meaningful information about the *logic involved*, as well as the significance and the envisaged consequences of such processing for the data subject"[10].

This information must at least be given to the data subject before any automated decision is made[11], but may also be required by the data subject at any time later on[12]. In addition, the processor of personal data should allow *ex post* contestation and:

> "implement suitable measures in order for recipients of automated decisions to be able to express their point of view and to contest the decision".[13]

[9] Regulation 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46 (General Data Protection Regulation), OJ [2016] L 199/1.
[10] GDPR, art. 13(2f), 14(2g) and 15(1h), our underlining.
[11] GDPR, art. 13(2f), and 14(2g).
[12] GDPR, art. 15(1h).
[13] GDPR, art. 22(3).

Those articles of the GDPR do not explicitly require the data processor to provide an explanation of decisions made, but they may be interpreted as imposing such an obligation.[14] This interpretation is confirmed by a recital of the GDPR which provides for:

> "(…) 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 (…)".[15]

However, the type of explanation to be given by the processor of personal data is not clear. In its interpretative guidance on the meaningful information to be given, the European Data Protection Board, which groups the Data Protection Authorities of the Member States, notes that:

> The growth and complexity of machine-learning can make it challenging to understand how an automated decision-making process or profiling works. The controller should find simple ways to tell the data subject about *the rationale behind, or the criteria relied on in reaching the decision*. The GDPR requires the controller to provide meaningful information about the *logic involved*, not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm. The information provided should, however, be sufficiently comprehensive for the data subject to understand the reasons for the decision."[16]

This interpretation leaves uncertainty as to the type and content of explanations to be given by data processors, as the "rationale behind the decision" and the "criteria relied upon" are not the same and imply different technical solutions, as described later. In addition, the level of precision of the explanation that should be given to data subjects is not specified.

3.2. Consumer protection and Platform-to-Business Regulation

Along with personal data protection law, explainability obligations are also derived from economic law, such as consumer protection in B2C relationships or its extension in some B2B relationships. The recently revised Consumer Rights Directive imposes new and additional specific information requirements for contracts concluded on online marketplaces whose functioning is mainly based on ML algorithms. When ranking different offers, the provider of the online market place should give the consumer:

> "general information, made available in a specific section of the online interface that is directly and easily accessible from the page where the offers are presented, on the *main parameters* determining ranking (…) of offers presented to the consumer as a result of the search query and the *relative importance* of those parameters".[17]

The Consumer Rights Directive further specifies that:

---

[14] L. Edwards and M. Veale, "Enslaving the algorithm: From a 'right to an explanation' to a 'right to better decisions'?", *IEEE Security & Privacy* 16(3), 2018, 46-54; G. Malgieri and G. Comandé, "Why a right to legibility of automated decision making exists in the general data protection regulation", *International Data Privacy Law*, 7(4), 2017, 243-265.

[15] GDPR, recital 71 (our underlining).

[16] Guidelines of the European Data Protection Board of 3 October 2017 on Automated individual decision-making and Profiling as revised on 6 February 2018, WP251rev.01, p. 25 (our underlining).

[17] Directive 2011/83 of the European Parliament and of the Council of 25 October 2011 on consumer rights, OJ [2011] L 304/64, new art.6a(1a) inserted by Directive 2019/2161 (our underlining).

> "Traders enabling consumers to search for goods and services, such as travel, accommodation and leisure activities, offered by different traders or by consumers should inform consumers about the default main parameters determining the ranking of offers presented to the consumer as a result of the search query and their relative importance as opposed to other parameters. That information should be succinct and made easily, prominently and directly available. Parameters determining the ranking mean *any general criteria, processes, specific signals incorporated into algorithms or other adjustment or demotion mechanisms used* in connection with the ranking."[18]

As the business users of online intermediation apps or online search engines may suffer from the same imbalance of power as the consumers, a new specific Platform-to-Business Regulation imposes a quasi-identical transparency obligation on the providers of online intermediation services and search engines in B2B relationships. The providers of online intermediation services should

> "set out in their terms and conditions the *main parameters* determining ranking and the reasons for the *relative importance* of those main parameters as opposed to other parameters".
>
> The descriptions (…) shall be sufficient to enable the business users or corporate website users to obtain an adequate understanding of whether, and if so how and to what extent, the ranking mechanism takes account of the following: (a) the characteristics of the goods and services offered to consumers through the online intermediation services or the online search engine; (b) the relevance of those characteristics for those consumers (…)"[19]

Similarly, the providers of online search engines have to:

> "set out the *main parameters*, which individually or collectively are most significant in determining ranking and the *relative importance* of those main parameters, by providing an easily and publicly available description, drafted in plain and intelligible language, on the online search engines of those providers (…)"[20]

To facilitate compliance, the Commission has to adopt guidelines explaining the transparency requirements but the Platform-to-Business Regulation already clarifies that:

> "(…) Providers should therefore outline the main parameters determining ranking beforehand, in order to improve predictability for business users, to allow them to better understand the functioning of the ranking mechanism and to enable them to compare the ranking practices of various providers. The specific design of this transparency obligation is important for business users as it implies the identification of a limited set of parameters that are most relevant out of a possibly much larger number of parameters that have some impact on ranking. This reasoned description should help business users to improve the presentation of their goods and services, or some inherent characteristics of those goods or services. The notion of main parameter should be understood to refer to any *general*

[18] Directive 2019/2161 of the European Parliament and of the Council of 27 November 2019 as regards better enforcement and modernisation of EU consumer protection rules, OJ [2019] L 328/7, recital 22.
[19] Regulation 2019/1150 of the European Parliament and of the Council of 20 June 2019 on promoting fairness and transparency for business users of online intermediation services, OJ [2019] L 186/55, art.5(1) and (5), our underlining.
[20] Regulation 2019/1150 on promoting fairness and transparency for business users of online intermediation services, art.5(2), our underlining.

> *criteria, processes, specific signals incorporated into algorithms or other adjustment or demotion mechanisms used* in connection with the ranking".[21]

In addition, the Platform-to-Business Regulation places limits to this explanation obligation as the providers of online intermediation services and online search engines are not:

> "(…) require to disclose algorithms or any information that, with reasonable certainty, would result in the enabling of deception of consumers or consumer harm through the manipulation of search results. This Article shall be without prejudice to Directive 2016/943".[22]

3.3. Financial regulation

Some legal rules are designed for particular sectors in order to have more detailed norms tailored to the needs and characteristics of each sector. In the financial sector, an investment firm that engages in algorithmic trading should notify such type of trading to the financial regulator so that the authority:

> "(…) may require the investment firm to provide, on a regular or ad-hoc basis, *a description of the nature of its algorithmic trading strategies, details of the trading parameters* or limits to which the system is subject, the key compliance and risk controls that it has in place [...] and details of the testing of its systems. The competent authority [...] may, at any time, request further information from an investment firm about its algorithmic trading and the systems used for that trading."[23]

Moreover, when an investment firm engages in a high-frequency algorithmic trading technique, it should:

> "store in an approved form accurate and time sequenced records of all its placed orders, including cancellations of orders, executed orders and quotations on trading venues and make them available to the competent authority upon request".[24]

3.4. Law of Council of Europe

In May 2018, the Committee of Ministers of the Council Europe decided to modernise the Convention for the protection of individuals with regard to the processing of personal data (Convention 108+). The Convention now includes a right for data subjects:

> "to obtain, on request, knowledge of the *reasoning underlying data processing* where the results of such processing are applied to him or her".[25]

The Explanatory Report of the Convention 108+ clarifies that:

---

[21] Regulation 2019/1150 on promoting fairness and transparency for business users of online intermediation services, recital 24 (our underlining).
[22] Regulation on promoting fairness and transparency for business users of online intermediation services, art.5(5). Refer to Directive 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJ [2016] L 157/1.
[23] Directive 2014/65 on Markets in financial Instruments, art.17(2), sub-para.2.
[24] Directive 2014/65 on Markets in financial Instruments, art.17(2), sub-para.5.
[25] Convention 108+, art.9(1c), our underlining.

"Data subjects should be entitled to know the reasoning underlying the processing of data, including the consequences of such a reasoning, which led to any resulting conclusions, in particular in cases involving the use of algorithms for automated decision-making including profiling. For instance, in the case of credit scoring, they should be entitled to know the *logic underpinning the processing* of their data and resulting in a "yes" or "no" decision, and not simply information on the decision itself. Having an understanding of these elements contributes to the effective exercise of other essential safeguards such as the right to object and the right to complain to a competent authority".[26]

IV. RATIONALE OF THE RULES AND THEIR IMPLEMENTATION IN MACHINE LEARNING MODELS

4.1. Why AI-specific obligations?

The rules detailed in section 3 fulfil several objectives with respect to explainability.

First, the addressees of an AI-based decision need to understand the rationale behind it. They will be able to either contest the decision or to take any further action that they feel necessary. For example, in the case of credit scoring, a bank needs to explain the reason of a denial to its client. The client can therefore either contest the denial if prohibited criteria are used, contact another bank which is more in line with his profile (e.g. because the first bank is particularly strict on the loan-to-value ratio), look for a cheaper real estate or simply pay off debts. In B2C or B2B relationships, customers also contest decisions before a Court or simply terminate a business partnership if necessary.

Second, the public authority must be allowed to exercise effective control on the legality of a private decision which is contested.[27] This is directly related to the first objective, as a client will not be able to effectively contest a decision if the judge cannot determine whether it is based on prohibited criteria.

Third, explainability obligations and requirements act as incentive for decision makers to rely on criteria that are not prohibited, since decisions are easier to contest. In other words, explainability increases the effectiveness of the whole legal system. This of course requires that the second objective (effective control by the public authority) is achieved.

4.2. Implementation in ML models

Table 2 organises the explainability obligations reviewed in section 3 with a four-level taxonomy (from weakest to strongest obligations). Level 1 requires to indicate the *main features* used for a decision and their *relative importance*. Level 2 requires to provide all features involved in the decision (not only dominant ones). Level 3 requires to explain how those features are combined to reach a decision. Level 4 requires a complete knowledge of the full model. Interestingly, some of the legal obligations refer to specific

[26] Explanatory Report of Convention 108+, para.77.
[27] Communication White Paper on Artificial Intelligence, p.14; Explanatory Report of Convention 108+, para.77.

decisions, whereas others target the model in itself.[28] Some obligations may be difficult to interpret because they use an inaccurate vocabulary. For example, the interpretative guidelines of the European Data Protection Board on the GDPR refer to the "rationale behind" or the "criteria relied on in reaching the decision" which are technically different.

Table 2: Summary of the legal obligations on XAI, adapted with permission.[29]

Main features	- Directive 2011/83 on Consumer Rights, art. 6(a): obligation to provide "the *main parameters*" and "the *relative importance* of those parameters" - Regulation 2019/1150 on promoting fairness and transparency of online intermediation services, art. 5: obligation to provide "the *main parameters*" and "the *relative importance* of those parameters"
All features	- Guidelines on Automated individual decision-making and Profiling: obligation to provide "the *criteria* relied on in reaching the *decision*"
Combination of features	- Guidelines on Automated individual decision-making and Profiling: obligation to provide "the rationale behind the decision"
Whole model	- Directive 2014/65 on Markets in Financial Instruments, art. 17: obligation to provide "information [...] about its algorithmic trading and the systems used for that trading"

When confronting each type of explainability with the current ML models, it can be shown that the weakest form of explainability can be achieved by most ML models (linear but also black-box models), while the strongest form can only be achieved by some interpretable ML models.[30] Thus, the trade-off between explainability and accuracy is mainly present for the strongest form of explainability.

If, as suggested by the Commission White Paper on AI, information provision is imposed for high risk AI systems,[31] the main question is which form of explainability will be imposed for those systems and, as a consequence, which types of ML models will be *de facto* excluded from use.

[28] See also, S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the general data protection regulation", *International Data Privacy Law*, 7(2), 2017, pp. 76-99.
[29] A. Bibal, M. Lognoul, A. de Streel and B. Frenay, "Legal Requirements on Explainability in Machine Learning", *Artificial Intelligence and Law*, 2020, forthcoming, Section 3.
[30] As explained in A. Bibal, M. Lognoul, A. de Streel and B. Frenay, *supra*.
[31] Communication White Paper on Artificial Intelligence, p.20.

V. ISSUES FOR FURTHER DISCUSSION

This issue paper has shown that there are *many dimensions of transparency and explainability* in law and in computer science:

- One dimension is whether an explanation needs to be given or whether some form of auditability is enough;

- Another dimension is the beneficiaries of the explanation obligations: users, developers, enforcers;

- An additional dimension is the type of explanation that may be given on the main features used to make a decision, all the processed features, a comprehensive explanation of the decision or an understandable representation of the whole model;

The paper has also shown that there are already *various/several rules and obligations* in the current EU legal framework on Explainable AI:

- Some rules, which are old, are not AI-specific and are mostly related to transparency obligations in consumer protection law for private decisions;

- Other rules, which are more recent, are specific to automated systems (which includes AI tools) and mostly related to transparency or explainability obligations in consumer protection and B2B fairness rules.

However, the technical implications of those rules for ML models are not clear yet. Therefore, it is key to clarify the meaning of those rules in close dialogue with the computer science community and industry before adopting new rules or principles.

This paper has also shown that rules are often *risk-based* and that the scope of the explanation to be provided is linked to the impacts of the decision to be adopted with an AI algorithm. This is a good governance principle that should be applied in the implementation of existing rules and the development of any possible new rules as the Commission proposed in its White Paper on AI.

As regards this risk-based approach, any implementation of existing rules or development of new rules should always apply a proportionality principle. Proportionality is a general principle of EU law and also a principle which combines the protection of users with the stimulation of innovation.

Finally, this paper has shown that behind any explainability obligations, a *series of trade-offs* should be decided by the legislator when making the rules and by the enforcers when implementing the rules: between the level of accuracy of a model and its possibilities of explanation, between the rights of users and the rights of the AI owners (notably on IP rights). Legislators and enforcers should be mindful of those trade-offs and find the best possible balance.

cerre

Centre on Regulation in Europe

Avenue Louise, 475 (box 10)
1050 Brussels, Belgium

+32 2 230 83 60

info@cerre.eu

cerre.eu

@CERRE_ThinkTank