Artificial Intelligence and Law (2021) 29:149–169
https://doi.org/10.1007/s10506-020-09270-4

ORIGINAL RESEARCH

Legal requirements on explainability in machine learning

Adrien Bibal[1] · **Michael Lognoul**[2] · **Alexandre de Streel**[2] · **Benoît Frénay**[1]

Published online: 30 July 2020

Abstract
Deep learning and other black-box models are becoming more and more popular today. Despite their high performance, they may not be accepted ethically or legally because of their lack of explainability. This paper presents the increasing number of legal requirements on machine learning model interpretability and explainability in the context of private and public decision making. It then explains how those legal requirements can be implemented into machine-learning models and concludes with a call for more inter-disciplinary research on explainability.

Keywords Interpretability · Explainability · Machine learning · Law

1 Introduction

As deep learning and other highly accurate black-box models develop, the social demand or legal requirements for interpretability and explainability of machine learning models are becoming more significant (Pasquale 2015; Doshi-Velez and Kortz 2017). Interpretability can be defined as the ability for a model to be understood by its users (Kodratoff 1994). For instance, decision trees with a small number of nodes can be considered interpretable, while support vector machines and neural networks are often considered as black boxes. However, despite these

✉ Adrien Bibal
adrien.bibal@unamur.be

Michael Lognoul
michael.lognoul@unamur.be

Alexandre de Streel
alexandre.destreel@unamur.be

Benoît Frénay
benoit.frenay@unamur.be

[1] PReCISE - Faculty of Computer Science - NADI, University of Namur, rue Grandgagnage 21, 5000 Namur, Belgium

[2] CRIDS - Faculty of Law - NADI, University of Namur, Rempart de la Vierge 5, 5000 Namur, Belgium

intuitions, interpretability has yet to be defined formally in the literature (Bibal and Frénay 2016; Lipton 2016). Such a definition is hard to provide as it may depend, among other things on semantics and the level of expertise of the model's users. Furthermore, in machine learning, interpretability and explainability have often been used as synonyms of each other (Bibal and Frénay 2016). Nowadays, the two terms are beginning to have different meanings (Guidotti et al. 2018), with interpretability describing the fact that the model is understandable by its nature (e.g. decision trees) and explainability corresponding to the capacity of a black-box model to be explained using external resources (e.g. visualizations).

In law and ethics, the definitions are not precise either. The European Commission notes that: "explainability of the algorithmic decision-making process, adapted to the persons involved, should be provided to the extent possible [...] In addition, explanations of the degree to which an AI system influences and shapes the organizational decision-making process, design choices of the system, as well as the rationale for deploying it, should be available, hence ensuring not just data and system transparency, but also business model transparency" (Communication from the Commission of 8 April 2019, Building Trust in Human-Centric Artificial Intelligence, COM(2019) 168).

This review paper aims at clarifying the meaning of explainability in law and studies how the legal requirements on explainability could be interpreted and applied in machine learning. This means finding how the concept of explainability that is discussed in legal texts can be translated in machine learning solutions. This also means presenting how the machine learning literature implements the technical solutions derived from this translation. Section 2 reviews the main legal requirements on explainability of machine learning models and decisions. Section 3 presents the possible translation of explainability from the legal to the machine learning literature, as well as the machine learning challenges that emerge from the legal requirements. Finally, Sect. 4 concludes by discussing the conceptual difference between the legal requirements on explainability and their technical implementation and by proposing future directions in machine learning related to these challenges.

2 Legal requirements on explainability

Explainability obligations depend on who makes the decisions, and on the degree of automation of the decision-making process. Indeed, requirements are stronger for public authorities than for private firms. They are also stronger when the decision-making process is completely automated (i.e., when no humans are in the loop). As the desired technical outcomes of legal requirements are not always clearly understandable from the legal texts, they need to be clarified on the basis of their objectives. In that perspective, this section analyses in turn explainability obligations that exist in private decision-making (Sect. 2.1), in public decision-making (Sect. 2.2), and the reasons for such requirements (Sect. 2.3).

2.1 Weaker explainability requirements in B2C and B2B

While public decision-making, in administration and justice, always needs an explanation (see Sect. 2.2 below), private decision-making in Business-to-Consumers (B2C) and Business-to-Business (B2B) relationships only needs explanation when a specific law requires it. This section considers two different types of laws that can impose explanation obligations on private companies, namely horizontal (transversal) and vertical (sectoral) rules. The first ones apply to all sectors of the economy, while the second ones only apply to specific sectors providing more detailed rules to better take into account their characteristics.

2.1.1 Horizontal rules and explainability requirements

The main explainability obligations come from data protection law (in the European Union, the General Data Protection Regulation 2016/679, GDPR). 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. For instance, an automatic refusal of an online credit application is subject to such obligations (art. 22(1) and recital 71 of the GDPR).

In this case, the processors of personal data have the obligation to give certain information to recipients of decisions. One type of information relates to explainability and is defined as "meaningful information about the logic involved, as well as [...] the envisaged consequences of such processing for the data subject" (art. 13 (2f) and 14 (2g) of the GDPR). This information must be given to the data subjects at the time of the collection of personal data, before any automated decision is made. The same information may also be required by data subjects at any time, before and/or after such a decision is made (art. 15 (1h) of the GDPR). In addition, processors of personal data should implement suitable measures in order for recipients of automated decisions to be able to express their point of view and to contest the decision *ex post*, after the decision is made and communicated to its recipient (art. 22(3) of the GDPR).

Those articles of the GDPR do not explicitly require data processors to provide the explanation of decisions made, but the different obligations imposed on processors of personal data can be interpreted as imposing such explanation. This interpretation is confirmed by the recital 71 of the GDPR which provides for the right to obtain an explanation of fully-automated decision in order to be able to challenge the decision. The existence of an explanation requirement in the GDPR is still debated among legal scholars as explainability is only specifically mentioned in a non-binding recital and not in a binding article of law. The majority of scholars support this requirement of explanation (Goodman and Flaxman 2016; Malgieri and Comandé 2017; Edwards and Veale 2018; Selbst and Powles 2017). They argue that Recital 71 should be used to complement and explain the binding requirements of the Regulation, on the basis of a systemic interpretation of the text (i.e. a type of interpretation of legal texts that focuses on the law as a whole, given its context and objectives). However, a minority of scholars reject the existence of a right to explanation for

the following reasons (Wachter et al. 2017). They argue that the European Parliament wanted the requirement for data controllers to explain their automated decisions inside the binding part of the text (Article 22), but this was finally not agreed during the political negotiations leading to the adoption of the GDPR. Hence, they argue, the term explanation was voluntarily placed within the non-binding Recital 71 by the European legislator. Thus, the final interpretation will have to be given by the Court of Justice of the European Union at some point in the future.

The type of explanation to be given by the processors of personal data is not clear either. In their interpretative guidance on the meaningful information to be given, the data protection authorities in Europe note that the processors "should find simple ways to tell the data subject about the rationale behind or the criteria relied on in reaching the decision," but not "a complex explanation of the algorithms used or the disclosure of the full algorithm. The information provided should, however, be sufficiently comprehensive [...] to understand the reasons for the decision" (Guidelines of the European Data Protection Board of 3 October 2017 on Automated individual decision-making and Profiling, p. 25). This interpretation leaves uncertainty on 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. In addition, the level of detail of the explanation that should be given to data subjects is not specified.

Next to data protection law, explainability obligations in B2C relationships may also derive from consumer protection law. As consumers are in a situation of weakness and lack bargaining power in their relations with businesses, several rules protect them from unfair practices. In the European Union, a reform of consumer protection law, adopted in 2019, imposes on online marketplaces an obligation to provide "the main parameters determining ranking [...] of offers presented to the consumer as result of the search query and the relative importance of those parameters as opposed to other parameters" (new art. 6(a) of Directive 2011/83 on Consumer Rights). The reform clarifies that "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" (recital 22 of Directive 2019/2161 on better enforcement and modernization of EU consumer protection rules).

In parallel, the European Union has adopted very similar obligations for online intermediation services and search engines to the benefit of their business users (B2B). Indeed, the business users of such services are in a situation of weakness that can be compared to the one of consumers, in their relations to this type of service providers. Providers of online intermediation services have to "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". 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" (art. 5 of Regulation 2019/1150 on promoting fairness and transparency for business users of online

intermediation services). The Regulation clarifies that "the notion of main parameter should be understood to refer to any general criteria, processes, specific signals incorporated into algorithms or other adjustment or demotion mechanisms used in connection with the ranking" (recital 24 of Regulation 2019/1150).

2.1.2 Sectoral rules and explainability requirements

Some legal rules are designed for particular sectors and contain more detailed norms tailored to the needs and characteristics of each sector. For instance, this is the case for the financial and insurance sectors. Regarding the trading of financial instruments, the investment firm that engages in algorithmic trading should notify it 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." 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" (art. 17(2) of the Directive 2014/65 on Markets in financial Instruments).

Regarding the provision of insurance services to consumers, the Belgian law states that insurance providers must inform their subscribers, in an individual and understandable way, of the segmentation criteria used to determine a tariff and the extent of the guarantee. The insurers also have to inform their customers of the criteria that might have an impact on the future of the insurance policy. Furthermore, in the case of a proposal for a modification of the tariff or of the extent of the guarantee, due to a modification of the risk that an insured person represents, the insurer has to motivate his proposal on the basis of the data and criteria used to assess the modification of the risk (art. 46 of the Belgian law of 4 April 2014 on insurances).

2.2 Stronger explainability requirements in G2C

When decisions are adopted by public authorities such as administrations and judges in Government-to-Citizens relationships (G2C), providing explanations on those decisions is always compulsory, and the legal obligations for explainability are stronger than in B2C. In law, this type of requirement is called 'motivation'. Among public authorities, the obligations are stronger for judges than for administrations. This subsection analyses the requirements of motivation for administrative decisions and, then, for judicial decisions.

2.2.1 Administrative decisions and explainability requirements

Administrative decisions must comply with a principle of formal motivation (Wiener 1969), requiring that all factual and legal grounds on which the decision is based should be mentioned and explained. The motivation has to be clear, precise and reflect the real motives behind a decision (e.g. the Belgian law of 29 July 1991 on the formal motivation of administrative decisions). This requirement is imposed at the European Union level by the Charter of Fundamental Rights, which states that "every person has the right to have his or her affairs handled impartially, fairly and within a reasonable time by the institutions, bodies, offices and agencies of the Union. This right includes: [...] the obligation of the administration to give reasons for its decisions"(art. 41 of the Charter of Fundamental Rights of the European Union).

The intensity of the motivation depends on the level of discretionary power enjoyed by the administrative authority (Autin 2011). If an administrative decision is made on the basis of objective conditions, the required motivation is weaker, so that the administration only has to explain in its decision that the conditions required by the applicable legal text are fulfilled. An example could be the award of a university degree. If all the credits of the curriculum are passed by a student, the university can limit its motivation to that finding to give the degree. When administrative bodies have more discretionary power, they have to motivate more their choices and legal reasoning. For example, staff selection requires more precise and specific motivation. Another example of more extensive motivation requirements for administrative decisions could be the award of contract after a public tender. Among the various proposals submitted by applicants, the administrative authority has to choose one, and explain precisely why it chooses that one over another one. In this regard, European law provides that public contracting authorities should inform each candidate and tenderer of decisions reached concerning the conclusion of the public procurement including the grounds for any decision. On request from the candidate or tenderer concerned, the contracting authority should "inform: (a) any unsuccessful candidate of the reasons for the rejection of its request to participate, (b) any unsuccessful tenderer of the reasons for the rejection of its tender, [...] (c) any tenderer that has made an admissible tender of the characteristics and relative advantages of the tender selected as well as the name of the successful tenderer [...], (d) any tenderer that has made an admissible tender of the conduct and progress of negotiations and dialogue with tenderers" (art. 55 of the Directive 2014/24 on public procurement).

When the administrative decision-making process is automated, additional explainability requirements may apply. One of the most comprehensive set of rules is in the French law which provides that "the administration gives to the person subject to the individual decision adopted on the basis of an algorithmic process, upon request of such person, in a intelligible manner and without prejudice of any trade secret protected by law, the following information: (1) the degree and the manner to which the algorithmic process contributed to the decision-making, (2) the data processed and their sources, (3) the parameters used for the process and, where appropriate, their weighting, applied to the individual case, (4) the operations carried

out by the processing" (art. R. 311-3-1-2 of the French Code on the relationships between the public and the administration).

An example of such an automated administrative decision-making process is the French software Parcoursup that determines which studies students should start, on the basis of their background, results in high school, available places in the chosen fields of studies, etc. When this software produces outputs for students, the French Code on the relationships between the public and the administration explained above should apply in principle. However, there is a specific derogation for Parcoursup, in order to protect the secrecy of the deliberations of the selecting teams. This derogation limits the information to be given to recipients of the decisions to the administrative documents used to make the decision, and forbids the disclosure of the weighting of parameters used to make the decisions, as well as the disclosure of the operations carried out by the processing (art. L. 612-3 of the French Code on education).

2.2.2 Judicial decisions and explainability requirements

Judicial decisions must also comply with the principle of motivation. This obligation is imposed by several laws, in particular the European Convention on Human Rights. The European Court of Human Rights decided in various cases that: "in accordance with Article 6(1) of the Convention, judgments of courts and tribunals should adequately state the reasons on which they are based" (Cases Salov v. Ukraine, request no 65518/01, 6 September 2005, Sect. 89; Boldea v. Romania, request no 19997/02, 15 February 2007, Sect. 23; Gradinar v. Moldova, request no 7170/02, 8 April 2008, Sect. 107). In addition, the European countries have similar obligations in their Constitutions (e.g. in Belgium, art. 149 of the Constitution, and art. 79 of the Code on judicial proceedings).

The judicial motivation requirement is more stringent than the one applicable to administrative decisions (Alonso 2012). Judges have to explain all the factual and legal grounds on which their decisions are based, but they also have to answer all the arguments made by the parties during the trial. As judges need to interpret and apply the relevant laws to given cases, they need to strongly motivate how they make a specific legal decision, and why they retain the various arguments of the parties supporting their claims. However, the level of detail required for the answers of judges to the arguments of the parties is dependent on the circumstances of the case (European Court of Human Rights, Garcia Ruiz v. Spain, request no30544/96, 21 January 1999, Sect. 26). If a judgment is produced by machine learning tools, the same rules apply in relation to the motivation of the judgment, as these rules do not focus on whom (i.e. a judge or a machine) makes the decision but only on the fact that a judgment is made.

2.3 Why legal requirements on explainability?

Previous sections showed that European and national laws already contain several obligations on explainability and, given the ethical importance of the issue, those rules may

be strengthened in the future (Commission White Paper on AI, COM(2020)65, p. 20). Some rules apply generally, to all types of decision-making, while other rules, often stricter, apply specifically to automated decision-making. Stricter rules apply to automated decision because, as precised in the Commission White Paper on AI (p. 11), errors and biases may have much larger effects in AI decision making than in human decision making. Moreover, it seems that many humans trust less AI systems than other humans. Both types of rules are often general and imprecise. This means that clarifications will have to be given by the enforcers of the rules, and ultimately by the judges, in case of conflict on the meaning and implications of a particular explainability obligation. To decide on the interpretation of an unspecified legal rule, enforcers and judges rely on legal texts, but also on the goals pursued by the rules. Legal obligations on explainability pursue in general two main objectives. The first one benefits the recipients of the decisions, while the second one benefits the public enforcers or the judges.

The first objective of explainability rules is to allow the recipients of a decision to understand its rationale and to act accordingly (Alonso 2012). Indeed, it is very difficult, if not impossible, to react to a decision when the reasoning and process that led to the outcome are unknown. In B2C or in B2B relationships, customers can act by changing providers and/or by contesting the decision before a Court when they think the decision is based on illegal grounds. For instance, if a customer seeks credit from her bank and such credit is denied, the applicant needs to receive meaningful explanation of the denial (e.g. the income is not sufficient). On that basis, the customer can decide to go to another bank relying on other (and more favourable) criteria or to contest the negative decision before courts if it was based on prohibited selection criteria (such as race or gender in some cases). In G2C relationships, the recipient of the decision cannot "vote with their feet" and change administration or judge when dissatisfied with the criteria used by the public authority, but can always contest the legality of the decision before a superior judge.

The second objective of explainability is to allow the public authority, before which a private or a public decision is contested, to exercise a meaningful effective control on the legality of the decision (Commission White Paper on AI, p. 14). Going back to the previous example of credit denial, the judge has to know the criteria on which the refusal was based to determine whether prohibited criteria were used to refuse the credit. In addition, even if a specific decision is not contested, more transparency and explainability increase the incentives of decision makers not to rely on illegal criteria as it would be more difficult (but not impossible) for them to hide the use of such illegal criteria, and hence easier to condemn them if they were using them. Reflecting the traditional view that "sunlight is the best disinfectant", transparency and explainability increase the effectiveness of the whole legal system by facilitating the identification of its violation.

3 Impact of the legal requirements on ML explainability

The legal requirements on explainability explained in Sect. 2 raise several challenges in machine learning at varying degrees. This Section shows how the legal requirements on explainability can be expressed in machine learning terms. It also shows how difficult it may be to comply with these legal requirements.

In order to introduce the technical understanding of the requirements imposed by the law, Sect. 3.1 presents some background on machine learning as well as some technical vocabulary. Section 3.2 proposes a technical interpretation of legal requirements in B2C and B2B. Those requirements relate to the weaker requirements of Sect. 2.1. Finally, some technical solutions that can be provided to the stronger requirements on explainability that are encountered in administrative and judicial decisions (G2C) of Sect. 2.2 are presented in Sect. 3.3.

3.1 Background on machine learning

This section introduces some background (and associated terms) needed to understand the impact of legal explainability on machine learning. Figure 1 presents a typical machine learning pipeline. As this paper is concerned with decisions, the pipeline is focused on supervised learning. It starts with data (Fig. 1(1)), also called a *dataset*) that are generally gathered by experts. These data contains two parts: (i) the targets to predict, which can be a continuous variable (e.g. the amount of a fine to be paid) or a categorical variable (e.g. guilty or not), and (ii) a set of instances (e.g. persons) characterized by features. The data are provided to a training algorithm (Fig. 1(2)) that optimizes the mathematical parameters of a model (i.e. the mathematical expression that is learned to make decisions, see Fig. 1(3)) given the data at hand. When the model is trained, it can be used with instances that have not been used for the training phase (called *unseen* instances) to predict the unknown target value of these instances (Fig. 1(4)). When a set of predictions have been made, performance measures are run on the result to assess the quality of the model (Fig. 1(5)). In the context of category prediction (a task called *classification*), a typical performance measure is the accuracy, which corresponds to the amount of correct predictions over all predictions that have been made by the model.

Despite the fact that regulating any module of the model production process affects the learned model, the notion of explainability studied here relates to explanations that can be provided on the model and its decisions. More precisely, two kinds of models can be described: interpretable models and black-box models. Interpretable models are models that are understandable either because they have a simple mathematical expression (e.g. linear models) or because their representation allows users to understand their mathematical expression (e.g. decision trees). On the contrary, black-box models are models with a complex mathematical expression that, moreover, do not possess a representation that can ease their understanding (Bibal and Frénay 2016). In the context of black-box models, which are not interpretable by definition, the way to improve understanding is through explanations.

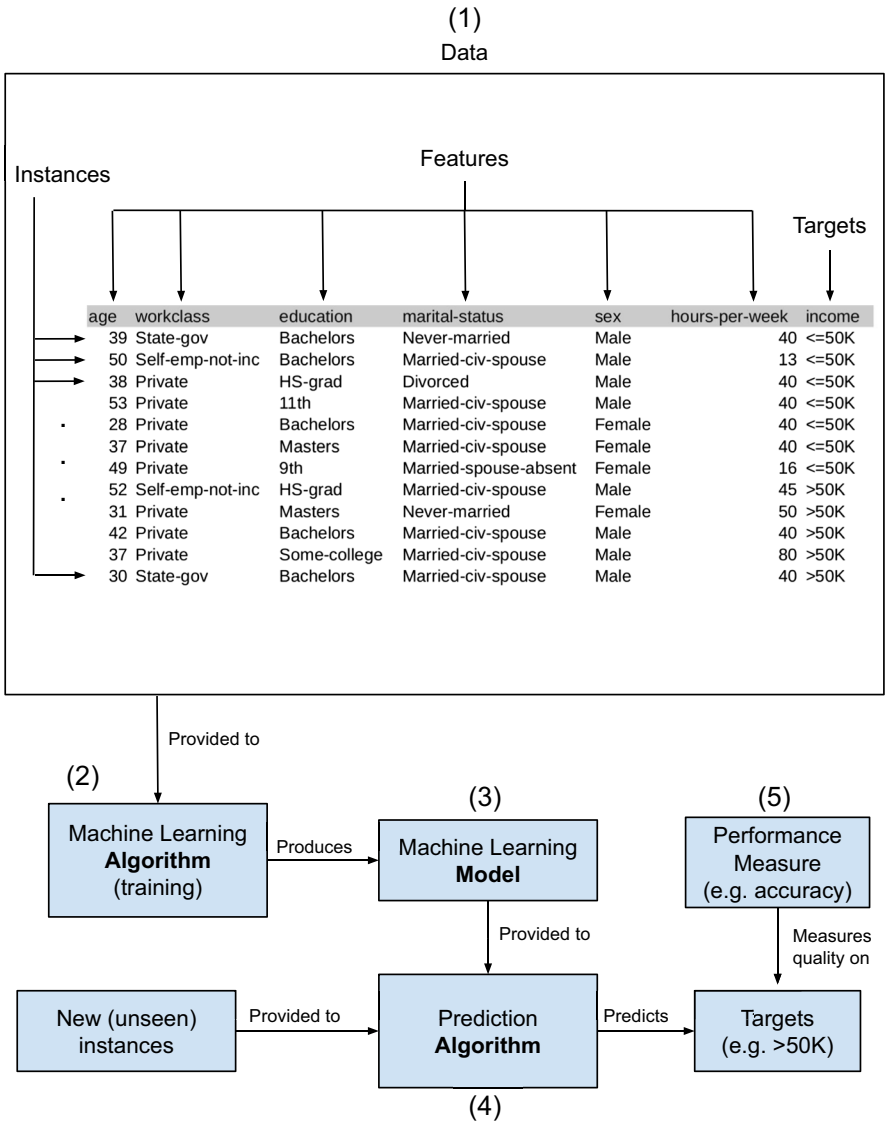

Fig. 1 The user classically provides structured data in a tabular format (1). The columns of the data table correspond to features of the instances in the rows. In this example based on the Adult dataset (Dua and Graff 2017), the instances are people that are characterized by socio-demographic features. The target is a special column containing what should be predicted (whether a particular person earns more or less than \$50 K/year, in this example). The data is provided to a training algorithm (2) that will learn a model (3). When the model is learned, it can be used to make predictions on instances that have not been used for training (called unseen instances) (4). Performance measures (such as the accuracy of the predictions) are then computed to evaluate the performance of the model (5)

Explainability is therefore the capacity of a model to be explainable by using methods that are external to the black-box model (e.g. visualizations, approximating it with interpretable models, etc.) (Guidotti et al. 2018; Mittelstadt et al. 2019).

Furthermore, three hierarchical elements of the model can be focused by legal requirements. Figure 2 shows these three views with a schematic decision tree as example. Note that the hierarchy presented in this paper follows the legal requirements. Indeed, other hierarchy of model explanation can be proposed (e.g. see Lepri et al. 2018). The first view of the model that we present is the whole model, which is, in the example of Fig. 2, the complete decision tree (see Fig. 2(1)). When using this kind of model to reach a decision, a first question Q_1 is asked. If the answer is yes (or true), the question Q_2 is asked, and Q_3 otherwise. This process continues until the end of the tree (also called a *leaf*) is reached, where a decision D_i is taken. The second view of the model that can be targeted by requirements is a particular decision made by the model (see Fig. 2(2)). Finally, the features that are involved in a particular decision can also be the focus of legal requirements (see Fig. 2(3)). In that case, it is not asked how the features are combined to make the decision, but only to provide the list of features that are used to make a decision.

The different ways legal requirements on explainability in B2C and B2B decisions can be considered in machine learning are presented in Sect. 3.2. The technical solutions to the stronger legal requirements in G2C are discussed in Sect. 3.3.

3.2 Weaker requirements: different explainability levels

As shown in Sect. 2, there is no unique definition of explainability in law. Some explainability requirements relate to the model while others relate to the decision (Wachter et al. 2017; Selbst and Barocas 2018). In addition, some explainability requirements merely relate to the features used by the model to adopt the decision, while others go further and relate to the way the features are combined to make the decision. Furthermore, there can be technical ambiguities regarding legal texts and their interpretation. For instance, the interpretative guidelines on the GDPR by the data protection authorities refer to the "rationale behind" or the "criteria relied on in reaching the decision"

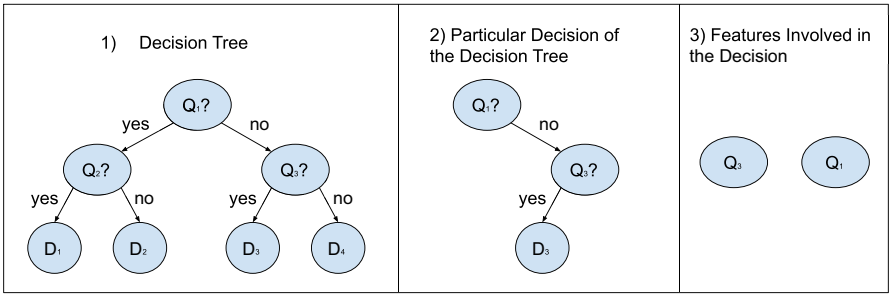

Fig. 2 Weaker requirements can focus on three views of a model: (1) the whole model, (2) a particular decision of the model (here when Q_1 = no and Q_3 = yes), or (3) the features involved in a particular decision (e.g. the age of person (X_1) and the salary (X_3) if they are used in questions such that "is his age lower than 18?" (Q_1) and "is his annual salary lower than 50 k/year?" (Q_3))

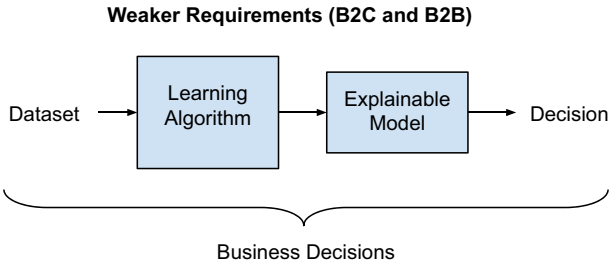

Fig. 3 Input and output of the learning process with legal requirements on explainability in B2C and B2B

Table 1 Summary of the legal texts used as examples in Sect. 3.2

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 for business users 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"
Belgian law of 4 April 2014 on insurances, art. 46: obligation to provide "the segmentation criteria"
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"

(Guidelines on Automated individual decision-making and Profiling, p. 25), which correspond to two technically different requirements.

Therefore, this section analyzes in turn the technical understanding of four levels of legal requirements: providing the main features that are used in the model or in a decision (Sect. 3.2.1), providing all features that are used in a particular decision (Sect. 3.2.2), providing the feature combination that is used to make a decision (Sect. 3.2.3) and providing an interpretable model (Sect. 3.2.4). The inputs and outputs of the learning process constrained by the weaker requirement in B2C and B2B are presented in Fig. 3 and the way directive and regulation examples are technically interpreted is summed up in Table 1.

3.2.1 Requirements on the main features

From a machine learning point of view, the legal texts in B2C and B2B (see Sect. 2.1) refer to four levels of requirement. The first and weakest requirement asks to provide the "main parameters" of the model, or used by the model to take a particular decision. "Parameters" in those legal texts correspond to features of instances in machine learning (see Fig. 1 for a background on machine learning). The methodology used to make the distinction between the main features and the other features is not described in the legal texts. Many machine learning models make it possible to extract the main features they use, even black-box models. The new art. 6(a) of Directive 2011/83 on Consumer Rights states that the "main default parameters" should be provided, without the obligation to instantiate for a particular decision. This means that the main features used by the entire model, not for a particular decision, should be provided.

Providing the main features used in a model is well-developed in the machine learning literature. For linear models, such as linear regression models, a kind of interpretable (or transparent) model, one only has to look at the weights that have been learned for determining the main features that are used. Indeed, given d features $\mathbf{f}_1$, $\mathbf{f}_2$, ..., $\mathbf{f}_d$ for predicting a target $\mathbf{t}$, the goal of the linear model training algorithm is to find the weights w_1, w_2, ..., w_d, such that the linear combination $(w_1 * \mathbf{f}_1) + (w_2 * \mathbf{f}_2) + \cdots + (w_d * \mathbf{f}_d)$ best predicts $\mathbf{t}$. If the d features are transformed in order to be in the same scale (a transformation called *scaling*), sorting the absolute value of the computed weights provide a ranking of the feature importance in the model. In particular, a weight w_j of zero means that the feature $\mathbf{f}_j$ is not used. For instance, if $\mathbf{t}$ correspond to house prices to predict, $|w_{number\ of\ rooms}| = 5$ and $|w_{house\ age}| = 2.5$ mean that the feature "number of rooms" is twice as important as the feature "house age" when predicting house prices. Some works go further and try to determine the features with a non-zero weight that are particularly relevant in a given linear model (e.g. Yu and Liu 2004; Frénay et al. 2014). In that context, a feature is considered strongly relevant if, by removing it, the performance of the model drops. Some features can also be characterized as weakly relevant if they bring new information, but only if other features are removed (John et al. 1994; Kohavi and John 1997; Frénay et al. 2014). These techniques for studying feature relevance in models such as linear models are important because such simple models are widely used in academia, as well as in industry.

In the case of black-box models, features may also be sorted by importance. For instance, random forests (Breiman 2001) use an out-of-bag error during the learning of the decision tree ensemble that can be used to rank features by importance. More precisely, when learning the decision trees in the forest, different sub-sets of training instances are used for learning each tree. The out-of-bag error is defined as the average error made in the prediction of each instance x_i for each decision tree that has not been trained using this x_i (Breiman 2001). In order to know what are the important features, values of each feature $\mathbf{f}_j$ are perturbed (i.e. the values of the feature are changed randomly) and the out-of-bag error is computed. A feature $\mathbf{f}_j$ is considered important if the perturbation of its values increases the out-of-bag error, with respect to the out-of-bag error computed without the perturbation. This idea of

perturbing feature values to assess the importance of each feature used in a model can in fact be extrapolated and applied to any machine learning algorithm (Fisher et al. 2018). In the case of computer vision, where images are used as input, it is less relevant to extract the main features (i.e pixels) used by a model. Instead, one is rather interested in the internal representation used by the model (the extracted features in hidden layers of neural networks). However, extracting the internal features of neural networks is a challenging problem in the machine learning literature. For deep convolutional neural networks, techniques such as saliency maps (Simonyan et al. 2013) and Grad-CAM (Selvaraju et al. 2017) can be used, yet they only extract the main features for specific decisions.

3.2.2 Requirements on all features

The second requirement level is to provide all the features used to take a particular decision. For instance, this is the case of the obligations arising from the GDPR as interpreted by the data protection authorities, under the terms "the criteria relied on in reaching the decision" (Guidelines on Automated individual decision-making and Profiling, p. 25). This means providing all features with a non-zero coefficient in a linear model, or the features in a specific decision path of a decision tree, without necessarily providing the whole tree. Providing all features involved in a particular decision may be motivated by the need to verify the absence of features (or proxies) that are forbidden to use by law (e.g. those that illegally discriminate people).

While it is still possible for all models (using perturbation, for instance, in the case of black-box models) to produce such list of features used, the issue lies in the size of the list. Indeed, providing all features with a non-zero coefficient in a linear model is straightforward, but processing thousands of them as a human is difficult. In order to avoid this issue, some machine learning algorithms incorporate a trade-off between the model accuracy and its complexity. For instance, a technique called Lasso makes it possible to set as many weights w_j as possible to zero when learning a linear model (effect called *sparsity*), while keeping a good enough predictive accuracy (Tibshirani 1996). This makes the resultant linear models much easier to understand. In practice, the balance between accuracy and complexity of the model has to be tuned by the user, depending on his needs. If no means to control the model complexity are provided in the learning algorithm, the problem slides from the machine learning side to the information visualization side, where questions about how to efficiently present information to users is the core issue.

3.2.3 Requirements on the combination of features in a decision

The third requirement level is about the complete explanation of a decision. For instance, this is the case of the obligations arising from the GDPR as interpreted by the data protection authorities, under the terms "the rationale behind the decision" (Guidelines on Automated individual decision-making and Profiling, p. 25). This means providing not only the features used in a decision, but also their combination used to make the particular prediction.

As developed in the literature on interpretability and explainability, this requires to use transparent models such as decision trees or linear models, to create new ones (e.g. supersparse linear integer models (SLIM) Ustun et al. 2013a, b; Ustun and Rudin 2016) or to create ways to explain black-box models (e.g. local interpretable model-agnostic explanations (LIME) Ribeiro et al. 2016). One typical example of interpretable models are the sparse linear models. Most of the time, sparsity in linear models is achieved using Lasso through the ℓ_1-norm (minimize the sum of all weight absolute values, $\sum_i |w_i|$), which is an approximation of the difficult-to-optimize ℓ_0-norm (minimize the number of non-zero weights). SLIM are models that optimize the ℓ_0-norm by transforming the problem. Indeed, instead of optimizing weights w_j with real values, those weights can now take values among a finite set of integer values. Thanks to that transformation, SLIM are more interpretable than classical models by being sparser and by using only integers (instead of reals) as weights, while obtaining similar accuracy scores.

In the case of black-box models, model-agnostic ways to understand those models can be considered. LIME is a technique used to understand specific decisions of a black-box model through the use of an interpretable model. For instance, a specific decision on an instance i made by a black-box neural network can be understood by approximating the decision through local decisions (i.e. decisions that are made on instances similar to the instance i, see Fig. 4). This local model, explaining local decisions, should be interpretable. This local model, that can be a linear model, does not globally explain the black box, but instead provides clues on why a specific decision has been taken by the black-box model. This can be compared to the

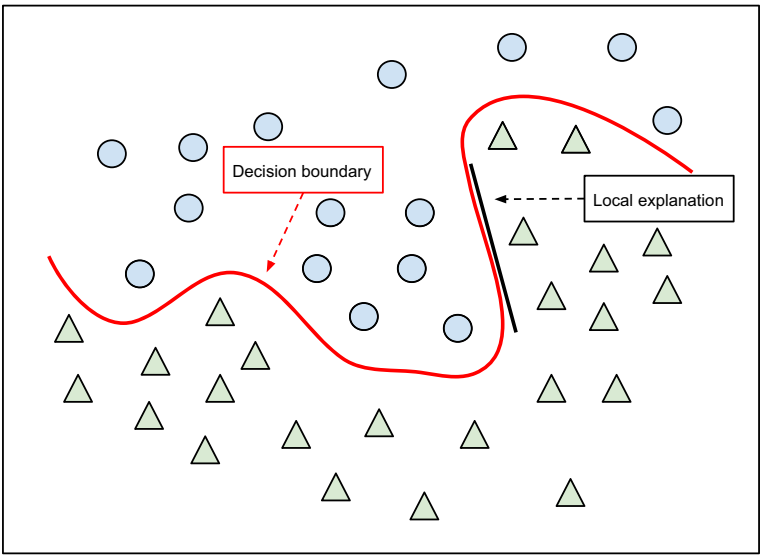

Fig. 4 Figure inspired by Ribeiro et al. (2016). Local explanation of a complex decision boundary (i.e. separating circles and triangles) by using a linear model. The linear model is easy to understand (using the relative value of the weights), but only provides an explanation on the complex decision boundary locally, where it is used

explanation of a particular path in a decision tree: the explanation of the path is local and does not globally explain the whole tree.

3.2.4 Requirements on the whole model

A global understanding of the model would be the maximal explainability requirement that can be asked for a model. Indeed, as a total understanding of the model is required, local explanations cannot suffice. In that case, the legal requirement would constrain the possible usable models to interpretable ones. This kind of requirement exists in the case of financial algorithms, as in addition to provide "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," investment firms can be asked to provide information "about its algorithmic trading and the systems used for that trading" (art. 17(2) of the Directive 2014/65 on Markets in Financial Instruments).

Moreover, in machine learning, Rudin (2019) argues for the need to use interpretable models (such as linear or rule-based models Guidotti et al. 2018), instead of explaining black boxes, in the case of high-stake decisions. This is justified by the fact that the drop in accuracy caused by the choice of an interpretable model can be marginal, for the benefit of having a model that can be understood and trusted by its users.

Section 3.2 presented a translation in vocabulary from the legal literature to the machine literature through four requirement levels related to the weak (B2C and B2B) legal requirements. The four levels were (i) providing the main features used in a decision or the model, (ii) providing all features processed by the model, (iii) providing a comprehensive explanation of a specific decision taken by the model and (iv) providing an interpretable model. The next section focuses on the stronger (G2C) legal requirements and the new machine learning problems that emerge.

3.3 Stronger requirements: new machine learning problems

In addition to the different levels of explanation described in Sect. 3.2, legal motivations need to be given when administrative decisions are made by the model. In the case of administrative decisions, models are now required to provide the law articles behind each decision. This may require to learn the link between

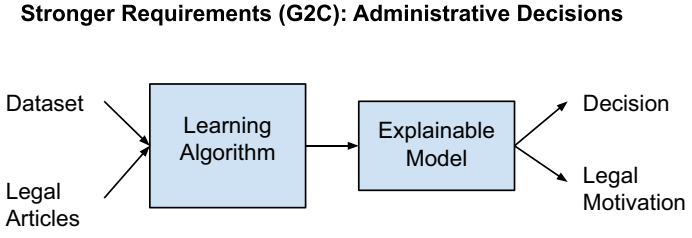

Fig. 5 Inputs/outputs of the learning process for administrative decisions

decisions and legal rules supporting these decisions (see Fig. 5). For instance, decision trees would have to output the legal bases supporting the paths leading to decisions.

On top of this, according to requirements applicable to judicial decision making, judicial decisions need to respond to the arguments submitted by the parties. In this context, a situation described by facts is provided as input to the model, along with textual arguments from both opposing parties. The model has then to output the decision, supported by legal articles as for administrative decisions, while at the same time answering all arguments (see Fig. 6). This means that the provided arguments have to be considered by the model, such that the processed arguments logically support the decision.

Note that the interpretability/explainability problem is re-framed in the context of administrative and judicial decisions. Indeed, in addition to the need for interpretable/explainable models, the stronger requirements ask for the processing of heterogeneous data (i.e., different types of data) for producing not a single output (the decision) but two (decision and legal articles related to the legal motivation) or three (decision, articles and arguments supporting the decision) outputs. This section presents examples of how the machine learning literature tackles the automated judicial decision by only considering the factual description (Sect. 3.3.1), the facts and legal articles (Sect. 3.3.2) and all three possible data elements, i.e. the facts, legal articles and arguments (Sect. 3.3.3).

3.3.1 Explaining judicial decisions with facts only

In the AI and law literature, explainability has not always been linked to the necessity to provide legal motivation and to answer arguments. For instance, Ashley and Brüninghaus (2009) extract facts (called Factors) from case texts in order to predict the decision on the case. In order to do that, the authors derive issues related to the extracted facts by using a domain model. Such domain models are trees defining how facts must be combined to define an issue (such as "trade secret misappropriation"). Given the extracted issues and facts, an algorithm called IBP (i) predicts whether the plaintiff or the defendant is favored and (ii) only provides an explanation of how these facts and issues are used to make the prediction. This kind of explainability is similar to the one discussed in Sect. 3.2.

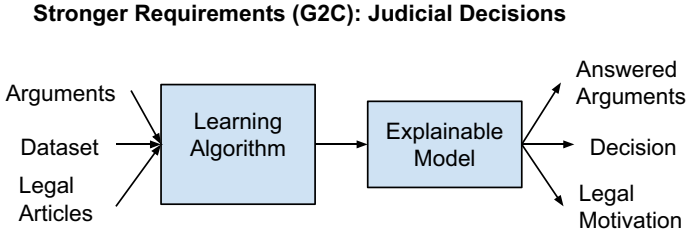

Fig. 6 Inputs/outputs of the learning process for judicial decisions

3.3.2 Explaining judicial decisions with facts and legal articles

A more problem-oriented way to see the stronger requirements of explainability is through multi-task learning. Multi-task learning is a way to learn a global model by splitting the learning into smaller tasks to learn (Zhong et al. 2018). This results in a set of small tasks that is easier to learn than learning solely the global task. Luo et al. (2017) propose to use a neural network (with a mechanism called *attention*) to predict the charges in criminal cases while also providing legal articles supporting the decision. The neural network is defined in such a way that it solves two tasks: charge prediction and relevant legal articles extraction. Following the same idea, Zhong et al. (2018) define the sub-tasks as learning (i) the applicable legal articles, (ii) the charges and (iii) the terms of penalty of a legal judgment, based on a textual fact description. In other words, from the fact description as sole input, multiple output are provided, such as the decision and the relevant articles supporting the decision. One should note that high-performing models used for multi-task learning are often not interpretable. For instance, in the work of Luo et al. (2017) and Zhong et al. (2018), black-box deep neural networks are used to solve the different tasks.

With the objective of making the automated judicial decisions interpretable, Li et al. (2018) use a Markov logic network (MLN) (Singla and Domingos 2005) to predict the outcome of divorce judgments. Their algorithm first extracts logical rules, among other preprocessing steps, from case texts. Then, these rules are weighted and ordered in the MLN such that following the network of rules makes it possible to predict a case outcome. This model is interpretable as humans can understand how the decisions are made.

3.3.3 Explaining judicial decisions with facts, legal articles and arguments

Most of the time, machine learning techniques in the literature do not consider the parties arguments. Despite that, one should note that argument mining and generation is an ongoing work in the literature (Branting 2017; Palau and Moens 2009). One next step could therefore be to design argument mining and generation as two new sub-tasks in a multi-task framework.

In the context of the European Court of Human Rights, Aletras et al. (2016) consider the three elements of interest (description of the situation, the applicable laws and the arguments of the parties) as a text in order to predict if a given article of the European Convention of Human Rights has been violated. They use n-grams on the whole text to train a SVM with a linear kernel in order for their model to be interpretable. By using an interpretable model, they are able to provide how the elements of the case contributed to the decision. As these elements are legal articles and arguments, they can provide clues on how these articles and arguments were used (through the use of certain n-grams by the model) to make the decision.

Ye et al. (2018) consider the generation of court views as a sequence to sequence (Seq2Seq) problem, where a fact description and the charges (which correspond to the decision of another model) are provided as input, and a corresponding court view corresponding to the rationale is generated. The Seq2Seq problem is commonly seen in language translation. In that context, a first sequence of words in a certain

language is provided to the machine learning model and a second sequence of words corresponding to the first sequence but in another language is produced. Court view generation can therefore be seen as a machine translation problem, where the court view would be a "translation" of what can be read in a fact description.

4 Conclusion, discussion and research directions

This paper presents how the law constrains machine learning models regarding their interpretability and explainability. The vocabulary used in law is not always determined, nor consistent in its strength. The constraints on explainability, in their weakest form, can be formulated in a four-level fashion: (i) providing the main features used to make a decision, (ii) providing all the processed features, (iii) providing a comprehensive explanation of the decision and (iv) providing an understandable representation of the whole model.

In the case of requirements related to administrative and judicial decisions, most of the work focuses on interpretable/explainable models, models that provide legal articles supporting their decisions, or both. However, models that provide answers to the arguments of the parties, alongside the decision, are not well studied in the machine learning literature. One clear direction, though, is the use of natural language processing (NLP) to solve the problem, as fact descriptions, legal articles and arguments are often in a text format. Even the explanation of a model's decision can be considered as an NLP problem through, e.g. Seq2Seq learning (Ye et al. 2018).

Note, that in Ye et al. (2018), the explainability of a model judicial decision is provided by the text generated by the Seq2Seq model. However, the Seq2Seq model, which is a deep neural network model, is not itself interpretable. Two different views on explainability are therefore to be put forward.

In the first view, the machine learning point of view, interpretability and explainability are defined on the abstract mathematical model that is used to make the decision (Bibal and Frénay 2016). For instance, decision tree models are considered interpretable because their tree representation makes it easier for humans to understand the abstract mathematical model behind it. By following paths in the tree, users follow a mathematical formula, although in an easier way.

In the second view, that rather corresponds to the legal point of view, explainability can be defined as meaningful insights on how a particular decision is made. In that second view, it is not necessarily required to provide an interpretable representation of a mathematical model, but most importantly to provide a train of thought that can make the decision meaningful for a user (i.e. so that the decision makes sense to him).

This distinction is crucial for drawing the future directions of administrative and judicial decisions made by machine learning models. Indeed, the problem is framed differently for machine learning researchers. In the first view, interpretable/explainable models are used to understand the mathematical processes behind decisions. This requires to develop interpretable models or to make it possible to explain black-box models (as developed in Sect. 3.2). In the second view, providing the human interpreter with an explanation of the decision that makes sense to him is the main

objective, even if the output is not an explanation of the mathematics behind the decision as such. This is the point of view adopted by the Seq2Seq solution, and seems to be the explainability requirement wanted in law.

Following this analysis, we call for a close inter-disciplinary dialogue between the legal and machine learning communities in order, on the one hand, to specify the undetermined terms of the law in the light of their objectives and, on the other hand, to develop new techniques allowing machine learning models to comply with the different level of explainability required by law. Furthermore, this exchange may also help machine learning researchers to more clearly define (and solve) the new problems related to the strongest legal requirements.

References

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V (2016) Predicting judicial decisions of the European court of human rights: a natural language processing perspective. PeerJ Comput Sci 2:e93

Alonso C (2012) La motivation des décisions juridictionnelles : exigence(s) du droit au procès équitable. In *Regards sur le droit au procès équitable*. Presses de l'Université Toulouse 1 Capitole

Ashley KD, Brüninghaus S (2009) Automatically classifying case texts and predicting outcomes. Artif Intell Law 17(2):125–165

Autin J-L (2011) La motivation des actes administratifs unilatéraux, entre tradition nationale et évolution des droits européens. Revue française d'administration publique 1:85–99

Bibal A, Frénay B (2016) Interpretability of machine learning models and representations: an introduction. In: Proceedings of the European symposium on artificial neural networks, computational intelligence and machine learning (ESANN), Bruges, Belgium, pp 77–82

Branting LK (2017) Data-centric and logic-based models for automated legal problem solving. Artif Intell Law 25(1):5–27

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Doshi-Velez F, Kortz M (2017) Accountability of AI under the law: the role of explanation. arXiv preprint arXiv:1711.01134

Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml

Edwards L, Veale M (2018) Enslaving the algorithm: from a "right to an explanation" to a "right to better decisions"? IEEE Secur Priv 16(3):46–54

Fisher A, Rudin C, Dominici F (2018) All models are wrong but many are useful: variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. arXiv preprint arXiv:1801.01489

Frénay B, Hofmann D, Schulz A, Biehl M, Hammer B (2014) Valid interpretation of feature relevance for linear data mappings. In: IEEE Symposium on Computational Intelligence and Data Mining (CIDM), pp 149–156

Goodman B, Flaxman S (2016) EU regulations on algorithmic decision-making and a "right to explanation". In: ICML workshop on human interpretability in machine learning, New York, USA

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv 51(5):1–42

John G. H, Kohavi R, Pfleger K (1994) Irrelevant features and the subset selection problem. In: International conference on machine learning (ICML), pp 121–129

Kodratoff Y (1994) The comprehensibility manifesto. AI Commun 7(2):83–85

Kohavi R, John GH (1997) Wrappers for feature subset selection. Artif Intell 97(1–2):273–324

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2018) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 31(4):611–627

Li J, Zhang G, Yu L, Meng T (2018) Research and design on cognitive computing framework for predicting judicial decisions. J Signal Process Syst 91:1159–1167

Lipton ZC (2016) The mythos of model interpretability. In: ICML workshop on human interpretability of machine learning, New York, USA

Luo B, Feng Y, Xu J, Zhang X, Zhao D (2017) Learning to predict charges for criminal cases with legal basis. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp. 2727–2736

Malgieri G, Comandé G (2017) Why a right to legibility of automated decision-making exists in the general data protection regulation. Int Data Priv Law 7(4):243–265

Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In: Proceedings of the conference on fairness, accountability, and transparency (FAT), pp 279–288

Palau RM, Moens MF (2009) Argumentation mining: the detection, classification and structure of arguments in text. In: Proceedings of the international conference on artificial intelligence and law (ICAIL), pp 98–107

Pasquale F (2015) Black box society, the secret algorithms that control money and information. Harvard University Press, New York

Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the ACM SIGKDD, pp. 1135–1144

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–2015

Selbst AD, Barocas S (2018) The intuitive appeal of explainable machines. Fordham Law Rev 87:1085–1139

Selbst AD, Powles J (2017) Meaningful information and the right to explanation. Int Data Privacy Law 7(4):233–242

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 618–626

Simonyan K, Vedaldi A, Zisserman A (2013) Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint aDrXiv:1312.6034

Singla P, Domingos P (2005) Discriminative training of markov logic networks. In: National conference on artificial intelligence (AAAI), pp 868–873

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol) 58(1):267–288

Ustun B, Rudin C (2016) Supersparse linear integer models for optimized medical scoring systems. Mach Learn 102(3):349–391

Ustun B, Traca S, Rudin C (2013a) Supersparse linear integer models for interpretable classification. arXiv preprint arXiv:1306.6677

Ustun B, Traca S, Rudin C (2013b) Supersparse linear integer models for predictive scoring systems. In: Proceedings of AAAI late breaking track

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 7(2):76–99

Wiener C (1969) La motivation des décisions administratives en droit comparé. Revue internationale de droit comparé 21(21):779–795

Ye H, Jiang X, Luo Z, Chao W (2018) Interpretable charge predictions for criminal cases: learning to generate court views from fact descriptions. In: Proceedings of the conference of the North American chapter of the association for computational linguistics: human language technologies, pp 1854–1864

Yu L, Liu H (2004) Efficient feature selection via analysis of relevance and redundancy. J Mach Learn Res 5:1205–1224

Zhong H, Zhipeng G, Tu C, Xiao C, Liu Z, Sun M (2018) Legal judgment prediction via topological learning. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp 3540–3549