FACULTY OF LAW, ECONOMICS AND FINANCE

LAW RESEARCH PAPER SERIES

No. 2023 - 12

INDIGO WORKING PAPERS

The Interplay Between Lawfulness and Explainability in the Automated Decision-making of the EU Administration

Author(s)

Davide Liga
University of Luxembourg
davide.liga@uni.lu

This article is part of the INDIGO project. The authors are grateful for funding support by the NORFACE Joint Research programme on Democratic Governance in Turbulent Ages and co-funded by AEI, AKA, DFG, FNR, and the European Commission through Horizon 2020 under the Grant Agreement No 822166.

The Interplay Between Lawfulness and Explainability in the Automated Decision-making of the EU Administration

Davide Liga
University of Luxembourg

Abstract

This work has two main goals, on the one side it explores the nature of *explainability* in the attempt to clarify the ambiguous use of this concept and how eXplainable AI (XAI) methods fit into this concept. On the other side, the work describes the legal framework which currently regulates explainability of automated decisions in the context of the European administration, showing to what extent a selection of famous XAI methods meets the requirements of such legal framework.

1. Introduction

Automated Decision Making (ADM) refers to the use of technology to make automatic or semi-automated decisions, i.e., making decisions with limited or no human intervention. The increasing availability of data, combined with more powerful computing capabilities, recently opened a new era of Artificial Intelligence (AI) and Machine Learning (ML), and this was accompanied with a significant increase in the use of ADM systems.

As the use of ADM systems continues to grow, there are also growing ethical concerns being raised around the fairness and transparency of these artificial systems, and around the potential for unintended biases or dangerous misuses. These ethical concerns directly affect the legal dimension, and the necessity to regulate these technologies appropriately.

While these ethical and legal concerns can be considered crucial in any automatised context, their importance is even higher when the automated decision is generated by a public body or institution. This paper focuses on this aspect, considering the use of automated decision systems in the context of European administrative law. In Section 2, we will shortly refer to some related studies. Then we will will discuss the concept of explainability in Section 3, showing why this concept is often connected or overlapped with a range of other concepts, some of which are particularly important in the legal domain. In Section 4, we will describe how the concept *explanation* is instantiated in the context of AI models. In Section 5, we will instantiate the previously discussed concepts in the context of EU Law, describing the interplay between AI explainability and lawfulness. In Section 6, we will describe some famous methods of XAI, showing how some of the most popular methods works from a technical point of view, trying to describe what are the outcomes and limitations of such approaches.

2. Related Works

In recent years a growing number of works has been dedicated to the field of explainability and XAI, due to the increasing relevance of AI systems in people's life. Moreover, due to the increasingly important role of the so called black box models (models which are intrinsically opaque), a huge portion of these studies have been dedicated to understanding how to treat these models and make sense of their predictions and behaviours. Under this growing need of explainability, some popular XAI methods emerged, such as LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017). However, due to the ambiguous and versatile nature of the word "explanation", many scholars have been proposing different interpretations of explainability, with the consequence that a lot of different taxonomies have been proposed to define and classify XAI methods. In this regard, an ambitious work has been proposed by (Speith 2022), which tries to make sense of the various taxonomies and classifications of XAI methods. In our work, we will start from this definitional level of analysis, trying to further clarify what is an explanation, and why there has been so much confusion and overlap between explanability and other concepts. We will also see how this idea of explanability is connected to more specific concepts which are crucial in the legal domain and in legal XAI.

With regard to this intersection between XAI and law, there have been only few studies which analysed the intersection between explainability and law in the field of ADM and EU administration. A crucial work in this regard is (Fink and Finck 2022), which has been of great inspiration for our work and describes ADM in EU administration (Hofmann 2021) by showing the most important legal basis concerting ADM for EU bodies, focusing on both primary and secondary legislation. Some previous works has been dedicated to shed some light on a similar direction, like (Hacker and Passoth 2020) and (Bibal et al. 2021). However, we believe that more effort is needed to address the interconnection between law and XAI methods, especially because these methods are increasingly variegated and show different ways in which explainability can be addressed. For this reason, this work is an attempt to move some steps towards this direction, trying to connect legal requirements with some specific XAI techniques.

3. Explanation and Explainability

One of the problems in the field of XAI is defining what explainability means and what are its relations with other related terms such as "understandability", "interpretability", "transparency".

The Oxford English dictionary defines "explanation" as "a statement or account that makes something clear". Etymologically, the word "explain" is associated with the Latin verb "explanare" which is composed of the prefix "ex" (i.e. *out*) and "planus" (i.e. *plain*), which refers to the idea of *making things plain*. This is contextual with regards to XAI, as the underlying aim of this field is to make the decisions of AI systems clear or understandable to humans.

However, the scientific community proposed different meanings for *explanation* (Guidotti et al. 2018). Moreover, the word *explainability* is often used in reference to (or even in place of) other close or overlapping concepts. When talking about *explainability* in the legal context, the term can be even associated with specific goals such as "justification", "accountability", "fairness", "privacy". We argue that the reason why the term explainability is often used in combination or in reference to other concepts is this multidimensional nature of the explainability.

3.1. Explanation and its Dimensions

The ambiguous use of the term "explainability" is somehow due to the fact that *explanation* is in itself a multidimensional concept whose dimensions can be intertwined. From a very general and abstract perspective, an explanation implies that there is an interaction between a *source* (delivering some piece of information, i.e. the explanation) and a *destination* (receiving the explanation), a *target* (the object of the explanation), and a *rationale* (the reasons and the goals of the explanation).

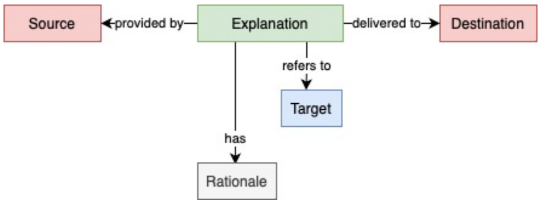

Figure 1: The main dimensions of an explanation.

At the most abstract level, explanations have at least one rationale, namely providing some clarity about the explanation' target[1], we can see this as the very basic rationale of any explanation. In other cases, the rationale can be more specifically related to the context of the explanation: for example, in the context of ADM in the European administration, the rationale of an explanation might be that of providing some kind of assessment with regard to the fairness of an automated decision. In other words, the rationale can be very simple and basic (aiming at providing just clarity) or more complex (being directly connected to the aims or goals of a given explanation). We will clarify this aspect further after.

3.2. Types of Explainability

If an *explanation* is an exchange of information which has the goal to clarify some target, *explainability* is the capacity of some target to be *explainable*. By definition, something is explainable if it can be explained, where *can* usually refers to the intrinsic capability of the target or to an extrinsic possibility[2]. Moreover, the explainability of some target can be seen in a multifaceted way, since it reflects the multidimensional nature of the word explanation. We argue that there are four notions of explainability. The explainability can be acquired, intrinsic, external and contextual.

For example, one can refer to the explainability provided by the source (we call it *acquired explainability*). Supposing that an EU body is using an XAI method to provide an explanation of a specific automated decision from an AI system employed by the EU body itself. In this scenario, some explainability will be provided by the relative XAI algorithm (in this sense, the XAI method/algorithm will be the source of the explanation).

Another notion of explainability is referred to the nature of the target itself (this is the *intrinsic explainability*). For example, supposing that we are using an AI model or algorithm to produce a specific automated decision, our artificial model or algorithm will have a specific level of explainability depending on its nature (e.g. depending on whether it is a transparent model or a blackbox model). This explainability is not acquired from the explanatory process (i.e. by an explanation's source), instead it is an intrinsic quality of the target.

A further notion of explainability is referred to the destination's capability of understanding the target (*external explainability*). As an example, suppose that a decision made by a deep learning algorithm has to be evaluated by people who have no knowledge about AI. In this scenario, we might refer to a lack of explainability of the algorithm's decision because of the illiteracy of the destination. In other words, in this case our notion of explainability will be directly connected to what we called the explanation's *destination*.

Finally, explainability can also depend on the specific kind of explanation which is acceptable in a specific context (*contextual explainability*). This notion of explainability is very much dependent on the underlying rationale of the context in which the explanation is envisaged. For ex-

[1] It is important to remark that sometimes the clarifying information is not needed (what we called "destination" might not need such information to have a better understanding of the target).

[2] Adjectives with the suffix -able/-ability (abilitative adjectives) are multifaceted by nature, since the potentiality channelled by their suffixes have different meanings (possible, capable of, suitable for, allowed to, causing/resulting in). The meaning of -able adjectives depends on the context, on the nature of the adjective itself, and on the object modified by the adjective.

ample, one might say that there is a lack of explainability because a specific rationale is not satisfactorily explained.

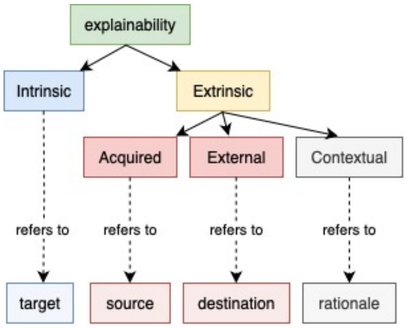

Figure 2: Four different notions of explainability.

In other words, the explainability can depend on each one of four dimensions surrounding the concept *explanation*, as illustrated in Figure 1. Consequently, all these notions of explainability can coexist in the same scenario, showing different analytical angles for the explanation.

For example, suppose that we are dealing with an AI system used by a European administration and that we are interested in understanding why the system took a very specific decision. In this scenario, the *acquired* explainability might be provided by an XAI method and is instantiated by the information provided by the XAI method itself. The *intrinsic* explainability will be determined by the kind of algorithm used by the AI system to produce the decision (is it a transparent AI system, or a black box system?). The *external* explainability will be related to the agents who will receive the explanation (are they capable of understanding the explanation?). The *contextual* explainability will be related to the specific rationale of the explanation (e.g., there might be a requirement to provide an assessments that the decision to be explained was fair and not discriminatory). In this scenario, the explainability of the system will be the result of these different interconnected analytical angles.

To sum up, we can reformulate our previous definition of explainability: *explainability* is the intrinsic or acquired capacity of something to *be explained* (with some purposes or rationales) to some agent.

3.3. Explanations' Rationales and Transparency

As mentioned before, the rationale of an explanation can deeply determine what can be acceptable as an explanation. While the basic rationale of an explanation is to provide some clarity, understandability, or interpretability (i.e. making the target clear, understandable or interpretable by the destination), in some context, this basic clarification is just one of the steps toward a more complex rationale.

In this regard, the rationale can be very much specific to the context in which the automated decision is taken. For example, in some context, we might want our systems to be capable to explain why their automated decisions are aligned to principles such as "privacy", "fairness", "accountability". Other rationales can instead be very abstract and general, like that of providing *trust* (i.e. making the target trustworthy).

In the field of XAI, explanations can also be referred to data, which means we can have specific rationales dedicated to the dimensions of data. For example, one might want to explain data in order to make sure that they are "relavant" (for the task of the AI system which will employ such data), or "representative" (to avoid discriminatory or biased outcomes from the AI system which will leverage such data). In this sense, "relevance" and "representativeness" are other kind of rationales.

In other words, the *explainability* can be connected to different concept because the underlying explanation can be aimed towards different goals (i.e. it can have different rationales).

Transparency A special example of explanation rationale is transparency, which is an instance of complex rationale, since it is a concept which can have different meanings. For example, according to Lipton (Lipton 2018), there are different notions of transparency in the field of AI:

- Simulatability
- Decomposability
- Algorithmic Transparency

Simulatability emphasizes the ease of mentally reproducing the model's decision process. Decomposability highlights the ability to dissect and understand the model's components. Algorithmic transparency focuses on the clarity of the underlying algorithm.

In the context of ADM, especially in EU Administration, EU bodies are required to exert their power by fostering transparency, also in order to grant citizens with a sufficient amount of information such that they are able not only to comprehend their position after the decision is made, but also to challenge the decision itself before the institutions. Therefore, an automated system used by an EU Administration to perform automated decisions, should be capable of providing some degree of transparency for its decisions, assessing whether an AI system has an acceptable level of transparency w.r.t. one or more of the three kinds of transparency mentioned before, depending on the given context.

Moreover, transparency is a complex rational because its scope often overlaps with the scope of other rationales such as "accountability", "trust", and so on. In fact, crucially for ADM in EU administration:

- Transparency ensures that the EU bodies are *accountable* for their actions. It allows the public to verify that EU institutions are functioning properly and are not abusing their power. This also includes how EU bodies use and manage personal and data information.
- Transparency fosters *trust*. When the public can see how decisions are made, it helps to build confidence in the EU bodies and administration.
- Transparency supports the principle of *participatory democracy*. When information is freely available, citizens are better equipped to engage in dialog and decision-making processes.

- Transparency is a hallmark of *good governance*. It contributes to efficiency, effectiveness and rule of law. It allows for scrutiny, which ensures that best practices are being followed and can act as a deterrent to corruption. When there is a high level of transparency, it is more difficult for unethical behavior to go unnoticed.

Moreover, the European Treaties foster the EU institutions to conduct their work as openly and as closely as possible to the citizen, for which transparency is an essential requirement.

It should also be remarked that transparency is very much related but not equal to interpretability, althought some works use them almost in an interchangeable way (Lipton 2018). In the context of explainability, we think that interpretability should be more related to the *subjective* capacities of the destination, while transparency should be more related to the *objective* intrinsic qualities of the target. Similarly, “transparent” is not equal to “understandable”.

As an example, we might consider a very transparent machine learning algorithm like a decision tree. Decision trees are generally considered intrinsically transparent and interpretable *white box* models, because one can see exactly what there is in each of their branches. However, they can also be very complex in their structure or in the interpretation of what each branch represent, which would make them less interpretable *for some people*. In this sense, even if their intrinsic (objective) transparency would not be contested, their interpretability might still be contested (subjectively) because of their complex structure.

4. eXplainable AI

Another source of confusion in the field of XAI is related to the different ways of categorising XAI methods. As we mentioned, XAI methods can be applied to both AI models and data. A famous example of XAI method applied to data is the so-called Explanatory Data Analysis (EDA), which focuses on providing useful insights about data and datasets (as we will see later, this is an important aspect for the lawfulness of AI systems). However, most of the studies on XAI are currently focusing on AI models, either to provide these models with some post-hoc explainability (i.e. providing an explanation for the models’ decisions) or to provide integrated explainability (i.e. creating models which are instrinsically designed to be more interpretable or more transparent)[3].

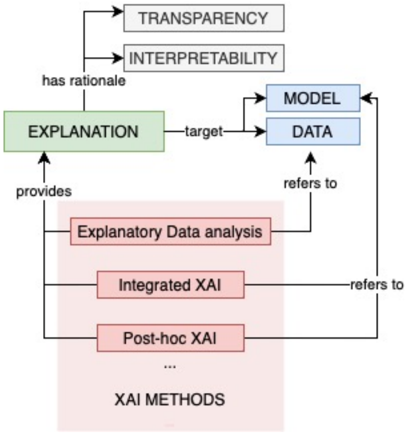

Figure 3: XAI’s explanation scope.

Therefore, it is important to notice that XAI methods are not just used to deal with black-box (i.e. opaque) models. Instead, XAI methods have the more comprehensive goal of enhancing transparency and interpretability of any AI model, even those which are possibly already intrinsically transparent. In fact, *transparent* is not synonyms of *understandable*. Transparent models might still need some XAI method to make them more understandable.

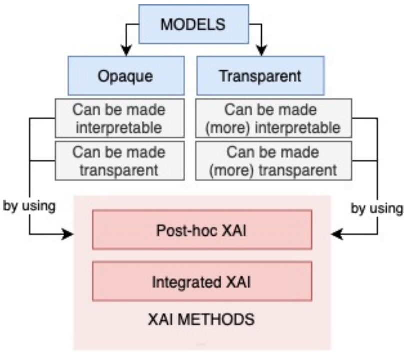

Figure 4: XAI methods can be used both on opaque and on already transparent models.

4.1. Trade-offs in XAI

Explainability cannot be accurately characterized as a binary attribute, ‘explainabile’ vs ‘not explainable’. The attribution of explainability is more similar to a gradient or spectrum of values ranging from high to low, rather than a dichotomous discrete categorization.

Moreover, explainability is often a compromise, since more explainable systems can have less performative outcomes. Highly complex models (like deep learning, random forests, or gradient boosting machines) often give better predictive performance but have a low interpretability because they involve many parameters and complex structures. On the other hand, simpler models (like linear or logistic regression) are easily interpretable but might not perform as well on complex tasks. Figure 5 is a famous graph proposed by (Arrieta et al. 2020) showing how the field of XAI tries to

[3] It can be useful to remark that the formers are related to the *external* explainability mentioned earlier, while the latters are related to the *intrinsic* explainability.

find the right compromise in this trade-off between performance/accuracy and explainability/interpretability.

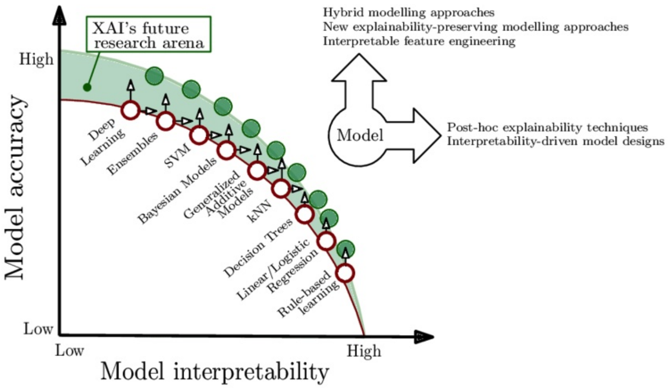

Figure 5: The trade-off between accuracy and interpretability for some types of AI systems. Image taken from (Arrieta et al. 2020).

Apart from the above-mentioned trade-off between *Prediction Accuracy vs. Interpretability*, there are other important trade-offs to consider in the field of XAI.

An important trade-off is *Transparency vs. Usability/Scalability*. In fact, full transparency might require disclosing all aspects of an AI model, which could affect usability by overwhelming non-expert users with unnecessary details. Additionally, creating fully transparent models could require significant computational resources, challenging scalability or efficiency.

Another important trade-off is *Privacy vs. Explainability*: providing detailed explanations may also risk disclosing sensitive details from the training data, causing privacy concerns. On the other hand, obscuring this element for the sake of privacy can compromise the system's explainability.

A further kind of trade-off is *Explainability vs. Time and Compute Resources*, since acquiring highly interpretable models or explanations can be computationally intensive and time-consuming.

4.2. Categories of XAI Methods

To the best of our knowledge, the most complete and comprehensive categorisation of XAI methods is the one proposed in (Speith 2022), which shows the most common ways in which scholars classify XAI methods.

Inspired by (Speith 2022), Figure 6 shows an illustration of different ways in which XAI methods can be categorised.

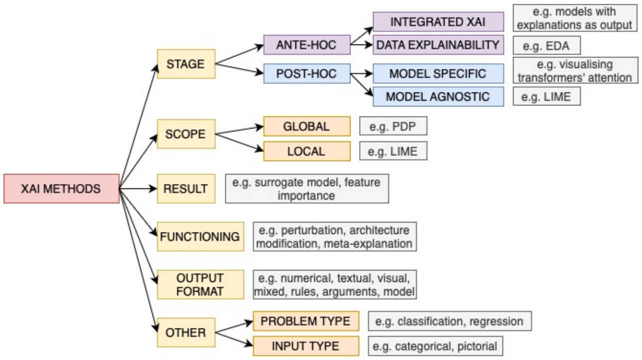
XAI METHODS
STAGE
ANTE-HOC
POST-HOC
INTEGRATED XAI
e.g. models with explanations as output
DATA EXPLAINABILITY
e.g. EDA
MODEL SPECIFIC
e.g. visualising transformers' attention
MODEL AGNOSTIC
e.g. LIME
SCOPE
GLOBAL
e.g. PDP
LOCAL
e.g. LIME
RESULT
e.g. surrogate model, feature importance
FUNCTIONING
e.g. perturbation, architecture modification, meta-explanation
OUTPUT FORMAT
e.g. numerical, textual, visual, mixed, rules, arguments, model
OTHER
PROBLEM TYPE
e.g. classification, regression
INPUT TYPE
e.g. categorical, pictorial

Figure 6: Taxonomy of explainable methods, inspired by (Speith 2022).

Stage One of the main categorisation is related to the "stage" on which the XAI method is dedicated: some XAI methods focus on the "post-hoc" stage (the stage which occurs *after* the model's output or automated decision), while other XAI methods focus on the "ante-hoc" stage (the stage which occurs *before* the automated decision). Post-hoc methods are becoming very popular due to the necessity of explaining black-box models such as those based on Deep Neural Networks (DNNs). Since these models are intrinsically opaque, post-hoc XAI methods tries to shed some light on their behaviour by analysing their output. *Post-hoc* methods can, in turn, be categorised by whether their **applicability** is *model-specific* (i.e. whether the XAI methods can only work for specific models) or *model-agnostic* (i.e. whether the XAI methods can work for any models). For example, explanations based on the visualisation of the attention mechanism can be applied only with neural networks which employs the attention mechanism (e.g. transformers). On the other side, methods such as Local Interpretable Model-Agnostic Explanations (LIME) (Ribeiro, Singh, and Guestrin 2016) and SHAP ((Lundberg and Lee 2017)) can be applied to any model. As far as *ante-hoc* methods are concerned, they can either be dedicated to the explanation of data (as for the previously mentioned Explanatory Data Analysis, EDA) or to the improve the transparency and interpretability of models directly at the modelling stage (which is what we call "integrated XAI").

Scope Another important way of categorising XAI methods is by referring to whether the produced explanations have a local or global scope. On the one hand, *local* explanations provide explainability about why an AI model made a specific single prediction, for example by focusing on how features contributed to that particular outcome. A *global* explanation, on the other hand, describes the overall behavior of the model, providing a general understanding of how the model makes predictions based on all the features across all instances. A famous example of XAI method which provides local explanations is the previously mentioned LIME, while an example of XAI method which provides global explanations are Partial Dependency Plots (PDP). We will have a look to these methods in Section 6.

Results XAI methods can be categorised depending on the kind of output they generate. For example, some XAI methods provide explanations in terms of feature importance. This is the case of famous methods such as LIME and SHAP, which generate graphs to visualize the most important features (i.e. which features have contributed the most in the generation of the automated decisions). Apart from feature importance, another example of result which can be produced by an XAI method are surrogate models, which are simpler and more interpretable models which are generally used to approximate the behaviour of more complex and opaque models (as we will se later, LIME employs surrogate models to generate its explanations).

Functioning XAI methods can also be categorised depending on their main underlying functioning. In this regard, one of the most important kinds of XAI approach is based on *perturbations*, which consist in perturbating the input of a model in order to see how the model behaviour is affected by these perturbations, potentially signalling the importance of some input features as opposed to others. Another example of functioning is the one which is based on the *modification of the model's architecture* (Arrieta et al. 2020). XAI methods functioning in this way simplify complex models by altering their architecture. A further example of functioning are explanations based on previous explanations, also called *meta-explanations* (Samek and Müller 2019). Another category functioning is based on *leveraging the structure* of the model to provide explanations (e.g., using the gradients of a DNN) (Samek and Müller 2019).

Output format Another way of categorising XAI methods is by simply referring to what kind of outputs they provides. Some XAI methods provide *numerical values*, other methods provide *textual values*, others provide *visual representations* such as graphs or diagrams. There are also models which combine different options providing *mixed outputs*. Other kinds of inputs types are *rules*, *arguments*, and even other *models*.

Other Apart from the above mentioned taxonomies, there are other ways of classifying XAI methods. As noticed in (Speith 2022), another potential way of categorising models concerns for which *kind of problem* the XAI method is conceived (e.g. regression, classification). Another way of categorising XAI methods is by referring to the type of *input data* the method employs.

5. Lawful explanations

In this section, we discuss about explainability and lawfulness. We will describe some legal basis which regulate explainability in EU Administration and ADM. At the same time, we will show how these legal provisions are met by the current XAI methods, which we described in the previous sections.

5.1. Duty to Give Reasons

In the context of ADM in the EU Administration, there are "long-standing and deeply rooted explanation duties in administrative law" (Fink and Finck 2022). A pillar of EU administrative law is in fact the so-called *duty to give reasons*. According to this duty, EU bodies are requested to provide reasons for their decisions.

The duty to give reasons is rooted in different legal basis. For example, art.296 Treaty on the Functioning of the EU (TFEU) states that legal acts "shall state the reasons on which they are based". Moreover, art.41 of the Charter of Fundamental Rights (CFR), which is focused on the procedural side, describes a range of rights to ensure the more general right to good administration, stating that there is an "obligation of the administration to give reasons for its decisions". Furthermore, the duty to give a reason is well rooted also as a general principle of law, and the Court of Justice of the EU (CJEU) often refers to the duty to give a reason, which is seen both as a way for EU bodies to exert their power to review the legality of decisions and as a way for citizen to have enough information to assess whether the decisions affecting their lives are well-founded (possibly challenging them if that is not the case).

More precisely, the duty to give reasons requires decision-making authority to state the facts and the most decisive legal considerations, also mentioning relevant counter-argument. According to the CJEU, the reasons provided by the EU bodies must be appropriate to the content of the decision and to the interests of the individuals affected by such decision, which means that decisions having negative and important consequences on an individual require more explanations.

Giving reasons with XAI The problem here is that for an EU body to provide reasons for an automated decision generated by an AI system, the AI system must have some degree of explainability. Moreover, as noticed by (Fink and Finck 2022), "the fact that AI is used may actually be a reason to increase the decision-maker's reasoning obligations". Given that one of the key aspects mentioned by the CJEU in reference to the requirements of the statement of reasons provided by EU bodies is that these reasons have the crucial goal of allowing decision review, a crucial XAI dimension to consider is the one related to the **scope** of the XAI methods. In fact, in case an EU body is requested to review a decision made on a single individual, the provided explanation will probably need a *local scope*, through which the EU body can say why the automated decision had some given outcomes. Moreover, the EU body will probably need to assess the *global scope* of the AI system too, especially in case the local explanation led the EU bodies to judge the automated decision negatively (e.g. unfair or discriminatory). In this scenario, a global explanation could be used to determine whether the AI system tends to reproduce the same unfair or discriminatory automated decision over more individuals, especially if they belong to a minority or to some potentially discriminated group.

Moreover, supposing that global explanation methods show that the AI system addresses a specific group of individuals unfairly, this could mean that the underlying training data on which the AI system was trained on was not sufficiently accurate, relevant or representative. In this regard, according to the art.10(3) of the recent AI Act, "training, validation and testing data sets shall be relevant, representative, free of errors and complete". In this scenario, another group of XAI methods which will be relevant for addressing and evaluating the (unfair) automated decision, would be the one we defined as *data explainability* methods, whose goal is to provide explainability at the level of data (as for the previously mentioned EDA).

5.2. Right to an Explanation

In the last few years, a huge topic of debate has been related to the the existence of a *right to an explanation*. This debate raised from the interpretation of art.22 of the European General Data Protection Regulation (GDPR). Art.22 introduces a prohibition on the use of "solely automated decision-making", stating some exceptions on the third paragraph, art.22(3). For these exceptions, a "right to an explanation" is envisaged in recital 71. The problem here is that the Legislator has decided to add this statement in a recital, i.e. in a non-legally binding provision. This opened a huge debate among legal experts in the attempt to determine whether or not such right actually exists (Goodman and Flaxman 2017) (Wachter, Mittelstadt, and Floridi 2017) (Fink and Finck 2022).

As far as the ADM in EU administration is concerned, the GDPR does not actually apply in the context of the EU administration. In the context of EU administration, the relevant law is the Regulation 2018/1725 (called "EUDPR"), which regulates how EU institutions, bodies and agencies should process personal data. However, the provisions related to the right to an explanation mentioned for the GDPR are identically replicated in the EUDPR: art.22 and recital 71 of the GDPR are equivalent to art.24 and recital 43 of the EUDPR. This means that GDPR and EUDPR share a common ambiguous formulation for the alleged, previously-mentioned "right to an explanation".

As clarified in (Fink and Finck 2022), only case law from the CJEU will clarify whether and to what extent such right exists. However, since this right is defined explicitly (although in a non-legally biding way), it can be inferred that it has at least a "political" or symbolic nature, aiming at shaping some future directions both in terms of legislation and in terms of case law.

Right to an explanation using XAI In case the enforceability of a *right of an explanation* is defined by the CJEU, this right would certainly make the use of XAI methods even more important in the data protection procedures where the the ADM is totally automated. In practice, this would mean that XAI methods would be required to provide explanations in the context of data protection, for the exceptions specified in art.24 EUDPR. These explanations would aim, for example, at describing how an AI system process data, and for which purposes.

5.3. AI Act Requirements

The recent AI Act (AIA) is another important piece of law for the scope of this work, since it is applicable also in the context of the EU administration and it fosters the enforcement of explainability for AI systems. More precisely, the AIA proposed a risk-based approach to regulate the use of AI systems. The deployment of AI systems which are considered at higher levels of risk is subject to stricter requirements and obligations. Among these requirements, a crucial role is played by transparency: the higher the risks of an AI system the higher the level of required transparency.

The AIA is very much focused on the concept of transparency. For example, in art.52, a *right to be informed* is defined, which states that "AI systems intended to interact with natural persons" must be "designed and developed in such a way that natural persons are informed that they are interacting with an AI system". In other words, art.52 creates a simple *duty to inform* the user about the fact that they are interacting with an artificial system, which is quite similar to other analogous provisions in product liability law (where products are required to have some informative statements). Art.13 of AIA is probably more relevant for the scope of our work, since it directly addresses the need to provide explainability (not just informative statements). The article states that "high-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately".

It is relevant that the the legislator decided to say "sufficiently", showing that transparency (similarly to explainability) is a gradual continuous value, and not a discrete dichotomous categorisation. It is also relevant that the envisaged goal is to "enable users to interpret the system", acknowledging the subjective counterpart of the term "transparency", usually referred to as "interpretability".

Going more into the details of art.13, there is an obligation for the providers of high-risk AI systems to provide instructions containing "characteristics, capabilities and limitations of performance of the high-risk AI systems", which includes among other things the intended purpose of the AI system, as well as the level of accuracy, robustness and cybersecurity. Moreover, the instructions must also include "the performance as regards the persons or groups of persons on which the system is intended to be used" and "when appropriate, specifications for the input data, or any other relevant information in terms of the training, validation and testing data sets used, taking into account the intended purpose of the AI system".

Importantly, this requirements set out in art.13 are not to be interpreted as a right to an explanation (mentioned in the previous section), but rather as an *obligation of explainability*, which must be addressed by AI system providers. The providers of AI systems must ensure a certain degree of explainability for their systems, which they can achieve through the use of XAI methods.

XAI methods for the AIA XAI methods can be useful to facilitate the compliance with the requirements set out by AIA. In this regard, XAI methods can be used by providers of high-risk AI systems to generate some of these informative instructions. For example, this might be the case for the requirement related to the performance with regards to specific groups of people, since some features of the data can be showed to be relevant for the robustness of the model with regard to specific data points. In this sense, a combination of local and global methods would be needed, similarly to what we said earlier w.r.t. the *duty to give reasons*[4].

Another important aspect in this regard concerns the previously mentioned requirements in terms of input data, including training and validation datasets. To understand better this requirement, recital 44 can be a complementary source of information, since it specifies that data must be “relevant, representative and free of errors” as well as “complete in view of the intended purpose of the system”. Moreover, requirements are further specified in art.10(2) of AIA, related to “Data and data governance”, where some requirements are laid out, which concern the practices that data governance should employ. These practices should concern, for example, data collection, design choices, any relevant data preparation and manipulation. Moreover, data governance and management should concern “a prior assessment of the availability, quantity and suitability of the data sets that are needed” and “examination in view of possible biases”.

In this regard, data explainability (like EDA) can surely be used to describe the relevance of data with respect to the intended goals of the system. In other words, to comply with these requirements providers of high-risk AI systems will probably need to address different XAI methods, both those related to the explainability of the underlying models (e.g. integrated XAI, post-hoc XAI) and those related to the explainability of the employed data (e.g. EDA).

This *obligation of explainability* set out in art.13 is even more important when considering *point e* of art.13(3), which states that the information should include “the expected lifetime of the high-risk AI system and any necessary maintenance and care measures to ensure the proper functioning of that AI system, including as regards software updates”. This requirement is closely connected to the data governance, since some AI systems might require a periodic update of the underlying training data, which means that data explainability will be periodically needed.

5.4. XAI as a compromise

It is important to underline that the legal framework described so far seems to give an important role to the performance of the models. For example, we mentioned the “the performance as regards the persons or groups of persons”, which seems to be a reference to the risk of some systems to be unfair, discriminatory or simply non-representative with respect to specific groups. However, this is where it becomes clear that in some cases AI providers will need to face a compromise, given the trade-off between explainability and performance mentioned in Section 4.1 and described in Figure 5. For example, we might have cases in which the contested opacity of some AI system might be justified by a higher capacity of such a system to be fair. Similarly, we might have cases in which the required transparency of some high-risk AI system might produce a lower capacity of such a system to generate fair decisions. In these cases, the compromise will probably require AI providers to use a combination of different XAI methods, tackling explainability from different perspectives at the same time, including the modelling stage (integrated XAI), the post-modelling stage (post-hoc XAI), as well as data explainability.

6. Methods of XAI

We will now describe some popular methods of XAI. In particular, we will describe two famous model-agnostic XAI methods, namely LIME and SHAP. Moreover, we will describe a well-known XAI method for global explanations called Partial Dependency Plot (PDP). While describing these methods, we will discuss how they can meet the legal requirements set out in the previous section.

6.1. LIME

LIME, or Local Interpretable Model-Agnostic Explanations (Ribeiro, Singh, and Guestrin 2016), is a method for explaining the predictions made by any machine learning model. LIME creates interpretable explanations by approximating the prediction surface locally, around the outcome to be explained. To do this, LIME generates a new dataset of perturbed samples, obtains the predictions for these from the original model, and then applies a simple model (e.g. a linear model) to these samples. The coefficients of the simpler model serve as the explanation and can help in understanding how each feature affects the prediction for the specific instance to be explained. The advantage of LIME is that it provides model-agnostic and locally faithful explanations, helping to interpret complex models.

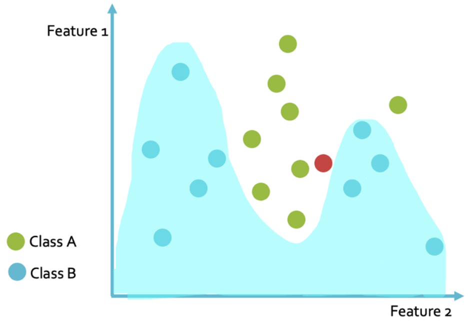

Figure 7: Illustrative example of non-linear decision boundary of a complex (black-box) model. The red data point is the one for which we want an explanation.

To understand how LIME works, suppose we have a black-box model which generates some complex (non-linear) decision boundary on our data points. To keep this scenario simple, we can consider a simple binary classification. The decision boundary might look similar to the one depicted in Figure 7, and we might have a data point for which we want to know why a decision has been made. For

[4] The duty to give reason in Administrative Law is very much related to the obligation to an explainability of the AI Act, even if the former is intended for EU bodies only, while the latter includes EU bodies as well as private stakeholders.

example, in Figure 7, the red data point is classified as belonging to class A (green), because it fall outside the decision's "blue area", which represents our decision boundary (we coloured the point red just to say that it is the data point we want to target). Intuitively, this target data point could be seen as the single automated decision which a citizen might want to have an explanation for.

To give some explanations about why a decision was taken (i.e., why the red dot was classified under class A), LIME will focus on the local boundary in the proximity of the targeted data point, as illustrated in Figure 8. This is crucial because by zooming in the vicinity of a specific data point we can approximate a linear decision boundary, which is more explainable than the complex non-linearity of the global model.

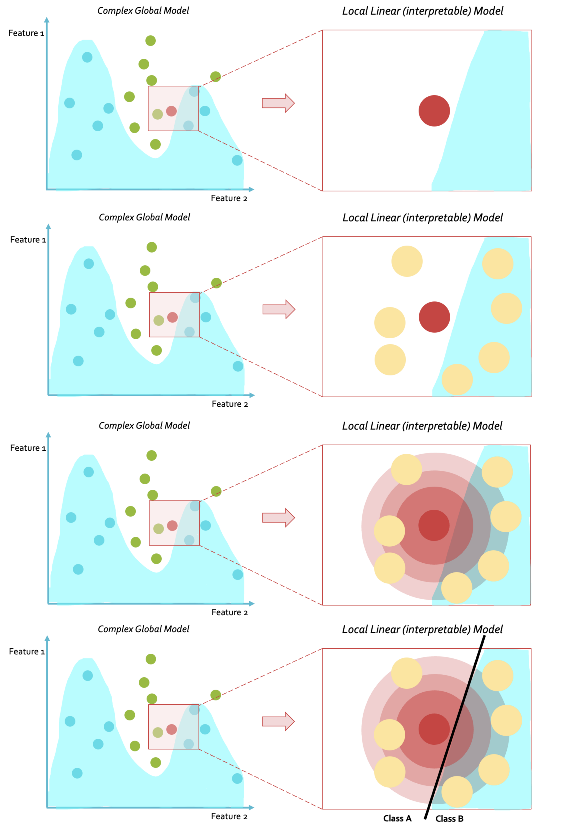

Figure 8: Illustrative example of the local boundary targeted by LIME.

As can be seen from Figure 8, LIME roughly performs 3 steps:

- It focuses on the vicinity of the targeted data point;
- It creates some perturbation on the data points (the yellow dots);
- It weight the data points depending on their vicinity to the target;
- It create a surrogate linear model which approximate the behaviour of the complex model.

Since LIME creates a *linear* surrogate model which approximate the (local) behaviour of the complex black-box model, it generates numerical coefficients which can explain the contribution of each feature in determining why the data point fall on one side of the decision boundary as opposed to the other side. In this scenario, LIME will generate some visualisations which indicate the contribution of each feature (i.e. the feature importance) in the prediction of our targeted data point, as illustrated in Figure 9.

As we can see, LIME employs some of the aspects we mentioned in the previous sections: perturbations (as a way of *functioning*), surrogate models (which is an intermediate *result*), feature importance (which is the final *result*), visualisation (which is the *type of output*).

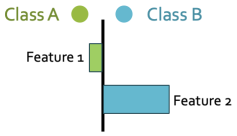

Figure 9: Example of visualisation of feature importance showing that feature 2 contributed significantly towards the prediction.

LIME and its usefulness for legal XAI From the point of view of the legal requirements we described in the previous section, LIME can certainly be helpful in meeting the legal requirements of the current legal framework governing ADM in the EU administration. Since LIME provide local explanations, it is particularly suitable to address those situations in which an individual wants to exert their right to contest an automated decision which significantly affected them. In this scenario, the EU bodies might want to use LIME to address what features determined the given decision, as a step towards the clarification of the righteousness of the automated decision. Importantly, while this might provide the targeted decision with additional explainability, potentially meeting the previously mentioned *duty to give reasons*, this local explainability might not be sufficient, or even not significant. In fact, one of the criticisms of LIME is that it lacks consistency. Specifically, LIME does not guarantee that if the model changes such that it relies more on a feature, the attributed importance for that feature should not decrease. This means that LIME's local explanations are sometimes uninformative, and this could and shold push an EU body to search for complementary explanatory insights, through the use of other XAI methods.

6.2. SHAP

Another very famous method which recently achieved enormous success is SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017), which is a unified measure of feature importance that assigns each feature an importance value for a particular prediction. SHAP values are based on the concept of a Shapley value from cooperative game theory. Their main characteristic is that they represent a fair distribution of the contribution of each feature to the prediction for a specific instance.

To understand this concept, it can be helpful to consider that the features employed in a machine learning algorithm have both an individual contribution towards the achievement of a specific prediction and a “collective” contribution (in the sense that their contribution is not just individual, but also correlated to the presence of other features).

Metaphorically, we can think of features as single individuals of a team. This team of individuals might have achieved a specific result (i.e. the prediction) and we might want to know which is a fair distribution of the merit for each individual (i.e. which is a fair distribution of the contribution). Shapley values, from cooperative game theory, answer this question by providing the so called “marginal contribution”. This term refers to the additional benefit or value that is gained from increasing a particular input or factor, while keeping all other factors constant.

SHAP employs this concept by considering different coalitions of inputs and by calculating the marginal contribution for all of them. The intuition behind this process, described in the previous metaphor, is illustrated in Figure 10.

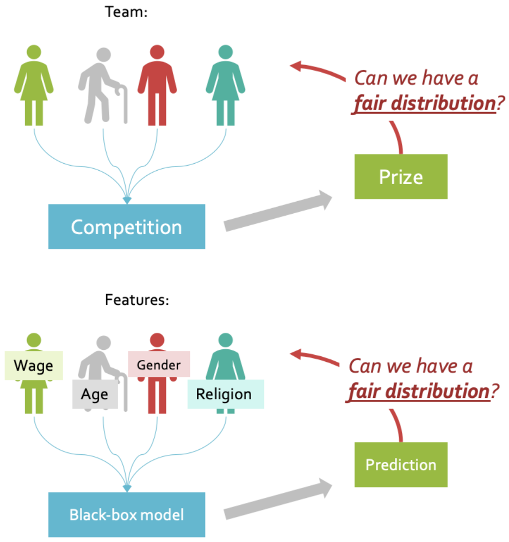

Figure 10: Shapley values, from game theory, address the problem of how to find a fair distribution of contribution. SHAP translate this concept in XAI, in order to find the features’ contribution for a specific prediction.

To calculate the contributions of each feature, SHAP divides features in coalitions, where each coalition is a subset of features. This is particularly important, because some features achieve better results when they are together (while perhaps their contribution is negligible when only one of them is employed). To stick with the previous metaphor, supposing we have a team like the one on the top of Figure 10, it might be the case that the green individual and the gray individual have the greatest contribution in the achievement of the prize when they operate together, while they might have a negligible contribution in case they operate separately.

For each coalition (i.e. for each subset of features) SHAP compares the difference in the prediction when removing a single element of the coalition (i.e. a single features). In this way, it is possible to calculate all marginal contributions of the features, having as a result a numerical representation visualized in a graph, where we can see which features contributed the most in a specific prediction.

Althought SHAP, like LIME, is mostly thought as a local method (since we are trying to explain single predictions), it can also be used with a global scope by aggregating local explanations, which is one of the advantages of SHAP w.r.t. LIME.

SHAP and its usefulness for legal XAI While both SHAP and LIME are frequently employed, SHAP have some characteristics that make it more suitable in certain contexts:

- **Consistency:** The main advantage of SHAP over LIME is consistency. SHAP values are consistent in their explanations, which means if we change a model to rely more on a feature, the attributed importance for that feature should not decrease. This consistency is lacking in LIME.
- **Game theoretic approach:** SHAP is based on game theory, which provides a more solid theoretical foundation and justification for the calculated importance values. This arguably makes SHAP’s interpretability stronger as compared to LIME.
- **Global interpretability:** Along with local interpretability (explaining individual predictions), SHAP also provides global interpretability (which can describe the behavior of the whole model).
- **Superior model agnosticity:** While both LIME and SHAP are model agnostic XAI methods, LIME has been criticized for the fact that the reliability of its results depends too much on the selected neighborhood size (which determines the weighting process described in Figure 8), which is in turn a factor that is strongly affected by the underlying model. SHAP, instead, does not have this issue, since it is based on the above-mentioned game theoretic approach.
- **Handling feature interactions:** another consequence of SHAP’s game theoretic nature is that SHAP handles interactions between features, which is extremely important in some cases where the contribution of a feature can be correlated with the values of other features (i.e. in case there is a high level of correlation among features).

From the legal point of view, the superior consistency of SHAP can certainly make it more suitable for some scenarios, like the one related to the AIA’s *obligation of explainability*, which obliges the providers of an high-risk AI systems to ensure levels of explainability which are consistent and coherent with the purposes of the AI system. In fact, providers of high-risk systems (including EU bodies) will presumably be motivated to mitigate the perceived risk of their systems, therefore trying to show that their design choices have been addressed with a look to specific numerical values which have been consistently adjusted following the numerical explanations provided by SHAP.

This might be useful not only when providers propose their systems, but also when providers update or improve their systems (according to the requirements set out in art.13 of the AIA, related to the lifetime and continuous maintenance of the AI systems). Clearly, the obligation to update their systems (as well as any need to improve a flawed system) can be guided through an explanatory process only if such explanatory process provides consistent responses to the newly introduced integrations.

6.3. PDPs

Partial Dependency Plots (PDPs) (Friedman 2001) offer a way to visually explore the relationship between a small number of input variables and the predictions made by a model. Similarly to LIME and SHAP, PDPs are model-agnostic (they can be used on any model). However, contrarily to LIME, PDPs operate on a global scope, i.e. they are a *global* XAI method. A PDP shows the marginal effect of a feature on the predicted outcome of a model, taking into account the average effect of all other features (this is why PDPs are a global XAI method). This is accomplished by systematically varying the values of the feature of interest, while holding all other features constant at their average values, and graphing the effect on the prediction. PDPs are particularly useful for visualising interactions between features and their impact on the prediction, and can be used with any type of machine learning model (it is a model-agnostic method).

For example, we might have a series of predictions generated by our model and we might want to know more about the relation between these predictions and the input features on which our model was trained, from a global point of view. Suppose, for example, that our predictions are related to court decisions in the field of criminal law, where the prediction is "approved" or "rejected", and suppose we have some features (for example, we might have both legal aspects and factual aspects). In this scenario, our features might be the nullity of the hearing (legal factor), the suspected criminal organisation of the defendant (legal factor) and the number of years the defendant has been already in detention (factual factor). In this example, we might have a list of predictions made by our model, which might look like Table 1 and which we might want to explain from a global point of view.

Feature 1	Feature 2	Feature 3	Result
no	Camorra	2	approved
yes	Cosa Nostra	12	rejected
yes	Ndrangeta	0	rejected
...	...	...	...

Table 1: Prediction examples. Feature 1 = nullity, Feature 2 = suspected criminal organisation, Feature 3 = years in detention.

For example, we might want to see what is the global behaviour of our model's prediction with regard to Feature 3 (the number of years the defendants already passed in detention in the past). In this scenario, we will follow the following simple steps:

- Choosing a set of fixed values for the selected feature (for example, from 0 to the maximum value found in our dataset, say 12).
- For each fixed value, we will create a modified dataset where all instances have the same fixed value for the selected feature, while keeping the original values for the other features.
- We will run the model's predictions for each modified dataset.
- We will calculate the average prediction for each unique fixed value of the selected feature, plotting it on a graph.

A drawback of PDPs is that they can be misleading when there are strong interactions or correlations between features or when missing data is not handled correctly.

PDPs and their usefulness for legal XAI Being a very intuitive and easily understandable method of global XAI, PDPs can be particularly useful for the purposes of the previously-mentioned *obligation of explainability*. However, it is important to noticed that this method of XAI should be employed in context in which there is not a strong correlation between features, because it could be a weakness in the robustness of the provided explainability.

Moreover, this method can arguably be useful to meet the *duty to give reasons* for the automated decisions performed by EU bodies and affecting single individuals (although in this case, reasons should probably be accompanied with some complementary local explanations directly connected to the single decision which affected the individual). In other words, also in this case, we can see that each method has advantages and limitations, and the optimal solution is often a combination of different XAI approaches.

7. Conclusion

This work tackles two directions, on the one side it tries to shed some light on the interconnection between explainability and other related terms such as "interpretability" and "transparency". In this regard, we showed why explainability is often used in combination or even in overlap with other terms, arguing that this versatility is somehow justified by the intrinsic multidimensional nature of the concept "explanation". On the other side, this work shows how explainability is practically instantiated in the context of Automated Decision Making (ADM) for the European Union administration, by referring to the legal basis which currently dominates the the explainability requirements for automated systems in EU bodies. In this regard, we showed some of the most important obligations and rights which EU bodies must address when using ADM systems, considering more specifically how XAI methods can fit these obligations. Furthermore, we discussed why EU bodies will likely need to address explainability requirements by employing different XAI methods in order to tackle different explanatory angles, given that there is no perfect-for-any-scenario approach.

References

Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion* 58:82–115.

Bibal, A.; Lognoul, M.; De Streel, A.; and Frénay, B. 2021. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law* 29:149–169.

Fink, M., and Finck, M. 2022. Reasoned a (i) administration: explanation requirements in eu law and the automation of public administration. *European Law Review* 47(3):376–392.

Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232.

Goodman, B., and Flaxman, S. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine* 38(3):50–57.

Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51(5):1–42.

Hacker, P., and Passoth, J.-H. 2020. Varieties of ai explanations under the law. from the gdpr to the aia, and beyond. In *International workshop on extending explainable AI beyond deep models and classifiers*, 343–373. Springer.

Hofmann, H. C. 2021. An introduction to automated decision-making (adm) and cyber-delegation in the scope of eu public law. *University of Luxembourg Law Research Paper* (2021-008).

Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16(3):31–57.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144.

Samek, W., and Müller, K.-R. 2019. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning* 5–22.

Speith, T. 2022. A review of taxonomies of explainable artificial intelligence (xai) methods. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2239–2250.

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