The Right to an Explanation under the GDPR and the AI Act

Bjørn Aslak Juliussen[1 [0000−0003−0046−8263]⋆]

Department of Computer Science, UiT The Arctic University of Norway, Tromsø, Norway

Abstract. The article provides a comprehensive overview of European regulations, the GDPR and the AI Act, focusing on the right to explanation for individual decisions inferred from high-risk AI systems and automated decision-making. It analyses the concept of the right to explanation in automated decision-making processes, emphasizing the legal obligations surrounding the provision of meaningful information pre- and post-decision. The paper examines explainable AI (XAI) methods in this context, categorizing them as intrinsic and post-hoc, with examples like decision trees, Shapley values and Local Intepretable Model-Agnostic Explanations (LIME). By analysing the legal and technical dimensions together, insights into the complex interplay between data protection, AI regulation, and the quest for transparency in the EU acquis are made.

Keywords: The Right to an Explanation · EU Law · XAI.

1 Introduction

Imagine these two scenarios: You receive a warning that your current employment contract is going to be terminated due to poor work performance, or you get a notification from the tax authority that they will be conducting a manual inspection of your tax returns for the last 10 years. A natural response to either of these two notifications is to ask: Why? In both scenarios, the majority of people would most likely view it as unacceptable to receive a response that only states that an AI system recommended terminating the contract and manually reviewing the tax returns.

The following sections evaluate if the affected person has a right to a meaningful explanation in the scenarios presented. Specifically, the right to an explanation under the GDPR (sec. 2) and the EU AI Act (sec. 3) are analysed [1, 2]. Due to their interrelation, both rights are examined together. The subsequent sections (sec. 4) assess various XAI methods to determine their compliance with these rights. Finally, section 5 summarizes the findings.

⋆ Email: bjorn.a.juliussen@uit.no

2 The Right to an Explanation of Automated Decision-Making under the GDPR

The right to an explanation of automated decision-making in the GDPR presupposes that the overall scope of the GDPR – the territorial and material scope – enters into effect and that the scope of Article 22 of the GDPR is activated.

Article 22 prohibits automated individual decision-making, including profiling, with several exemptions. In order for Article 22 to enter into effect, the decision must be based solely on automated processing and the decision must produce legal effects concerning a natural person or similarly significantly affect him or her.

The right to an explanation of an individual automated decision made in accordance with Article 22 of the GDPR is not part of the wording of Article 22. The right to an explanation of such a decision is part of Article 13 – 15 of the GDPR. The right to be informed of the reasons behind an automated decision is, thus, separated from the right not to be subject to such a decision.

Article 13 of the GDPR requires that controllers – the natural or legal person that determines the purpose of the processing and the means applied – provide specific information to data subjects when collecting their personal data directly from them.

Article 14 of the GDPR concerns the information the controller is required to inform the data subject about when the personal data has not been obtained from the data subject, e.g., from other data subjects, other controllers, where personal data is collected from sensors, or similar.

Both Articles 13 and 14 regulate the information that needs to be given to the data subject by the controller ex officio. The data subject is not required to perform any action, such as making a request, to obtain the information covered in Articles 13 and 14. Article 15, however, provides the right to access information for the data subject and regulates the information that the controller is required to provide when requested by the data subject.

Articles 13 and 14 require the controller to provide the data subject with the information at the time of collection of the personal data, while Article 15 requires the controller to provide the information at the time of the request for information, and no later than one month after the request has been submitted, under Article 12 (3).

Article 13 (1) (f), Article 14 (2) (g), and 15 (1) (h) all have the same wording regarding automated decision-making under Article 22:

> The controller shall provide the data subject with (...), **meaningful information about the logic involved**, as well as the significance and the envisaged consequences of such processing for the data subject".

The phrase 'meaningful information about the logic involved' is debated in the scholarship. Some view it as a right to explanations of automated decisions [3]. Others argue the right is minimal or nearly non-existent [4]. Some scholars suggest a contextual interpretation [5].

Article 22 – interpreted in light of recent case law from the Court of Justice of the European Union (CJEU) – is a right with a corresponding prohibition the controller needs to implement. The controller therefore needs to map out their processing operations and to have control over whether or not they have individual automated decision-making processes implemented. The complicating factor of the right to an explanation is the requirement to provide the data subject with "meaningful information about the logic involved" in automated decision-making.

What does "meaningful information about the logic involved" entail when a deployed AI system is used for automated individual decision-making? The right to meaningful information about the logic involved was not included in the predecessor of the GDPR, the data protection directive [6]. The right to meaningful information about the logic involved has not been interpreted by the CJEU, and there is only some scarce guidance on the right to meaningful information from the European Data Protection Board (EDPB) [7]. The right to receive meaningful information about the logic involved will, therefore, be interpreted in line with the existing legal sources in the following sections, mostly the wording of the GDPR – including the wording of the different official language versions – and legal literature.

A key question is whether "meaningful information" should be interpreted differently across articles 13, 14, and 15, despite the identical wording.

It is necessary to interpret the different articles not just according to their wording, but also in light of their objectives and context. Articles 13 and 14 of the GDPR regulate the information the controller needs to give to the data subject when collecting personal data, typically through the information provided to the data subject in a privacy policy. The objective of providing this information is to make the data subject aware of how his or her personal data is going to be processed. Since the information is given at the time of collection, before the actual processing by the use of automated individual decision-making has taken place, the "meaningful" criterion needs to be interpreted in this context.

The information given to the data subject about the intended automated processing at the time of collection under Articles 13 and 14 needs to be meaningful to enable the data subject to decide whether or not to consent to the processing – if the processing relies on Article 6 (1) (a). Moreover, the information provided to the data subjects should enable the data subjects to assess whether or not they should invoke specific rights when the processing has commenced.

Since the processing of personal data in the automated decision-making process has not commenced when the personal data is collected, meaningful information about the logic involved under Articles 13 and 14 would entail a general description of the overall AI system intended to be used in the automated decision-making process. Such a description could include information about how the AI model is trained, i.e. the training data and the type of AI algorithm used, typical outputs of the AI model, i.e. if the output is a prediction, classification, or generated content, and the sensitivity and the positive predictive value of the finished trained AI model, if possible. These various types of information

would be examples of information given that would enable the data subject to invoke their rights at the time of collecting the personal data and throughout the processing lifecycle.

The right to receive meaningful information about automated decision-making under Article 15 is part of a reactive right after the processing of personal data has taken place. According to Recital (63) of the GDPR, the purpose of the right to access in Article 15 is to enable the data subject to "verify the lawfulness of the processing". When the data subject submits an access request, personal data has been processed and the data subject wants to receive information about the actual processing.

"Meaningful information about the logic involved" will, thus, differ across Articles 13 and 14 and Article 15. What constitutes meaningful information will differ when the information is general information about the overall processing in an AI system prior to the processing, compared to information about the actual processing of personal data that has taken place within a system pursuing an access request from the data subject.

To illustrate with an example, when a data subject has had their loan application rejected due to profiling in an automated credit scoring process and submits an access request, the data subject does not – most likely – have a general curiosity about the processing but is enquiring about the individual decision that has taken place and why it was rejected.

To better understand the term "meaningful" in Article 15 (1) (h), it can be helpful to consider other official language versions of the GDPR, as all versions carry equal authenticity and require a uniform interpretation based on the real intention of their author, as established by the CJEU [8].

The term "meaningful" in the English version of the GDPR, is "aussagekräftige informationen" in the German version, "information utiles" in the French, "nuttige informative" in the Dutch, and "meningsfulde" in the Danish language versions of the GDPR. Some nuances are present in each of these language versions. The French, Dutch and Danish wording entails that the information should be understandable, helpful, and useful for the data subject. The German version entails that the information given should be sound and reliable. The information explaining the logic involved in the automatic decision-making process, therefore, needs to be an actual and reliable representation of the automated processing. The German wording "aussagekräftige" also supports the utility of the information. The information provided under an access request should enable the data subject to assess the lawfulness of the processing [9]. The "meaningful information" condition under Article 15 therefore carries elements of understandability, usefulness, reliability, and utility.

The meaningful information should be about the "logic" involved in the automated decision-making process. The wording of Article 15 concerns the "logic" of the automated processing and not the specific technology applied. Hence, the data subject does not have the right to obtain the name of the AI system or the AI system providers' name under an access request.

According to Recital (63) of the GDPR, the right to receive meaningful information about the logic involved in automated decision-making should not "adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject". The controller could therefore refuse to give information about trade secrets and copyright-protected material under the access request, but could not refuse an overall request for meaningful information about the logic involved by reference to Recital (63).

The information about the logic involved must be meaningful for the data subject. The explanation of the logic could be descriptive and at a high level. However, the level of abstraction in the explanation must be interpreted in line with the purpose of Article 15, to enable the data subject to assess the lawfulness of the processing, according to Recital (63). This needs to be assessed contextually and on a case-to-case basis. However, two examples can be given that do not allow data subjects to assess the lawfulness of the processing for automated decision-making. An explanation of the logic such as that the automated decision-making process "applies machine learning" or "applies AI" is too abstract for the data subject to assess lawfulness. On the other hand, descriptions such as "the automated decision-making processes uses a support vector machine to assess whether individual data points are placed on the maximum-margin hyperplane during the perceptron of optional stability". The latter description is too advanced for it to be meaningful for the data subject and the description does not make it possible for them to assess the lawfulness of the processing.

To recapitulate, both Articles 13, 14, and 15 contain a right to receive meaningful information about the logic involved when processing personal data as part of automated decision-making under Article 22 of the GDPR. However, Articles 13 and 14 of the GDPR apply to the collection of personal data, while Article 15 contains reactive rights that are dependent on requests from the data subject. Hence, Articles 13 and 14 on the one hand and Article 15, on the other hand, will provide different types of explanations where the first is more general ex-ante AI system descriptions and the latter is ex-post explanations of the output.

Article 15 requires the controller to provide "meaningful" information. This condition entails that the information must be understandable, useful, reliable, and helpful for assessing the lawfulness of the data processing for the subject. It must be evaluated contextually on a case-by-case basis.

3 The Right to an Explanation under the AI Act

An affected person has a right to an explanation of individual decision-making under the AI Act. According to Article 86 (1) of the AI Act:

> "[a]ny affected person subject to a decision which is taken by the deployer on the basis from the high-risk AI systems listed in Annex III (...), which produces legal effects or similarly significantly affects that person in a

> way that they consider to have adverse impact on their health, safety or fundamental rights shall have the right to obtain from the deployer clear and meaningful explanations of the role of the AI system in the decision-making procedure and the main elements of the decision taken".

High-risk AI systems in Annex III point 2 – critical infrastructure such as safety components in the management of critical digital infrastructure, road traffic, or in the supply of water, gas, heating and electricity – are not covered by the right to an explanation of an individual decision-making process under Article 86, according to Article 86 (1).

The right to an explanation is part of the remedies subsection in the enforcement section of the AI Act. The purpose of the right to an explanation is, thus, to act as a prerequisite to the providers and the deployers complying with the AI Act through affected persons requesting information. When the conditions under Article 86 – elaborated below – are fulfilled, the activating criterion for the scope of the right to an explanation is that the affected persons "consider" that the output from the AI system has an adverse impact on their health, safety or fundamental rights.

One relevant question is whether the deployer is only required to explain the role of the output of the AI system in the later decision process or whether the logic of the AI system needs to be explained. The wording of Article 86 (1) only requires explaining the role of the AI system in the decision-making process. However, if Article 86 (1) is interpreted in line with Recital (171) of the AI Act, it becomes evident that the explanation should be "clear and meaningful" and provide "a basis on which the affected persons are able to exercise their rights". The right to an explanation under Article 86 (1) does, thus, not only cover the simple role the output from the AI system had in the decision process but also – as far as it is feasible – the data, algorithm type, and other relevant aspects of the inferred output from the AI system.

Moreover, the right to an explanation under the AI Act "shall apply only to the extent" that the right is not covered under Union law, according to the AI AI Act Article 86 (3). If a data subject under the GDPR has the right to meaningful information about the logic involved in automated individual decision-making under Article 15 (1) (h) of the GDPR, the right to an explanation under Article 86 of the AI Act does not apply, according to Article 86 (3). The former section concluded that Article 15 (1) (h) of the GDPR contains a right to meaningful information about the logic involved in automated decisions and that this right contains a right to an ex-post explanation of individual decisions. The remainder of the legal analysis will presuppose that Article 15 (1) (h) of the GDPR provides a right similar to Article 86 (1) of the AI Act.

In which cases could affected persons rely on Article 86 (1) of the AI Act because their right to an explanation is not covered by the GDPR? The wording of Article 86 of the AI Act and the wording of Article 22 of the GDPR is not completely interchangeable. While Article 22 (1) of the GDPR covers a "decision based solely on automated processing", Article 86 (1) of the AI Act has a broader scope and covers decisions which are taken "on the basis of the output from a

high-risk AI system". In the Schufa Holding AG case [11], the CJEU has included profiling conducted by another entity as a decision under Article 22 (1) where the controller draws strongly on the conducted profiling. The wording of the GDPR "decision based solely" on automated processing could, thus, not be interpreted strictly by its wording and also covers situations where it is a market standard to draw strongly on a profiled or automated decision, even though the first decision is conducted by another entity. The AI Act Article 86 only covers decisions taken on the basis of a high-risk AI system. To illustrate the relationships between the right to meaningful information about the logic involved under the GDPR and the right to an explanation under the AI Act, consider these scenarios:

Imagine that an AI system is used for recruitment to filter applications and to analyse and evaluate job candidates. Such an AI system is classified as a high-risk AI system according to Annex III Section 4 (a). The system is fully automated and decisions to reject applications or accept them for interviews are based solely on the AI system. In this instance, the job selection process is based solely on automated decision-making under Article 22 (1) of the GDPR. Since the data subject has a right to receive meaningful information about the logic involved under Article 15 (1) (h) of the GDPR, the right to explanation under Article 86 (1) could not enter into effect, under Article 86 (3). When automated decision-making enters the scope of Article 22 (1) of the GDPR and the AI system applied is regarded as high-risk under the AI Act, Article 15 of the GDPR prevails.

Secondly, consider that an AI system is used solely to decide prices in an online e-commerce setting. The prices are based on the personal data submitted to the e-commerce site. According to guidance from the EDPB, such price setting could be regarded as a decision with similar effect under Article 22 (1) of the GDPR if the pricing discriminates certain people and groups from buying the product [12]. If such automated decision-making enters the scope of the GDPR Article 22 (1), the data subject can request an explanation about the logic involved under Article 15 (1) (h). However, the affected person does not have a right to an explanation under Article 86 (1) of the AI Act since the AI system used for price setting is not regarded as high-risk in accordance with Annex III of the AI Act. When automated decision-making enters the scope of the GDPR, but the AI system applied is not high-risk under the AI Act, the affected person has a right to an explanation about the logic involved under the GDPR and not under the AI Act.

Suppose that an AI system is used to influence the outcome of a local referendum. Such an AI system is regarded as a high-risk, according to Annex III 8 (b) of the AI Act. If such an AI system processes personal data, it could be argued that the automated decision-making does not legally or significantly factually affect a data subject, since the decision is a referendum and not a decision directed towards the data subject. Hence, the high-risk AI system in this specific use case does not enter the scope of Article 22 (1) of the GDPR. In this example, the affected person has a right to an explanation under Article 86 (1) of the AI Act, but not under the GDPR.

Consider an AI system that is high-risk under the AI Act, but not entering the scope of the GDPR. One example could include high-risk AI systems used for law enforcement purposes outside the scope of the GDPR. Another example could include high-risk AI systems not processing personal data. One example is if an AI system is used by a judicial authority to interpret the law, under Annex III 8 (a) and does not process personal data. In such an instance, the affected person has a right to explanation under the AI Act Article 86 (1) of the AI Act and not under the GDPR.

If an AI system infers a decision while not entering GDPR Article 22 (1) and is not regarded as a high-risk AI system under Annex III of the AI Act, the affected person does not have a right to explanation under either rule set. For instance, if a tax authority has a profiling system that flags individuals who are subject to manual inspection. It is possible to argue that the manual inspection is not a legal decision or a decision similarly affecting the individual, under Article 22 (1) of the GDPR. At the same time, the Schufa Judgement is not completely transferrable since it is not certain whether the authority "draws strongly" on the flag. Annex III does not list such an AI system as high-risk, meaning the affected person may not have a right to an explanation.

To conclude, the right to an explanation under Article 86 (1) of the AI Act covers outputs from high-risk AI systems in Annex III when the output forms the basis for a decision affecting natural persons. Such a right applies only to the extent that the right to an explanation is not otherwise provided in Union law, for instance in Article 15 (1) (h) of the GDPR. The examples above have covered situations where both the AI Act and the GDPR enter into effect and what this overlap in scope signifies for the right to an explanation. The next section will address how the right to an explanation in the GDPR and AI Act relates to various methods of XAI.

4 XAI Methods and the Right to Explanation under the GDPR and the AI Act

4.1 Explainable AI (XAI) Methods and Interpretable Methods

Based on the above legal analyses, the following sections will examine XAI methods in light of the right to an explanation under the GDPR and the AI Act.

XAI is an expanding and evolving research field [13–21]. One motivation behind the development of methods explaining the outputs of AI systems is to comply with regulations such as the GDPR and the AI Act [22, 23].

In the previous sections, it is established that an explanation under Article 15 (1) (h) is required to provide useful, reliable, and understandable information about the logic involved in automated decision-making. This explanation should provide the data subject with enough information to make the data subject able to assess the lawfulness of the processing of personal data in the automated decision-making process. An explanation under Article 86 (1) of the AI Act should be a "meaningful explanation of the role of the AI system in the decision-making procedure and the main elements of the decision taken".

Which XAI methods could provide such meaningful explanations under Article 15 (1) (h) of the GDPR and Article 86 (1) of the AI Act?

XAI is a term used for methods and ML models used for making ML models and their outputs understandable for natural persons [13]. In the XAI field, the terms interpretable and explainable are sometimes used interchangeably and sometimes used to denote different notions. In the next sections, the term interpretable will be used in line with the definition from Miller as "[t]he degree to which a human can understand the cause of a decision" [14]. Explainability will be applied as a term that relates to the interpretability of individual outputs from an AI system [13].

There are various methods to achieve XAI. Generally, it is possible to divide current XAI methods into intrinsic or post-hoc XAI methods [24, 25]. Intrinsic methods are interpretable on their own, due to their "simple" structure and self-explainable structure. Post-hoc methods apply such intrinsic methods on top of an uninterpretable AI method or utilise other methods to explain AI models that are not interpretable.

4.2 Instrinsic XAI Methods

Another typical manner to distinguish XAI methods is between XAI methods that make the whole trained AI model interpretable, and XAI methods that make the model output explainable. Since both Article 15 (1) (h) of the GDPR and Article 86 (1) of the AI Act revolve around the explainability of individual decisions, the next sections will focus on XAI models for model output explainability. Both intrinsic and post-hoc methods will be put under scrutiny.

An example of a potential intrinsic interpretable XAI method is to use "simple" models that are in themselves interpretable to draw inferences. One example is a type of supervised machine learning known as decision trees [13, 26]. A decision tree can, e.g., be applied to predict outcomes or classify data. When decision trees are applied as a supervised machine learning method, the algorithm discovers and represents the relationships between the data in the decision tree model [26]. The tree representation of the model makes the relationships between the different data interpretable and the individual output explainable, as long as the decision tree is not too large. One algorithm typically applied in decision trees, is the CART (Classification and Regression Trees) algorithm [27].

In short, the CART algorithm "builds" the three and the internal nodes by analysing how often a data point occurs in the training data. This is done until a pre-defining stopping criterion is reached. The CART algorithm decides the cut-off values by splitting the data into clusters of similar data and deciding which splits results into the most homogeneous, "similar", subnodes [27]. The decision on maximising the similarity in each of the two subnodes is made according to an index. In this index, the split on a specific data point results in the most different data in the two subnodes, of the data in the data set [26, 13].

When explaining an individual output of a decision tree, the deployer of a tree-based model starts in the inferred decision, the leaf node and goes back

in the tree model through the internal nodes to the start. To provide a textual interpretation, the different intermediate subsets are connected with "and". An intrinsic interpretable XAI method, such as a decision tree, thus, makes it possible with an explanation of the specific predicted output.

There are several other examples of intrinsic XAI methods [13]. However, to interpret whether intrinsic XAI methods comply with the identified legal requirements under Article 15 (1) (h) of the GDPR and Article 86 (1) of the AI Act, the logic established with the decision tree is sufficient.

The legal analysis of Article 15 (1) (h) of the GDPR established that the purpose of the access right is for the data subject to assess the lawfulness of the processing. Moreover, in order to assess the lawfulness, the explanation of the output of an individual decision-making process needs to be understandable, reliable, and have utility for the data subject.

Intrinsic models, such as decision trees, represent the relationships between the data points in a manner that corresponds to the inference being made. In contrast to post-hoc explanations, the explanations are therefore reliable. Moreover, provided that the decision tree is not too large, it is also a pedagogical and easily understandable explanation. Intrinsic explanations, provided that they are not too advanced, would represent explanations that comply with the objective and purpose of the right to receive meaningful information about the logic involved in individual automated decision-making in Article 15 (1) (h) of the GDPR.

In relation to Article 86 (1) of the AI Act, the purpose of the explanation is to provide a remedy for the affected persons who have been affected by high-risk AI systems in a manner that they consider have had an adverse impact on their health, safety, and fundamental rights. An explanation that explain the role the AI system has in the decision being inferred by a high-risk AI system, needs to – as long as it is feasible – be an actual representation of how the data is being processed within the AI system. An intrinsic XAI method such as a decision tree would thus be a method to explain an individual decision from a high-risk AI system in compliance with Article 86 (1) of the AI Act.

4.3 Post-Hoc XAI Methods

Post-hoc explanations, also referred to as model-agnostic XAI methods, separate the explanation of individual decisions from the model that is applied to draw inference [13]. The model that provides the interpretability and the explanations are put on top of the AI model that performs the tasks or inference. The next sections will address two such methods typically applied, Shapley values and local surrogate models, and examine if a post-hoc XAI method complies with Article 15 (1) (h) of the GDPR and Article 86 (1) of the AI Act. The reason behind the choice of these two methods is that they are local, meaning that they explain the individual decisions made rather than provide for the overall global model to be interpretable, and because they have a clear logic.

Shapley values is a theoretical concept from collaborative game theory used in the XAI field to provide an explanation of how much "influence" each parameter in the AI model has on the output of the model [28, 23] The best manner

to explain Shapley values is through the use of an example. Suppose that a natural person has applied for a loan and therefore has undergone a credit scoring process using ML. The credit score was 100 and negative and the application was rejected. An average person in the same neighbourhood, at the same age, and with similar income has an average credit score of 200 which will get the application accepted. An individual having their loan rejected would be curious about which of these features are most important for the output, which features they could improve to get their loan accepted, and – as in the example above why their application got rejected and their neighbours accepted. Shapley values is originally a method to calculate the division of a price between players that have won a game together based on how much each player contributed to winning the game [28].

In a deployed ML setting, the "price" refers to the individual inferred prediction, the "game" refers to the ML model, the gain refers to the actual predicted value minus the average predicted value, and the players are the different features in the ML model that "collaborate" to reach the specific output. The objective of the Shapley values is to explain the discrepancy between the inferred prediction, in the credit score example 100, and the average predicted credit score of 200 [13, 23].

Shapley values – building on collaborative game theory – is the average value of one player in a game, calculated on the performance of the collation with and without the specific player [28, 13]. In XAI, Shapley values make it possible to evaluate the value or contribution of specific features in the output of an ML model. This evaluation is done by calculating the average marginal contribution of the feature across possible coalitions with other features.

Suppose that years of education is a feature in the credit scoring. For simplicity, suppose that the credit score depends on three features: net salary, age, and years of education. We want to calculate the contribution of the *education* feature to the output of the credit scoring model. By selecting random data from the data set, *education* is replaced with random data points and the output is calculated. Then, the output of the model with and without the feature is calculated. However, the various features are interrelated and it is not just as simple as calculating the average of the education feature. The overall credit score would, e.g., be low even with a high education feature but with a low net salary. Thus, *education* needs to be interpreted in various coalitions with the other features, and the outcome of these different coalitions also needs to be averaged to calculate the Shapley value of the feature. This step is repeated and the average "value" referring to the increase or decrease in the output, the credit score, is repeated across different possible coalitions between features and values for education.

In terms of explainability, Shapley values are the average contribution of the feature across different coalitions and not the value of the feature if the feature is removed. Shapley values therefore offer explainability of individual decisions, but they are only an approximation of the feature's importance [13, 30].

Shapley values are an example of an explainability method that calculated feature importance, how important one feature is to reach a specific output of an ML model. Another method to explain individual outputs from black box ML models is local surrogate models. One such surrogate model is LIME (Local Interpretable Model-agnostic Explanations) [13, 31]. When it is not possible to interpret an ML model because it is a black box model, LIME make explanations possible by perturbing the input to the black box model and tests how the model performs around a specific output when the input is changed. This perturbed input data and corresponding output data from the model is used to create a model on top of the black box model that is intrinsically interpretable [31, 13]. The explanation of the individual decision from the black box model can then be interpreted by the intrinsically interpretable model trained on the input and output data from the uninterpretable model.

Are model-agnostic explanation methods acceptable explanations under Article 86 (1) of the AI Act and Article 15 (1) (h) of the GDPR? It was established under the German language version of Article 15 (1) (h) of the GDPR that the meaningful information about the logic involved must be "reliable". When using a surrogate model to explain another model, there is no guarantee that the explanation "matches" the processing conducted in the underlying black box model. The models on top of the black box model, such as the LIME method and Shapley values, are just approximations of input-output data in LIME and feature importance in Shapley values. However, the wording of both the GDPR and the AI Act in relation to the right to an explanation is open-ended. In the GDPR only information about the "logic involved" is required and in the AI Act, information about the role of the AI system in the decision-making is required to be given to affected persons. As a general conclusion both intrinsic explanations and post-hoc explanation methods comply with the right to explanation of decisions under the GDPR and the AI Act.

XAI is an evolving research field and the methods are becoming more advanced. As it becomes more technically feasible to explain individual automated decisions, the corresponding rights to receive such explanations, for instance in the GDPR and the AI Act, should evolve too. Today, these rights are open-ended reflecting the difficulty in explaining outputs from AI models. However, if such explanations are becoming more and more feasible as the XAI technology improves, a natural response is to specify and strengthen the right for natural persons to receive explanations of outputs inferred from deployed AI models, both under data protection law and in the AI Act.

5 Conclusion

The "right to an explanation" in the EU, the GDPR and the AI Act, are closely interrelated. Due to the open-ended wording of the two rule sets, both intrinsic and post-hoc explanations could be applied to comply with the requirements in Article 15 (1) (h) of the GDPR and Article 86 of the AI Act.

References

1. Regulation (EU) 2016 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 199/1.
2. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) OJ L 2024/1689.
3. Goodman, B. and Flaxman, S. EU regulations on algorithmic decision-making and a "right to explanation". *ICML Workshop On Human Interpretability In Machine Learning (WHI 2016), New York, NY. Http://arxiv. Org/abs/1606.08813 V1.* (2016)
4. Wachter, S., Mittelstadt, B. and Floridi, L. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law.* **7**, 76-99 (2017,6), https://doi.org/10.1093/idpl/ipx005
5. Selbst, A. and Powles, J. Meaningful information and the right to explanation. *International Data Privacy Law.* **7**, 233-242 (2017,12), https://doi.org/10.1093/idpl/ipx022
6. Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and of the free movement of such data [1995] OJ L 281/31.
7. EDPB Guidleines 01/2022 on data subject rights-Right of access., `https://www.edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-012022-data-subject-rights-right-access_en`
8. Parliament, E. Legal aspects of EU multilingualism. , `https://www.europarl.europa.eu/RegData/etudes/BRIE/2017/595914/EPRS_BRI%282017%29595914_EN.pdf`
9. Custers, B. and Heijne, A. The right of access in automated decision-making: The scope of article 15(1)(h) GDPR in theory and practice. *Computer Law and Security Review.* **46** pp. 105727 (2022), https://www.sciencedirect.com/science/article/pii/S026736492200070X
10. Directive (EU) 2016/943 of The European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure OJ/L 157/1.
11. Judgement of the Court (First Chamber) in Case C-634/21 OQ v Land Hessen and Schufa Holding AG ECLI:EU:C:2023:957
12. EDPB Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679 (wp251rev.01). , https://ec.europa.eu/newsroom/article29/items/612053
13. Molnar, C. Interpretable machine learning. (Lulu.com,2020)
14. Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. (2018)
15. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G. XAI—Explainable artificial intelligence. *Science Robotics.* **4**, eaay7120 (2019)
16. Speith, T. A review of taxonomies of explainable artificial intelligence (XAI) methods. *Proceedings Of The 2022 ACM Conference On Fairness, Accountability, And Transparency.* pp. 2239-2250 (2022)

17. Albahri, A., Duhaim, A., Fadhel, M., Alnoor, A., Baqer, N., Alzubaidi, L., Albahri, O., Alamoodi, A., Bai, J., Salhi, A. and Others A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*. (2023)
18. Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G. and Others Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys*. **55**, 1-33 (2023)
19. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J., Confalonieri, R., Guidotti, R., Del Ser, J., Diaz-Rodriguez, N. and Herrera, F. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*. **99** pp. 101805 (2023)
20. Islam, M., Ahmed, M., Barua, S. and Begum, S. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. *Applied Sciences*. **12**, 1353 (2022)
21. Linardatos, P., Papastefanopoulos, V. and Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*. **23**, 18 (2020)
22. Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J. and Others The role of explainable AI in the context of the AI Act. *Proceedings Of The 2023 ACM Conference On Fairness, Accountability, And Transparency*. pp. 1139-1150 (2023)
23. Hjelkrem, L. and Lange, P. Explaining deep learning models for credit scoring with SHAP: A case study using Open Banking Data. *Journal Of Risk And Financial Management*. **16**, 221 (2023)
24. Colaner, N. Is explainable artificial intelligence intrinsically valuable?. *Ai and Society*. pp. 1-8 (2022)
25. Vale, D., El-Sharif, A. and Ali, M. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI And Ethics*. **2**, 815-826 (2022)
26. Rokach, L. and Maimom, O. Data Mining with Decision Trees- Theory and Applications. (World Scientific,2014)
27. Crawford, S. Extensions to the CART algorithm. *International Journal Of Man-Machine Studies*. **31**, 197-217 (1989), `https://www.sciencedirect.com/science/article/pii/0020737389900278`
28. Shapley, L. A Value for N-Person Game. (1952), `https://www.rand.org/pubs/papers/P295.html`
29. Hausken, K. The Shapley value of coalitions to other coalitions. *Humanities And Social Sciences Communications*. **7**, 1-10 (2020)
30. Huang, X. and Marques-Silva, J. On the failings of Shapley values for explainability. *International Journal Of Approximate Reasoning*. pp. 109112 (2024), `https://www.sciencedirect.com/science/article/pii/S0888613X23002438`
31. Zafar, M. and Khan, N. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning And Knowledge Extraction*. **3**, 525-541 (2021)