Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making

©SHUTTERSTOCK.COM/VIKTORIA KURPAS

Digital Object Identifier 10.1109/MCI.2021.3129960
Date of current version: 12 January 2022

Ronan Hamon, Henrik Junklewitz, and Ignacio Sanchez
European Commission, Joint Research Centre, ITALY

Gianclaudio Malgieri
EDHEC Business School, FRANCE

Paul De Hert
Vrije Universiteit Brussel, BELGIUM

Abstract—Can satisfactory explanations for complex machine learning models be achieved in high-risk automated decision-making? How can such explanations be integrated into a data protection framework safeguarding a right to explanation? This article explores from an interdisciplinary point of view the connection between existing legal requirements for the explainability of AI systems set out in the General Data Protection Regulation (GDPR) and the current state of the art in the field of explainable AI. It studies the challenges of providing human legible explanations for current and future AI-based decision-making systems in practice, based on two scenarios of automated decision-making in credit scoring risks and medical diagnosis of COVID-19. These scenarios exemplify the trend towards increasingly complex machine learning algorithms in automated decision-making, both in terms of data and models. Current machine learning techniques, in particular those based on deep learning, are unable to make clear causal links between input data and final decisions. This represents a limitation for providing exact, human-legible reasons behind specific decisions, and presents a serious challenge to the provision of satisfactory, fair and transparent explanations. Therefore, the conclusion is that the quality of explanations might not be considered as an adequate safeguard for automated decision-making processes under the GDPR. Accordingly, additional tools should be considered to complement explanations. These could include algorithmic impact assessments, other forms of algorithmic justifications based on broader AI principles, and new technical developments in trustworthy AI. This suggests that eventually all of these approaches would need to be considered as a whole.

I. Introduction

Artificial Intelligence (AI) [1] has become increasingly important for many technological applications. To ensure that its use will benefit society as a whole, institutions, industry and human rights organizations, among others, are actively seeking robust regulation. This is especially crucial in the context of automated decision-making systems, which make decisions through technological means without human involvement [2]. Their deployment in application domains such as banking, employment, or healthcare, is raising concerns for the potential impact on the fundamental rights of citizens. The General Data Protection Regulation (GDPR) [3] introduced a set of measures to protect those rights for individuals in the European Union, when they are subject to automated decision-making systems involving the processing of personal data. Among them, the requirement for data controllers to provide meaningful information about the logics of the decision-making process and a justification of the outcomes, has opened an argument in legal circles of whether a right to explanation for data subjects can be inferred from the GDPR [4].

Many contemporary automated decision-making solutions rely on established practices of encoding expert knowledge [5], and consist in the processing of inputs based on pre-defined rules. Typically, the algorithm at the core of the decision-making process is well specified and documented, and the logic is known and can be audited to evaluate its compliance with respect to existing regulations. This situation is evolving rapidly with the uptake of advanced AI techniques, and in particular with the use of machine learning techniques that rely on implicit decisions rules extracted from large collections of data. With the increasing complexity brought by these techniques, the question of how much the outputs of a given algorithm are still understandable for a human, or are uniquely explainable, has evolved. Requiring AI-based decision-making systems to comply with a right to explanation for the data subject raises questions about the concrete implementation of explanatory mechanisms. The outstanding achievements of AI systems are counterbalanced with their complexity that significantly reduces the legibility of the logics involved in the process. Even if the scientific AI community is actively developing methodologies to encourage explainable and interpretable AI [6], evaluating the relevance of explanations returned by such approaches from a legal perspective, with regard to the right to explanation as the one derived from the GDPR, is still lacking a sound basis.

The aim of this study is to contribute to bridging this gap and to explore to what extent current and future AI systems can provide adequate explanations that would be admissible from a legal point of view, following the applicable EU data protection framework. This analysis reflects on the meaning of an explanation with regard to what constitutes its legibility. It also acknowledges the importance of the context in which the system is operating to judge its relevance and adequacy [7], [8]. The study is carried out using two use cases, which are based on typical scenarios for high-risk AI-based decision-making in credit scoring and medical imaging. These two applications, often mentioned in legal arguments around the GDPR [2], highlight the current technological shifts, from classical machine learning systems to much more sophisticated deep learning models, and from tabular data to high dimensional data (e.g., images). These transitions have a significant impact on the understanding of the mechanisms at play in the decision-making processes, in fact increasing the opacity of systems. They serve as an idealized yet plausible test-bed to base the legal discussion on firm technical grounds. This discussion is even more relevant in light of current policy initiatives regarding the regulation of AI, as illustrated by the recent proposal published by the European Commission [9].

This work is the in-depth continuation of a preliminary study [8], which reasoned about AI explainability and legal

Corresponding author: I. Sanchez (e-mail: ignacio.sanchez@ec.europa.eu). R. Hamon, H. Junklewitz and G. Malgieri all contributed equally.

requirements on the basis of the medical imaging use case. The conclusion was that current trends in machine learning introduce a level of complexity that is hard to reconcile with existing legal requirements, especially with regard to human legibility of explanations for non-expert data subjects. Furthermore, the paper reasoned that the quality of possible explanations might not be judged as adequate under the provisions of the GDPR. In this work, these preliminary findings are strengthened by systematically exploring the technical challenges linked to the usage of AI techniques in decision-making systems through two use cases. The novelty of this study lies in the interdisciplinary exploration of technical and legal issues to tackle a number of open questions and more general problems, including:

- ❏ What are the main challenges to providing explanations regarding the realities of contemporary AI?
- ❏ What are legal and technical implications if explanations cannot be provided, or can only be provided partially according to the legal requirements from the GDPR?
- ❏ How could a right of explanation be embedded and enforced in the context of auditing and regulatory processes such as an impact assessment?
- ❏ How could a regulatory setting be devised to compensate for the technical limitations?

The outline is as follows: Section II introduces the necessary background on legal requirements of automated decision-making and AI explainability techniques. Section III presents the core analysis of the challenges using the use cases as illustrative examples. Section IV exposes the doubts about the feasibility of a monolithic approach to requiring explanations in the future AI-based environment, and proposes some alternative approaches to mitigate the situation.

II. Background

In this section, background elements regarding the legal requirements set out in the GDPR on automated decision-making are provided, as well as an overview of the field of Explainable AI (XAI) that aims to complement AI systems with meaningful explanations of their mechanisms.

A. Legal Requirements on Automated Decision-making

The GDPR [3] came into application in May 2018 in the European Union (EU) and introduced meaningful novelties in EU legislation about automated decision-making systems based on the processing of personal data (information related to an identified or identifiable individual). Considering the large use of personal data in AI applications and the specific data protection rules about explainability and accountability of automated decisions, it would appear that the GDPR is, currently, the most advanced, robust, and comprehensive legal tool to regulate the use of personal data in AI. One of its main goals is to protect fundamental rights and human rights of individuals when they are subject to—among others—profiling and automated decision-making [10]. Looking at the text of the GDPR, the definition of profiling given in Article 4(1)(4) and the general principles of transparency, lawfulness, fairness, accuracy, and accountability (Article 5) seem very relevant elements for trustworthy AI applications.

Moreover, Article 22 provides a right not be subject to decisions based solely on automated decisions that produce legal or similarly significant effects, with three exceptions [11]: the consent of the data subject, the necessity for the performance of a contract, or an EU or national law that regulates specific cases of automated decision-making. In case these automated decisions are taken, the data subject should have always the rights to: contest the automated decision [12]; express their own view; have a human intervention in the automated decision-making systems [13]. In addition to these rights, recital 71 clarifies that—depending on the level of risk of the algorithmic application—the safeguards for data subjects should also include the auditing of the algorithm and the right to receive an explanation about the individual decision taken. The existence and modalities of such a right to an explanation have triggered an intense discussion among legal and technical scholars [4], [14]–[16] and led to its implementation by several EU member states in their national laws [17]. In addition to these explanation requirements, Articles 13(2)(f), 14(2)(g), and 15(1)(h) of the GDPR require that data subjects are informed about the existence of automated decision-making processes and about the underlying mechanisms (logics) behind the automated decision-making performed, as well as the significance and potential consequences of such processing [18].

Different levels of explanation can be identified: a general one about the logic of the AI system as a whole, and more granular explanations about how the AI application works in particular situations (e.g., for some specific cluster or group of individuals), or even in individual cases. These kinds of explanations can be also relevant to better exercise the right to contest the output of the algorithm and the right to human intervention. Indeed, the guidelines released on this topic by the European Data Protection Board (EDPB) state that human intervention implies that the human-in-the-loop should refer to someone with the appropriate authority and capability to change the decision [2]. It is clear how a requirement of technical explainability is relevant for these envisaged safeguards. Human supervision can only be effective if the person reviewing the process is in a position to assess the algorithmic processing carried out. This implies that such processing should be understandable. Furthermore, explainability is also key to ensuring that data subjects are able to express their point of view and are capable of contesting the decision. As stated in the EDPB guidelines, data subjects will only be able to do that if they fully understand how the automated decision-making was made and on what basis. In summary, considering the different levels of risk that automated decision-making can imply for rights and freedoms of data subjects, some scholars proposed the following multi-layered explanation model [16]: for low-risk scenarios, a general layer of explanation about the algorithmic functioning might help (inferred from Articles 13-15 GDPR); for medium-risk scenarios, an intermediate and group-based explanation of

the algorithm might be advisable; for high-risk scenarios, an individual explanation about the decision reached in a specific case might be necessary (according to recital 71).

Although the definition of personal data is so broad that the GDPR has been considered the "law of everything" [19], a holistic regulation of AI is still missing. A proposal for a comprehensive EU Regulation of AI [9] is currently being discussed based on the notion of 'Trustworthy AI', which incorporates both explainability and transparency as mandatory elements among a list of other important principles such as robustness, security, fairness, or safety [20]. In this context, the technical notions of explainability and transparency become crucial requirements to ensure trust in AI-based decision-making systems.

B. Explainable Artificial Intelligence (XAI)

Making AI systems more understandable first requires making them transparent, not only regarding the internal implementation of the models (e.g., parameters), but also with respect to the full development process (design choices, data uses, testing procedures, etc.) [21], [22]. Technical transparency alone is not enough to explain AI models that are often described as opaque systems, where relationships between inputs and outputs are beyond understanding. The concepts of explainability and interpretability have been introduced to describe how well a human could understand the decisions of an automated algorithmic system [7], [23], [24]. Albeit used interchangeably, explainability is mostly defined as being a subject-centric notion, whereas interpretability would be an AI model-centric one, mostly employed in the narrower context of machine learning models, and thus can be seen as a subset of the broader notion of explainability [23], [25].

The field of Explainable AI [26], [27] has recently gained a lot of traction, aiming to provide interpretable elements alongside predictions. It ranges from a growing literature of technical work on interpretable models and explainable AI [28]–[30], to an ongoing discussion about the precise meaning and definition of explainability and interpretability [6], [25], [31], and to more procedural questions about the evaluation of existing frameworks [24]. The meaning of what the technical literature refers to as explainability of an AI model is very different from the meaning of an explanation which is generally discussed in other social contexts (see [7], [23], [24], [31] for the ongoing academic discussion) or the broader notion of transparency put forward in similar contexts [20]. In practice, a distinction is made between approaches aiming at explaining the general mechanisms of a model (global explanation) and those explaining a decision on a particular instance (local explanation) [28], in line with the multi-layered explanation model previously discussed [16].

Machine learning models are typically categorized into interpretable and non-interpretable models. Interpretable models are designed to provide reliable and easy-to-understand explanations of the prediction they output from the start [29], [32]. Their interpretable nature comes from their simplicity, either by using only few parameters or operations, or by being conceptually understandable. Examples include linear and logistic regressions, decision trees or generalized additive models such as spline fitting. Non-interpretable models require the use of post-hoc techniques to generate explanations of decisions. This can be through the construction of interpretable surrogate models [28], or by extracting interpretable elements from a range of specific processing techniques. A classical way to provide insights about which parts of the data drive a particular behavior is to compute the contribution of features in the decision-making process, i.e., evaluating how much a decision is influenced by specific attributes of the input data (or features). For some interpretable models, this could be straightforwardly derived from the model weights, but dedicated post-hoc approaches have been proposed to evaluate the significance of data features in a more robust and systematic way, also applicable to more opaque models. This includes techniques such as stochastic permutation methods, local surrogate approximations like LIME [33], or the SHAP values [34], a popular technique using game-theoretic arguments to identify the effect of a single feature on an individual entry in comparison to the average model output computed on training data. Their use though is often limited to relatively low-dimensional problems, as they require a minimum computation capability that is not reachable for complex models. Explaining deep learning models requires specific techniques [27] that are capable of handling the high dimensionality of data, and the high number of parameters typically present in those models. These techniques should be often adapted to the type of data involved. For instance, image-based models often rely on visual maps over the image to highlight important patterns [35].

A major caveat of current approaches is the lack of interactivity of explanations, that cannot take into account the diversity of contexts [7] or adapt to the technical background of the recipient. However, research into explainable AI is very active and ongoing developments [28], [36]–[41] can be constantly observed, with some expectations that such limitations might be successfully addressed. Current EU legislation, and specifically the above-discussed GDPR provisions, have had a positive impact on the field of explainable AI, leading to an increasing number of publications specifically aiming at providing GDPR-compliant explanations [39], [42]–[44]. These provisions also introduce additional challenges for research, as the GDPR was mostly thought for 'tabular data' processing. For complex AI systems, the implementation of general safeguards (explanation, contestation, human-in-the-loop) appears more problematic.

III. Challenges of AI for explanations: insights through high-risk use cases

In this section, the broad technical challenges that question the feasibility of AI-based algorithmic explanations are explored and discussed with respect to legal requirements, by means of two use cases on credit scoring and medical imaging. As illustrated in legal discussions around the GDPR [2], both cases are prominent high-risk scenarios for which automated decision-making can have significant and immediate negative impacts on individual rights.

For each use case, one or several decision-making systems have been built based on machine learning techniques widely

used in current applications. These implementations, including the choice of techniques, models and data sets, are not meant to achieve the performance of the state of the art for the considered tasks, but rather to represent an accurate picture of the kind of tools used in such contexts. As such, the tasks are addressed only to provide a baseline for our interdisciplinary analysis, with no claim of generality or exhaustiveness.

Use case 1: Credit Scoring

Banking has been an early field of application for machine learning methods since the 1980s [45], [46] and can nowadays be considered one of the typical, mature fields where such methods are regularly employed in a business context. The need to learn from large databases of customer entries finding patterns or anomalies, has proven to be a suitable application for predictive machine learning models.

TABLE I Performance measures (in %) for the models trained on the FICO credit scoring data set.

Model	Accuracy	Precision	Recall	ROC
Decision tree (DT)	71	68	74	76
Random forest (RF)	73	72	71	80
Linear regression (LR)	73	71	73	80
Multilayer perceptron (MLP)	73	71	73	80

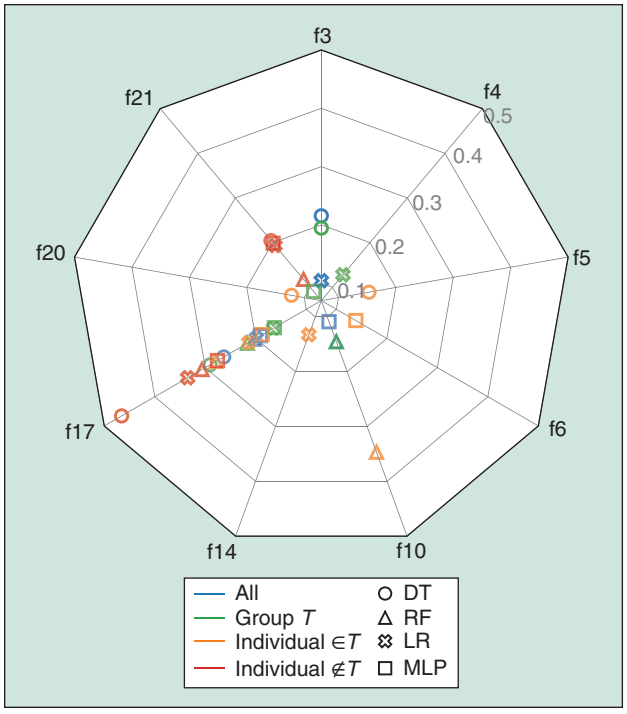

FIGURE 1 Importance of the two most significant features in the final decision for each of the models using SHAP values. Names of the features can be found in Figure 2(a). T denotes the group of frequent traders.

For reference and legal discussion of our findings, the typical real life scenario for this use case is the following:

> *A customer of a loan-lending company applies for a small HELOC (a specific type of loan where the borrower gets a credit using a house as collateral) dedicated to a major item such as debt consolidation, education or home improvement. He/she has to fill a questionnaire about his/her financial history and other personal information. The company informs the customer that an automated process is triggered during which the loan application will be evaluated using AI techniques.*

This scenario is very relevant in terms of data protection, as receiving an automated denial of a loan request should be under the scope of Article 22. It is indeed a decision based solely on automated processing and on profiling, which produces legal effects or similarly significantly effects on the data subject. Considering Article 22(2)(1), this automated decision-making system might be allowed because it might be considered necessary for the performance of the contract between the bank and the customer. Given that automated credit scoring might be considered data processing producing high risks for the rights and freedoms of individuals, according to the risk-based approach, the data controller might be required to implement higher safeguards, in addition to the three safeguards mentioned in Article 22(3) (see Sec. II-A).

The technical implementation of an automated decision-making system is based on the 'HELOC FICO data challenge' [47], which consists in anonymized data of HELOC applications made by US homeowners. The data set serves as a reference benchmark for the application of explainable AI methods in a high-risk automated decision-making context [32], [48]–[50]. It is made of entries from 10459 customers, ranging from account details to detailed information about the past financial transactions within the customer's credit line.

The two classes to be predicted can be qualitatively described as *high risk of future default* and *low risk of future default*. Four classical predictive models have been trained to solve this task: a decision tree (DT), a random forest (RF), a logistic regression (LR), and a multilayer perceptron (MLP).

Explanations should ideally be tailored to the need of the situation and the expected risk associated with the decision [16], [37], [41]. As a very basic form of an explanation, the performance metrics of the various models (see Table I) can give a first overview about the automated decision-making system in different situations. Furthermore, in Figure 1, a comparison of multi-layered explanations is displayed by estimating feature significance using SHAP values. The figure shows the importance of the two most significant features as a scatter plot over the features, for the decision taken by each of the four different models, and for four different layers. The four populations making up the layers of explanations comprise all individuals in the data set (global explanation), the most frequent traders defined as having at least 40% of the maximum trade frequency in the data (group-based explanation), and two individual customers, one belonging in the group of frequent traders and one not belonging to it (individual explanation). These explanations could be used to describe the typical behavior of the system depending

on the input data, for example as general information for an expert ex-ante, or, to justify a decision ex-post on an individual applicant's loan in response to a contestation.

Use Case 2: Medical Imaging

Computer-aided software has been used for decades in the medical environment to support medical professionals. The recent progress of AI suggests that these systems may become more and more autonomous in their decision-making, and help to automate some aspects of the medical decision-making process that are typically done by medical staff, to aid the diagnosis and facilitate care handling of patients. The growing integration of AI capabilities in automated decision-making gains particular relevance in light of the current COVID-19 pandemic, during which an uptake of AI in the health sector has been observed [51]–[56] with the aim to compensate for the limitations of health systems to handle exceptional high-stress situations. This includes various forms of automated triage consisting in defining the priority of a patient for immediate intensive care. These works are part of a bigger trend in machine learning to provide detection tools that not only achieve human performances, but also go beyond and are able to capture weak patterns in complex images not detectable by humans.

As introduced in [8], the following hypothetical medical scenario is considered:

> *A patient with COVID-19 symptoms admits himself/herself to the emergency ward of a hospital during a severe, ongoing local outbreak of the COVID-19 pandemic. The nurse tells the patient that blood tests indicate a possibility of a COVID-19 infection. Normally, such patients would be admitted after decision of a medical doctor to the intermediate care ward since a pneumonia induced by COVID-19 is likely to cause severe and rapid worsening of the patient. However, no doctor is available for at least several hours to make the decision. Instead, the patient could decide to let an automated decision-making system that has been exceptionally authorized to decide from a chest X-ray image, whether he/she is likely suffering from COVID-19.*

Receiving an automated diagnosis of COVID-19 falls under the scope of Article 22 as it is a decision based solely on automated processing and on profiling, which produces legal effects or similarly significant effects on the data subject. In particular, at least two effects can be identified: the psychological effects deriving from a diagnosis, and the medical consequences of such diagnosis (access to further healthcare, self-quarantine, etc.). This AI application might be allowed under Article 22 because either it is based on the explicit consent of the data subject, or it is "authorized by Union or Member State law". Accordingly, all measures mentioned in Article 22(3) should be implemented by the data controller, and the data subject has a right to receive a "general explanation" before the processing and upon his/her eventual request at a later stage. As stated in Section II-A, the higher the risk, the more specific and comprehensive the explanation of the automated decisions should be. Thus, since COVID-19 automated detection/diagnosis might be considered a data processing producing high risks for the rights and freedoms of individuals, the data controller might be required to implement stronger safeguards, including the right to receive an explanation about the decision reached.

The technical implementation of such a decision-making system is based on the deep learning architecture ResNet-50 [57], which was trained for COVID-19 detection following a methodology described in [56]. The automated decision-making system returns a probability of the patient being infected based on an X-ray image of the chest. The system achieves a global accuracy of 99% on previously unseen data, with a sensitivity of 85% to COVID-19, which is in line with published results for this kind of models [56]. A complementary study presenting a set of typical multi-layered explanations for this technically involved scenario is given in [8], providing examples ranging from informative ex-ante explanations of the system to counterfactual ex-post explanations of an individual decision.

A. Challenge 1: The Trend to Complexity in Machine Learning

In machine learning, a clear trend to higher complexity can be observed, with three major causes:

- ❏ data sets feeding AI systems are getting more and more heterogeneous and high-dimensional;
- ❏ models are made of compound architectures including a growing number of parameters;
- ❏ algorithms and techniques used for the development of models are getting increasingly sophisticated.

The use of XAI techniques to generate sound explanations adds an additional layer of complexity, as they can be themselves hard to understand.

1) Complexity of Data

Data sets are often referred to as the new oil to highlight their crucial role in the development of AI systems. Although relevant in some contexts, this analogy fails to transcribe the complexity and diversity of data (images, sounds, texts, tabular data, graphs, etc.) that significantly differ from the multi-purpose crude oil. Hence, a data set is only helpful to solve a limited range of applications in a given context. Historically, tabular data sets, consisting of entries composed of pre-defined text, value or category fields, have been predominantly used in early machine learning applications. The credit scoring data set is a typical example of such type of data (see Figure 2a).

The digitalization of equipment and the increased capacity of storage have led to the creation of bigger data sets that have grown to reach hundreds of thousands, or even millions of entries. Besides, it has led to the apparition of large collections of heterogeneous data, such as images, sounds, or texts, which differ from tabular data sets by their high dimensionality. For instance, an X-ray image from the medical imaging scenario is made of thousands of pixels providing a spatial representation of an organ, possibly including several color channels (see Figure 2b).

2) Complexity of Models

Models are at the core of machine learning systems, and consist in transforming input data into predictions using simple operations. For example, the logistic regression applied to the

FICO credit scoring data set consists in a weighted sum of all data features, followed by the application of the logistic function to return a probability of risk of default. The weights used in the sum, called parameters of the model, hold the values of the model, and define how the features will influence the outcomes. To address more challenging tasks, the architecture of models has evolved to represent more and more sophisticated relationships between inputs and outputs by increasing the number of operations and parameters. Deep learning methods are representative examples of this trend, with the multiplication of relatively simple layers of operations stacked together in flexible architectures designed to solve complex tasks.

Figure 3 displays a decision tree, considered as an interpretable-by-design model, trained on the FICO credit scoring

f1	ExternalRiskEstimate	72	72
f2	MSinceOldestTradeOpen	365	84
f3	MSinceMostRecentTradeOpen	11	4
f4	AverageMInFile	83	53
f5	NumSatisfactoryTrades	22	29
f6	NumTrades60Ever2DerogPubRec	0	0
f7	NumTrades90Ever2DerogPubRec	0	0
f8	PercentTradesNeverDelq	100	100
f9	MSinceMostRecentDelq	–7	–7
f10	MaxDelq2PublicRecLast12M	7	7
f11	MaxDelqEver	8	8
f12	NumTotalTrades	22	30
f13	NumTradesOpeninLast12M	1	2
f14	PercentInstallTrades	41	27
f15	MSinceMostRecentInqexcl7days	0	0
f16	NumInqLast6M	0	1
f17	NumInqLast6Mexcl7days	0	1
f18	NetFractionRevolvingBurden	49	43
f19	NetFractionInstallBurden	23	48
f20	NumRevolvingTradesWBalance	6	9
f21	NumInstallTradesWBalance	2	5
f22	NumBank2NatlTradesWHighUtilization	1	3
f23	PercentTradesWBalance	73	64

(a)

(b)

FIGURE 2 Example of input data for the two use cases. (a) FICO credit scoring data set (negative: red; positive: green); (b) COVIDx data set.

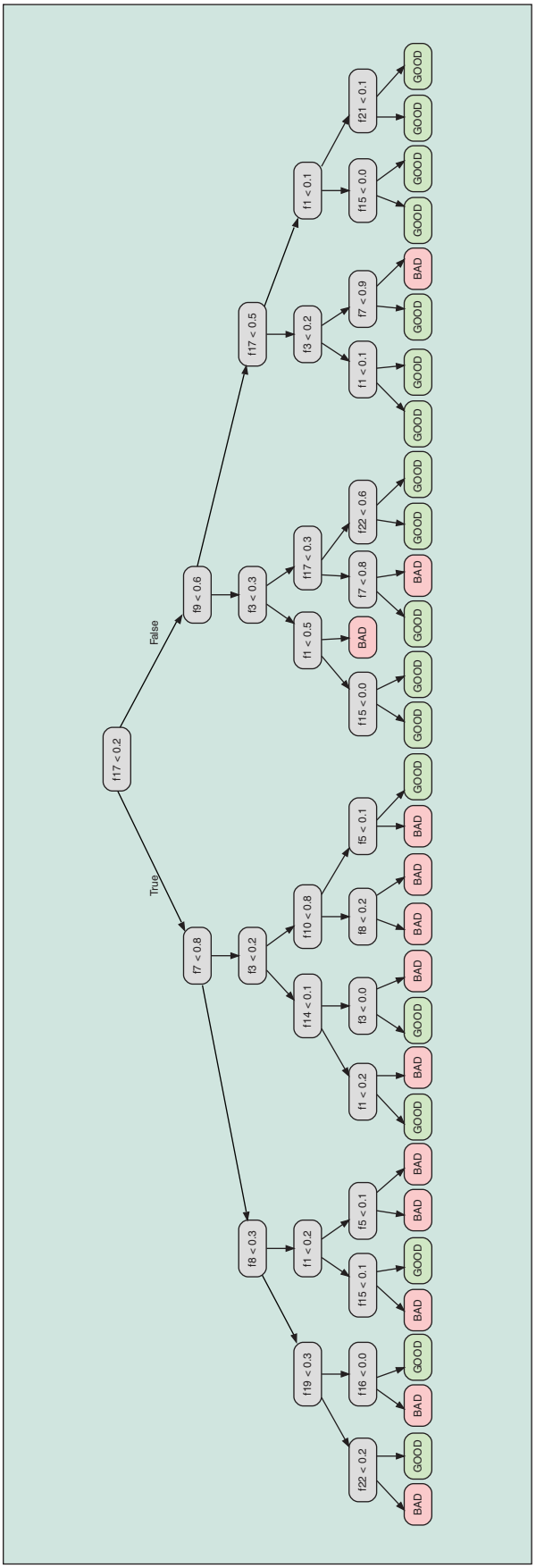

FIGURE 3 Decision tree trained on the FICO credit scoring data set. Names of features are given in Figure 2a.

data set: starting from the root node on the top, the tree is browsed following the branches according to the rules present on each node, until reaching a final decision. Despite their simplicity, their explainability can be questioned: even a simple decision tree as the one displayed in Figure 3 may already be too complex for providing legible explanations. Furthermore, the choice of an interpretable model might not always be possible. The complexity of a task may quickly limit the predictive power of interpretable models, and may require the use of more sophisticated and less interpretable models, such as random forests (see a larger discussion on the trade-off between performance and explainability in Section III-B).

Deep learning techniques also pose significant issues in terms of interpretability: the ResNet-50 architecture used for the medical imaging use case involves a succession of operations that includes hundreds of thousands of parameters, which exceed any reasonable limit on direct interpretation of predictions. Figure 4 displays some features that are captured by the model at different layers. In deeper layers of the architecture, images have smaller size and exhibit complex patterns that are not easily understandable or even readable by human auditors, even with medical knowledge.

3) Complexity of AI Algorithms

Developing an AI system requires undertaking a series of steps, known as the AI life cycle, each of them including a number of processes that may hinder the capacity of the final system to be explainable to end-users. At the development stage, three specific steps are of particular importance: data processing, training and evaluation.

Data processing aims firstly at cleaning all possible issues that may arise in the acquisition of data (missing values, text embedding, etc.), and secondly, at preparing the data for the training of models. For feature-based models, this last step consists in the extraction of features. The FICO credit scoring data set is for instance made of 23 handcrafted features, including the number of transactions or the percentage of unsuccessful transactions (see Figure 2a for the complete list of features). For deep learning based models, the features are learned by the model at the training stage, but pre-processing may involve the generation of additional samples, for instance by applying rotation, translation, color change, etc., to augment the training set and increase the accuracy of models.

The training stage consists in fitting the model to the data by adjusting the parameters to obtain a good prediction accuracy. This mainly relies on gradient descent optimization schemes, where parameters are iteratively updated such that the error between ground truths and predictions is minimal. This step implies the selection of a number of training parameters (e.g., number of iterations, thresholds, etc.) as well as additional settings of the models (e.g., the number of trees in a random forest), denoted hyperparameters. The choice of the set of hyperparameters has a strong influence on the performance of the resulting model, and is optimized using techniques such as cross-validation, bagging and bootstrapping [58].

Once a model is trained, its performance on previously unseen data is evaluated using metrics, whose interpretation may also be deceptive. In binary classification for example, a standard metric is accuracy, which measures how often the system is correct. It may be tempting to use measures to compare different systems, but the obtained values should be considered with caution. For example, if only 1% of samples are positive, then a system only outputting a negative prediction has an accuracy of 99%, which may suggest that this uninformative model is performing well. More generally, once a problem is getting harder, the more difficult it is to define good metrics, and interpret them correctly [59].

Development of AI systems is mostly driven by techniques with limited human intervention, especially in the case of deep learning techniques. Importantly, all these steps limit the capacity to 'reverse engineer' the results, preventing an efficient algorithmic auditing, such as an Algorithmic Impact Assessment, to study the impact on the resulting decision.

4) Complexity of Explanatory Techniques

Explanatory techniques are also subject to an increased complexity with a direct impact on the legibility of explanations for data subjects. They are indeed presented as a result to describe, justify, or explain a decision. In many cases, the employed technological solutions are in principle well known for practitioners of machine learning, especially in classical applications such as the FICO credit scoring scenario. The performance metrics mentioned previously (see Table I) are a

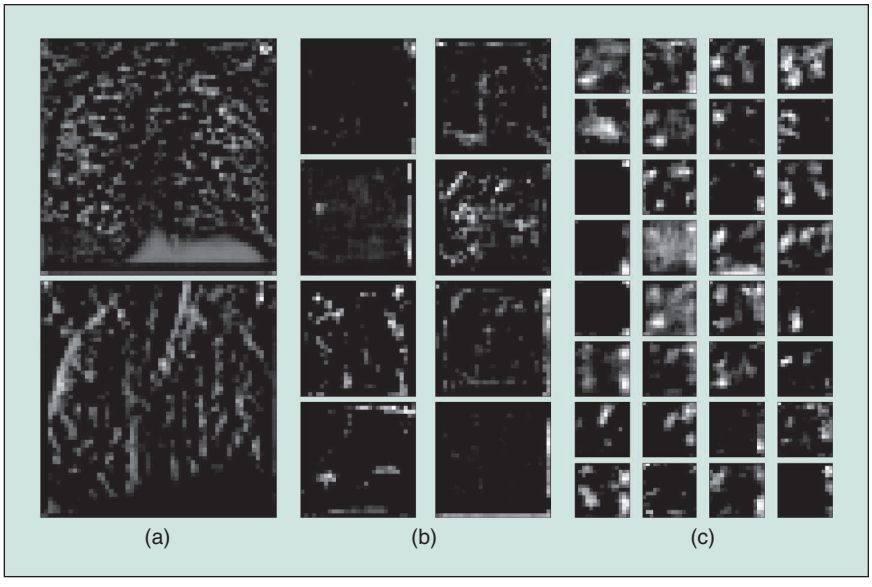

FIGURE 4 Internal representation of the X-ray image displayed in Figure 2b at a selected channel. (a) Layer 2; (b) Layer 3; (c) Layer 4.

standard way of assessing the validity of models, by giving a rough estimate of the quality of results, arguably, even for non-experts. However, Figure 1 highlights how the combined complexity of data, models and methods can also lead to ambiguous results. Different models rely on different features for the decision-making. Even for the same model, the mechanisms at play vary according to the population that is considered, and for two individuals belonging to different groups, the important features may significantly differ. This raises the question of the reliability of the decision that is taken, and also of the reliability of the explanation itself. The use of SHAP values serves as an example of increased methodological complexity, and other approaches might provide different outcomes, especially when stochastic behaviors are involved.

Following the generally presented trend, the complexity of these techniques tends to be aligned with those of the models. Techniques introduced for tabular data are in practice unusable for explaining deep learning models such as the one implemented for the COVID-19 scenario. Instead, specific techniques have to be used, such as occlusion maps [35] (see Figure 5), which consists in masking a portion of the image, and tracking the evolution of the prediction score. If the score increases, then the masked region contains elements indicating that the patient is not infected, and vice versa. With this type of information, abnormal behaviors could be detected, for example a detection that relies on regions with bone tissue. Other typical examples would be so called gradient maps or counterfactual images (see [8] for an illustration on the medical imaging scenario). For all these techniques, there is nonetheless no guarantee that indicated regions are indeed the ones that have been used, nor that they are indeed important for the final decision of the model. In particular, compensation effects may occur: a group of pixels may belong to a pattern that is indicative of an infection, while belonging to another pattern specific to negative samples at the same time.

Many feature importance methods are reasonably interpretable for an expert in AI, but already less so for an average data controller, and even less for a data subject aiming to understand the system behavior. This means the user should either trust an opaque method, or should be required to have explicit domain knowledge to understand the underlying mathematics. Furthermore, these techniques require fine parametrization, and

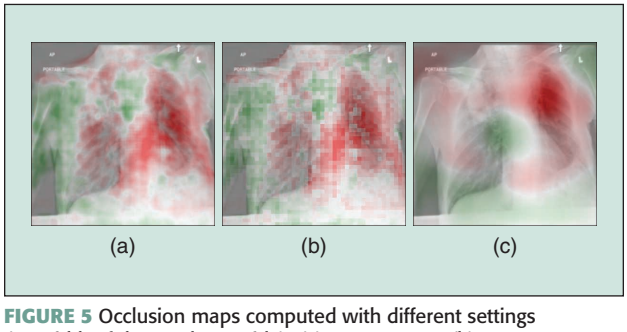

FIGURE 5 Occlusion maps computed with different settings (w: width of the mask, s: stride). (a) w = 5, s = 1; (b) w = 5, s = 5; (c) w = 30, s = 1.

are subject to a trade-off between accuracy and tractability. Occlusion maps, for instance, require selecting a size for the mask, and a step size for moving the mask all over the image. These parameters have a great influence on the final maps, and the computation required for generating them (see Figure 5).

B. Challenge 2: The Trade-off between Accuracy and Explainability

The relationship between interpretability and accuracy is a recurring topic in the machine learning community when discussing explainability of AI models. Making models explainable either by using inherently explainable models or by using post-hoc techniques is often expected to degrade the accuracy of the model. The scientific discussion on this topic is built on the requirements of two desirable properties of systems that are their effectiveness, i.e., their ability to perform various tasks with fewer mistakes, and their understandability, i.e., their capacity to provide interpretable elements of their inner mechanisms involved in the processing of data. The pursuit of these two objectives has actually proven to be contradictory, as interpretable methods often require constraints that may limit their accuracy, such as reducing the number of parameters. The effects of this trade-off have been particularly strong in recent years with the advent of deep learning techniques, in a context where increasing the complexity of models has been employed to obtain higher accuracies, or to solve more demanding tasks (see Section III-A). Making these models more interpretable in turn seems to come almost inevitably with a reduction of these high predictive capabilities.

A straightforward example is provided by looking at the performance metrics of the models used in the credit scoring scenario (see Table I). A gap is visible between more interpretable but less accurate models, such as decision tree or logistic regression, and more accurate but less interpretable models, such as random forests. On the other hand, the assumption that under given constraints, better results can only be achieved with a more complex model can be challenged, especially when good feature engineering is combined with simpler, but robust models [29]. Yet from another angle, the very notion of a complex and less interpretable model might depend on the point of view, constraints or situational context [31].

C. Challenge 3: Correlation or Causality?

Most machine learning models in use today are purely predictive models, based on known relationships between input and target variables. This effectively means that the uncovered relationships are solely based on the correlations found in the training data. This is a direct result of a design choice favoring predictive modelling over interpretative or explanatory modelling [60]. The latter is often the case in scientific statistical modelling, where the modelled relationship itself is of interest, for instance when describing a natural law, or when the goal of the modelling is to learn model parameters that describe reality. Conversely, predictive modelling is not concerned with the fitted parameters themselves, but with the accuracy of predicting

the correct results for future data, which for many purposes is the main task of contemporary narrow AI models. This means that in most machine learning models, there is no guarantee that the found correlations correspond to causal relations, only that they are maximizing a predictive score over the training data. It also leads to the well-known multiplicity of equally well accurate models, which can all be correctly predicting outcomes, with no way to properly assess which model might be the correct one, in an explanatory sense [61]. This is especially true for modern, large deep neural network models trained on huge data sets, whose model function is specifically designed to be able to represent any function, and which in turn usually have too many parameters for meaningful explanatory modelling.

Some of the features in the credit scoring data set show a strong correlation with others, and it is not clear what the underlying cause might be. Since many of the simpler explanatory methods rely on feature significance metrics, these correlations can be a problem. They might render irrelevant the explanatory power of statements about a certain feature to have caused a system decision, and wrongly inform data subjects about which elements of their financial behavior they should change to obtain a positive outcome for their loan application. This explains why some data protection laws that require indicating the factor or parameter that is mostly relevant [17] in the decision-making process might appear very difficult to implement in these complex AI application systems.

A visually striking example is provided in Figure 6 for the COVID-19 scenario, showing two images that have been obtained through a pixel alteration procedure in order to specifically alter the decision of the system to positive from negative, and vice versa. Each image displays the discriminant patterns that need to be added to the X-ray images to turn over the decision. These patterns follow specific correlations that have been learned by the system, which do not seem to have a clear causal relation to the actual lung disease, even if some elements of the anatomical shape of lungs are perceptible.

In the literature, the issue of correlation and causality in machine learning, and statistical prediction in general, is well known and widely debated. For instance, some have discussed whether this is a problem for the further development of AI [62], and others have tried to reintroduce full probabilistic modelling [63] and use methods from the statistical field of causal inference [64] in machine learning. It is undeniable that, for the time being, with the inability to provide causally grounded explanations, explanatory methods for AI will be fundamentally limited. This might be a problem for certain domains of applications, in particular in legal and medical fields [65]. This is surely a fundamental limitation if the aim is to rely solely on explanations to make automated decision-making systems compliant with legal frameworks. In other terms, if explanations are meant as the rational link between an effect (a decision) and its cause (the data subject's personal characteristics that are relevant for that decision and are automatically analyzed or predicted), such an explanation might appear more and more difficult to reach in models based on correlations as the ones described above.

IV. Towards Legal Explainability for a Trustworthy AI

The presented analysis has highlighted major challenges for the ability of AI systems to provide legible and legally applicable explanations. The legal and technical implications are in large parts the consequence of increasingly complex and opaque predictive machine learning models, trained with large collection of heterogeneous data, and applied in a growing range of application domains. The main reason of the success of these current machine learning systems is their undeniable outstanding performance. Such models have—often explicitly so—not been designed to be transparent or explainable in the first place, trading these elements for predictive power [61]. This situation is exacerbated if one additionally considers that current systems are mostly designed to rely on statistical analysis of correlations. These scientific realities should be taken into account at some point when assessing the risks of using a particular model, and not only their explanatory power.

In this section, an analysis of the challenges discussed in Section III is proposed, in light of the GDPR provisions. An outlook on how to reconcile these technical challenges with legal requirements follows, including two proposals:

- ❏ the use of Data Protection Impact Assessment (DPIA) to promote explainability;
- ❏ the justification of AI-based decision-making systems in the context of transparency requirements.

A. Beyond Explainable AI as a Singular Tool

The central aim at the outset of this study was to explore to what extent current and future AI systems can provide adequate and satisfactory explanations that would be admissible from a legal point of view in high-risk cases, exemplified by following the applicable EU data protection framework. Although it might appear that there is just one single form of explanation in the GDPR, there is no one-size-fit-all explanation in practice: each form of explanation (ex-ante, ex-post, expert-oriented or subject-oriented, more or less granular) strongly depends on the context of the problem and on the capacity of the data subject to interpret the results [15]. In general, achieving high performance in increasingly complex tasks is only achievable with opaque models, as explored in Section III. This can make the provision of context-adequate explanations a challenging task.

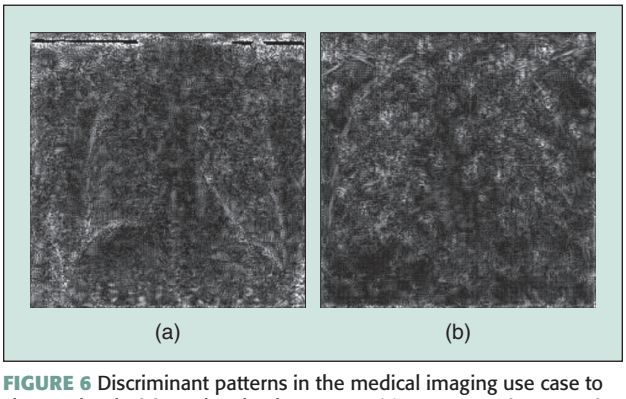

FIGURE 6 Discriminant patterns in the medical imaging use case to change the decision taken by the system. (a) From negative to positive; (b) From positive to negative.

Even the elementary machine learning systems implemented in the credit scoring use case may exhibit these contextual limitations, while in the context of a computer vision problem such as the medical imaging scenario developed in the COVID-19 use case, providing generally legible explanations may be simply impossible. In addition, the possibility to give a satisfactory, fair, and transparent explanation is often considered to depend on the possibility of showing a causal link between the input data and the final decision [66], which clearly has been illustrated to be fundamentally limited for current AI systems.

Explanations—in the traditional sense—are not only sometimes problematic, but also not sufficient to make AI socially and legally 'desirable'. In particular, several scholars reflected upon the "transparency fallacy" of algorithmic explanation [67], i.e., the risk that even a meaningful explanation could be not effectively received or understood by the data subjects (due to its technical nature or to the limited attention, interests or—even temporary—cognitive capabilities of the data subject) [68]. Thus, in line with our preliminary analysis [8], the quality of possible explanations in some cases might not be judged adequate under Article 22(3) of the GDPR. Ultimately, adopting a strict approach, this would mean that either more technologically advanced (and possibly obscure) forms of decision-making should be prohibited because their decisions are unexplainable, or, conversely, that we should tolerate AI-based decision-making systems that do not formally respect the transparency duties of the GDPR. These last considerations may lead to an impossible choice: technologically advanced decision-making systems are prohibited because they cannot comply with the GDPR explainability requirements; or AI-based decision-making systems that do not formally respect the transparency duties in the GDPR are tolerated.

Instead, a possible third way might be a technical and practical non-monolithic interpretation of a right to explanation in the context of automated decision-making systems. In order to do that, the right to explanation in the broader picture of the GDPR should be better contextualized. In particular, Article 22(3) and recital 71, when addressing the possible safeguards to reach accountable automated decisions, do not focus mostly on the right to explanation, but on a varied set of tools (including the right to contest, human in the loop, and algorithmic auditing). Further, in Articles 13–15 the "right to obtain meaningful information about the logic involved" in a system does not explicitly require this information to be a full explanation of the decision-making system as exemplified in our case studies. Beyond Articles 13–15 and 22 (and recital 71), the GDPR contains several notions that might influence the interpretation of accountability duties also in case of automated profiling and explanations. Data protection principles in Article 5 include the fairness principle, the lawfulness principle, the accuracy principle, the purpose-limitation principle, the data minimization and the storage limitation principle, the integrity and confidentiality principle, and the accountability principle, which seem all very relevant requirements towards more desirable AI.

In our view, in cases where the best possible technical explanation is not satisfactory, further tools should be implemented too. Firstly, they might rely on the disclosure of meaningful information about a Data Protection Impact Assessment (DPIA) on the algorithmic decision-making system (see Section IV-B). Secondly, apart from assessing the potential impact, a different set of tools, technical developments and approaches could be considered in order to mitigate the identified risks, and embed a potential lack of explanation in a framework of additional measures. These could be the result of a justification test on the algorithm, based on the mentioned data protection principles given in Article 5 (see Section IV-C).

B. Proposal: Explanations and the Data Protection Impact Assessment (DPIA)

As affirmed in the previous section, the impact assessment of AI applications might be not only a desirable solution, but also a feasible way to overcome many problems about the inscrutability of black-box systems, while reaching an adequate justification of algorithmic decision. The impact assessment model proposed by the GDPR is the Data Protection Impact Assessment (DPIA) that could also be used for AI applications. The DPIA, as regulated in Article 35 of the GDPR, is a tool to describe, assess, and mitigate the risks for data subjects resulting from the processing of personal data. As noted in [69], the DPIA comes in addition to the obligation for data controllers to manage risks, consisting in their identification, their analysis, their estimation, their evaluation, their treatment, and their regular reviews. In the case of possible infringement of the rights and freedoms of data subjects, a DPIA must be carried out. The EDPB indicated ten indices for high-risk data processing: evaluation or scoring, automated decision-making with legal or similar significant effect, systematic monitoring, sensitive data or data of highly personal nature, large-scale data processing, the combination of multiple data sets, the presence of vulnerable data subjects, the use of innovative technologies or organizational measures, extra-EU data transfers, and the limitation of data subjects' rights. In summary, applying the DPIA framework to AI systems would be essential to protect fundamental rights and human rights of individuals [70]. Some scholars have even tried to elaborate a DPIA model that could take into account all concerned human rights of individuals [10].

Both use cases presented in this paper clearly mark such high-risk situations for which the data controller would need to carry out a DPIA. For the credit scoring use case, the banking institution is processing financial records of customers alongside personal information (location, employment, possibly health, etc.) that might be considered either as sensitive data or as data of a highly personal nature, used for evaluation or scoring. Automated decision-making based on these data may have legal or similar significant effects, and this may prevent the data subject from exercising a right or using a service under a contract. Even more pronounced is the relevance of the medical image case study: the hospital is processing patient's health data, which are considered sensitive data. In addition, patients may be considered as vulnerable data subjects, and the data processing may be done on a large scale. The use of multiple data sets and of innovative technologies could also apply to many machine learning based systems.

Accordingly, in such situations, a DPIA should be necessary: the data controller should describe the data processing, assess the necessity and proportionality of it, analyze risks for data subjects (e.g., risks of discrimination, violation of the right to health, to freedom of thought, of movement, attack to their digital identity, etc.), and seek for the data subjects' opinion on the data processing. This procedure will also imply the description of the logic of the AI-based system, as well as the assessment of its impact on the fundamental rights and freedoms relevant for the application and the ways of mitigating it [70]. Making this report public (at least in some parts), would also significantly help to reach a better level of transparency and justification of AI applications [10].

The risks to rights and freedoms[1] might concern, in the credit scoring scenario, a) the discrimination of some social groups (based on ethnic origins, gender, sexual origins, social conditions) in the access to loans that might be embedded in AI biases; b) the automatic exploitation of individuals in situations of financial vulnerabilities or in situations of gambling addictions through systems of unsustainable interest rates; c) the economic, social and psychological disadvantages of families of individuals in accessing loans at unfavorable conditions (resulting in adverse effects to the fundamental right to health or even to conduct business). Mitigating these risks might be done also without requiring advanced explainability techniques: for example, the data controller could check the presence of eventual biases against social minorities or against gambling addicted people in the training data sets of the credit scoring systems. In addition, the data controller could check on a regular basis the effects of credit scoring outcomes to see whether it led to unfair economic, social or psychological damages to the consumers, seek for the opinion of the representatives of the data subjects (as Article 35(9) GDPR requires), or prepare periodical reports.

Regarding the medical imaging scenario, the risks to rights and freedoms might include, a) the discrimination of some social groups (based on ethnic origins, gender, sexual origins, social conditions) in their access to COVID-19 medical treatments[2]; b) false negatives that could lead to serious health damages to the patient, and to the group of people that might be infected by him/her; c) false positives that could lead to unnecessary medical treatments (that might both damage the health conditions of the patient and diminish the availability of drugs to people that might need them in a pandemic), and to psychological suffering of the patient and of his/her family; etc. Even in this case, addressing and mitigating these risks might be done without opening the 'black-box' of the AI application. For example, the data controller could check the presence of eventual biases against social minorities in the training data sets of the COVID-19 detection tool. The data controller can also check the effects of COVID-19 detection outcomes on a regular basis to see whether it led to false positives or false negatives and whether these false diagnoses refer to people belonging to some particular minorities (ethnicity, gender, etc.). In both cases, when these problematic aspects are detected, the data controller could provide some organizational mitigation measures, for example a second-step human-mediated decision in 'crucial' cases where the user belongs to a situation that has been observed (during the algorithmic auditing) to be at high risk of unfairness, inaccuracies and adverse effects in general[3]. In addition, the data controller could organize some periodical questionnaires on people who used this tool to check whether the system produced further damages, or perform some simulations to see whether particular people are discriminated or anyway adversely affected.

C. Proposal: Justification of AI-based Decision-making Systems in the Context of Transparency Requirements

AI systems can be judged from different perspectives, considering various aspects such as performance, robustness, transparency, or explainability. So far, the discussion regarding algorithmic explanations has focused on the latter, exploring the right to explanation. However, a broader consideration of those aspects becomes valuable in cases where no sufficient technical explanation can be provided according to Article 22(3) of the GDPR. Therefore, we argue that within the scope of a full AI risk assessment, as outlined in Section IV-B, one should in general not focus only on one particular risk. As an illustration, high performance in a critical application—such as in the medical imaging use case, where performance could save life—might warrant the risk of losing human legibility of results.

Beyond the trade-off between performance and explainability discussed in Section III-B, the robustness of AI systems is also highly intertwined with explainability. It is reasonable to assume that a robust model, systematically and properly checked, will yield more robust and trustworthy explanations, even if it might not be directly human-legible. This provides a crucial link between explainability and the broader topics of AI safety and cybersecurity. In the context of the latter, another trade-off might even be introduced between transparency and security, by increasing the attack surface against models [78] in providing full technical transparency, including ways to compromise the validity of explanatory methods [79]. It could also be argued that for an AI model lacking in explainability, conducting such proper technical system performance validations and increasing the overall robustness could be enough to provide information about the logics of the system, at least for a technical expert or an auditing process.

Likewise, addressing fairness issues in automated decision-making systems is obviously connected to addressing infringements on individual rights, as are the issues surrounding the legal requirements on explanations. Both fields could be jointly considered, and many explanatory methods are seen as instrumental to detect biases and unfair

[1]About the concept of "risk to a right" in the GDPR provisions about the DPIA see [71]–[75].

[2]Although this might appear as a mere hypothetical case, many studies in the US proved discrimination of health-related algorithms against black people. See for example [76].

[3]This approach might be criticized because it would imply to use a bigger amount of sensitive data in order to avoid discrimination. However, see the reply in [77].

decision-making [80], [81] (such as those providing group based explanations). There is no reason to prefer one of those provisions over the other, and seeing both fairness and explainability as common elements related to the individual rights of data subjects actually reinforces the notion that they should also be considered together when assessing risk.

In a justification test, the data controller could then explain why an antomated decision-making system (analyzed on the aggregated final effects on different data subjects, and analyzed in its purposes, intentions, etc.) is not unfair, unlawful, inaccurate, beyond the data protection principles listed in Article 5. Justifying a decision means not merely explaining the logic and the reasoning behind it, but also explaining why it is a legally acceptable (correct, lawful, fair, purpose-specific, accurate) decision, i.e., why the decision complies with the core of the GDPR principles in Article 5. Justification and explanation are not necessarily in conflict with each other: when explanations are not satisfactory or feasible, the data controller should in any case implement some additional accountability tools [67].

Approaching AI regulation from a broader angle is actually not only in line with the more general data protection principles, but also with recent European policy initiatives. In the proposal for a regulation of AI [9], a general risk-based approach is promoted, and the notion of trustworthy AI is put forward [20] to encompass explainability among other important aspects of increasing trust in AI such as robustness, safety, fairness, accountability, privacy, or data governance.

Promoting a full bag of measures to mitigate risks instead of focusing on a single element may also have beneficial effects on the way AI systems are designed. Developers and researchers of AI models could certainly be encouraged to overcome some explanatory problems with additional standardized requirements starting from a development principle of 'explainability-by-design'. Research should be promoted into new technical developments targeted to mitigate specific problems, for instance, into fields such as causal inference, automated provision of human legible explanations for complex systems, or methods for measuring uncertainty. At the same time, the known trade-offs between accuracy and performance on one side and explainability and transparency on the other, and the mutual importance of robustness, fairness, and explainability, may prompt an approach where in critical applications a range of risks could be weighed against each other, including explainability risks, ultimately approaching AI systems that are 'trustworthy-by-design'.

V. Conclusion

This article explores the challenges that current and future AI systems pose to the legal requirement of explainability in automated decision-making. The provisions of the GDPR that establish such a requirement were discussed, followed by a summary of the state of the art in the field of explainable AI. With this background, an interdisciplinary analysis was conducted, based on use cases from two high-risk scenarios—credit scoring and medical imaging.

Three major broad challenges to the ability of AI systems to provide sound explanations were identified: the growing complexity of data, models and algorithms, the trade-off between model performance and explainability, and the reliance of AI systems mostly on correlations in the data and not necessarily on real causal relationships. This analysis led to the conclusion that the quality of explanations in some cases might not be judged adequate under the GDPR.

To address this dilemma, a set of flexible proposals has been proposed, opting for a non-monolithic interpretation of a right to explanation in the larger context of the GDPR and ongoing policy developments on AI. Consequently, in cases where the best possible technical explanation is not satisfactory, further tools should be implemented. This includes a risk-based approach integrating explainability aspects, using tools such as the Data Protection Impact Assessment (DPIA) that could be put forward by the data controller. A second proposal discusses a justification test that would be based on the data protection principles stated in Article 5 of the GDPR, of which, for AI-based systems, the most relevant are performance, robustness and safety, and fairness. Following these lines of thought, this work advances the discussion from singular features of AI systems, such as their explainability or their robustness, to the general ideal of 'Trustworthy AI' as put forward by the European Union. Instead of enforcing single standards on one of these dimensions, the overall trustworthiness of an AI system might be the result of weighing the risks in terms of security, safety, fairness, transparency, and explainability.

References

[1] High Level Expert Group on Artificial Intelligence. "A definition of AI: Main capabilities and scientific disciplines," 2019.

[2] Article 29 Data Protection Working Party. "Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679," 2018.

[3] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)," 2016.

[4] K. M. E, "The right to explanation, explained," *Berkeley Technol. Law J.*, vol. 34, pp. 189, 2019, doi: 10.15779/Z38TD9N83H.

[5] "Atlas of automation. Automated decision making and participation in Germany," Algorithm Watch, Tech. Rep., 2019.

[6] F. K. Došilović *et al.*, "Explainable artificial intelligence: A survey," in *Proc. 41st Int. Convention Inf. Commun. Technol., Electron. Microelectron.*, 2018, pp. 210–215, doi: 10.23919/MIPRO.2018.8400040.

[7] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019, doi: 10.1016/j.artint.2018.07.007.

[8] R. Hamon *et al.*, "Impossible explanations? Beyond explainable AI in the GDPR from a COVID-19 use case scenario," in *Proc. ACM Conf. Fairness, Accountability Transparency*, 2021, pp. 549–559, doi: 10.1145/3442188.3445917.

[9] "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence," European Commission. 2021.

[10] A. Mantelero, "AI and Big Data: A blueprint for a human rights, social and ethical impact assessment," *Comput. Law Security Rev.*, vol. 34, no. 4, pp. 754–772, 2018, doi: 10.1016/j.clsr.2018.05.017.

[11] C. Kuner *et al.*, Eds., *The EU General Data Protection Regulation (GDPR): A Commentary*. London, U.K.: Oxford Univ. Press, 2020.

[12] E. Bayamlioglu, "Contesting automated decisions," *European Data Protection Law Rev. (EDPL)*, vol. 4, no. 4, p. 433, 2018, doi: 10.21552/edpl/2018/4/6.

[13] A. Roig, "Safeguards for the right not to be subject to a decision based solely on automated processing (Article 22 GDPR)," *European J. Law Technol.*, vol. 8, no. 3, 2017.

[14] S. Wachter *et al.* "Why a right to explanation of automated decision-making does not exist in the general data protection regulation," *Int. Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017, doi: 10.1093/idpl/ipx005.

[15] G. Malgieri and G. Comandé, "Why a right to legibility of automated decision-making exists in the general data protection regulation," *Int. Data Privacy Law*, vol. 7, no. 4, pp. 243–265, 2017, doi: 10.1093/idpl/ipx019.
[16] M. E. Kaminski and G. Malgieri, "Multi-layered explanations from algorithmic impact assessments in the GDPR," in *Proc. ACM Conf. Fairness, Accountability Transparency*, 2020, pp. 68–79, doi: 10.1145/3351095.3372875.
[17] G. Malgieri, "Automated decision-making in the EU Member States: The right to explanation and other 'suitable safeguards' in the national legislations," *Comput. Law Security Rev.*, vol. 35, no. 5, p. 105327, 2019, doi: 10.1016/j.clsr.2019.05.002.
[18] A. Selbst and J. Powles, "'Meaningful information' and the right to explanation," *Int. Data Privacy Law*, vol. 7, no. 4, pp. 233–242, 2017, doi: 10.1093/idpl/ipx022.
[19] N. Purtova, "The law of everything. Broad concept of personal data and future of EU data protection law," *Law, Innovation Technol.*, vol. 10, no. 1, pp. 40–81, 2018, doi: 10.1080/17579961.2018.1452176.
[20] High Level Expert Group on Artificial Intelligence. "Ethics guidelines for trustworthy AI," 2019.
[21] T. Gebru *et al.*, "Datasheets for datasets," 2020, arxiv:1803.09010.
[22] M. Mitchell *et al.*, "Model cards for model reporting," in *Proc. ACM Conf. Fairness, Accountability Transparency*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.
[23] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52,138–52,160, 2018, doi: 10.1109/ACCESS.2018.2870052.
[24] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, arXiv:1702.08608.
[25] W. J. Murdoch *et al.*, "Interpretable machine learning: Definitions, methods, and applications," 2019, arXiv:1901.04592.
[26] P. J. Phillips *et al.*, "Four principles of explainable artificial intelligence," Tech. Rep. 8312-draft, 2020, doi: 10.6028/NIST.IR.8312-draft.
[27] S. Jesus *et al.*, "How can I choose an explainer? An application-grounded evaluation of post-hoc explanations," in *Proc. 2021 ACM Conf. Fairness, Accountability Transparency (FAccT)*, 2021, pp. 805–815, doi: 10.1145/3442188.3445941.
[28] R. Guidotti *et al.*, "A survey of methods for explaining black box models," *Comput. Surveys*, vol. 51, no. 5, pp. 1–42, 2019, doi: 10.1145/3236009.
[29] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, p. 206, 2019, doi: 10.1038/s42256-019-0048-x.
[30] C. Molnar, "Interpretable machine learning," 2021. https://christophm.github.io/interpretable-ml-book/
[31] Z. C. Lipton, "The mythos of model interpretability," *Commun. ACM*, vol. 61, no. 10, pp. 36–43, 2018, doi: 10.1145/3233231.
[32] C. Chen *et al.*, "An interpretable model with globally consistent explanations for credit risk," 2018, arXiv: 1811.12615.
[33] M. T. Ribeiro *et al.*, "'Why should I trust you?': Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[34] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inform. Process. Syst.*, vol. 30, pp. 4765–4774, 2017.
[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. European Conf. Comput. Vision,* 2014, pp. 818–833, doi: 10.1007/978-3-319-10590-1_53.
[36] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *Proc. 35th Int. Conf. Machine Learning*, 2018, vol. 80, pp. 2668–2677.
[37] V. Arya *et al.*, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," 2019, arxiv:1909.03012.
[38] A. Bibal *et al.*, "Legal requirements on explainability in machine learning," *Artif. Intell. Law*, vol. 29, no. 2, pp. 149–169, 2020, doi: 10.1007/s10506-020-09270-4.
[39] M. Brkan and G. Bonnet, "Legal and technical feasibility of the GDPR's Quest for explanation of algorithmic decisions: Of black boxes, white boxes and Fata Morganas," *European J. Risk Regul.*, vol. 11, no. 1, pp. 18–50, 2020, doi: 10.1017/err.2020.10.
[40] P. Hacker *et al.*, "Explainable AI under contract and tort law: Legal incentives and technical challenges," *Artif. Intell. Law*, vol. 28, no. 4, 2020, doi: 10.1007/s10506-020-09260-6.
[41] C. Henin and D. L. Métayer, "A multi-layered approach for interactive black-box explanations," 2020, hal-02498418.
[42] S. Meacham *et al.*, "Towards explainable AI: Design and development for explanation of machine learning predictions for a patient readmittance medical application," in *Proc. Comput. Conf.*, 2019, pp. 939–955, doi: 10.1007/978-3-030-22871-2_67.
[43] S. Rathi, "Generating counterfactual and contrastive explanations using SHAP," 2019, arxiv:1906.09293.
[44] M. T. Keane and B. Smyth, "Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI)," in *Proc. Int. Conf. Case-Based Reasoning Res. Development*, 2020, pp. 163–178, doi: 10.1007/978-3-030-58342-2_11.
[45] M. Leo *et al.*, "Machine learning in banking risk management: A literature review," *Risks*, vol. 7, no. 1, p. 29, 2019, doi: 10.3390/risks7010029.
[46] X. Dastile *et al.*, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Appl. Soft Comput.*, vol. 91, p. 106263, 2020, doi: 10.1016/j.asoc.2020.106263.
[47] FICO – Explainable Machine Learning Challenge. [Online]. Available: https://community.fico.com/s/explainable-machine-learning-challenge
[48] M. Kłosok and M. Chlebus, "Towards better understanding of complex machine learning models using Explainable Artificial Intelligence (XAI) – Case of Credit Scoring modelling," Faculty of Economic Sciences, Univ. of Warsaw, Working Papers 2020-18, 2020.
[49] L. M. Demajo *et al.*, "Explainable AI for interpretable credit scoring," 2020, arXiv: 2012.03749.
[50] M. Bücker *et al.*, "Transparency, auditability, and explainability of machine learning models in credit scoring," *J. Oper. Res. Soc.*, 2021, doi: 10.1080/01605682.2021.1922098.
[51] P. Bachtiger *et al.*, "Machine learning for COVID-19—Asking the right questions," *Lancet Digital Health*, vol. 2, no. 8, pp. e391–e392, 2020, doi: 10.1016/S2589-7500(20)30162-X.
[52] K. Hao, "AI is helping triage coronavirus patients. The tools may be here to stay." MIT Technology Review, 2020.
[53] B. McCall, "COVID-19 and artificial intelligence: Protecting health-care workers and curbing the spread," *Lancet Digital Health*, vol. 2, no. 4, pp. e166–e167, 2020, doi: 10.1016/S2589-7500(20)30054-6.
[54] H. B. Syeda *et al.*, "Role of machine learning techniques to tackle the COVID-19 crisis: Systematic review," *JMIR Med. Informat.*, vol. 9, no. 1, p. e23811, 2021, doi: 10.2196/23811.
[55] M. Wang *et al.*, "Deep learning-based triage and analysis of lesion burden for COVID-19: A retrospective study with external validation," *Lancet Digital Health*, vol. 2, no. 10, pp. e506–e515, 2020, doi: 10.1016/S2589-7500(20)30199-0.
[56] L. Wang *et al.*, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Sci. Rep.*, vol. 10, no. 1, 2020, doi: 10.1038/s41598-020-76550-z.
[57] K. He *et al.*, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recogn.*, 2016, pp. 770–778.
[58] J. Franklin, "The elements of statistical learning: Data mining, inference and prediction," *Math. Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005, doi: 10.1007/BF02985802.
[59] R. Padilla et al., "Survey on performance metrics for object-detection algorithms," in *Proc. Int. Conf. Syst., Signals Image Process.*, 2020, pp. 237–242, doi: 10.1109/IWSSIP48289.2020.9145130.
[60] G. Shmueli, "To explain or to predict?" *Statistical Sci.*, vol. 25, no. 3, pp. 289–310, 2010, doi: 10.1214/10-STS330.
[61] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, 2001, doi: 10.1214/ss/1009213726.
[62] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[63] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452–459, 2015, doi: 10.1038/nature14541.
[64] B. Schölkopf, "Causality for machine learning," 2019, arXiv: 1911.10500.
[65] G. Comandè *et al.*, "Causality and explanation in ML: A lead from the GDPR and an application to personal injury damages," preprint, 2020.
[66] A. Rouvroy, "The end(s) of critique: Data behaviourism versus due process," in *Privacy, Due Process and the Computational Turn*. Routledge, 2013, pp. 157–182.
[67] L. Edwards and M. Veale, "Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for," *Duke Law Technol. Rev.*, vol. 16, p. 18, 2017.
[68] G. Malgieri and J. Niklas, "Vulnerable data subjects," *Comput. Law Security Rev.*, vol. 37, p. 105415, 2020, doi: 10.1016/j.clsr.2020.105415.
[69] Article 29 Data Protection Working Party. "Guidelines on Data Protection Impact Assessment (DPIA)," 2017.
[70] H. Janssen, "An approach for a fundamental rights impact assessment to automated decision-making," *Int. Data Privacy Law*, vol. 10, no. 1, pp. 76–106, 2020, doi: 10.17863/CAM.54393.
[71] I. Borocz, "Risk to the right to the protection of personal data: An analysis through the lenses of Hermagoras," *European Data Protection Law Rev. (EDPL)*, vol. 2, no. 4, pp. 467–480, 2016, doi: 10.21552/EDPL/2016/4/6.
[72] K. Demetzou, "Data Protection Impact Assessment: A tool for accountability and the unclarified concept of 'high risk' in the General Data Protection Regulation," *Comput. Law Security Rev.*, vol. 35, no. 6, p. 105342, 2019, doi: 10.1016/j.clsr.2019.105342.
[73] R. Gellert, "Understanding the notion of risk in the General Data Protection Regulation," *Comput. Law Security Rev.*, vol. 34, no. 2, pp. 279–288, 2018, doi: 10.1016/j.clsr.2017.12.003.
[74] M. Macenaite, "The riskification of European data protection law through a two-fold shift," *European J. Risk Regulation (EJRR)*, vol. 8, no. 3, pp. 506–540, 2017, doi: 10.1017/err.2017.40.
[75] D. Kloza *et al.*, "Towards a method for data protection impact assessment: Making sense of GDPR requirements," Policy Brief, 2020, doi: 10.31228/osf.io/es8bm.
[76] Z. Obermeyer *et al.*, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: 10.1126/science.aax2342.
[77] I. Žliobaitė and B. Custers, "Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models," *Artif. Intell. Law*, vol. 24, no. 2, pp. 183–201, 2016, doi: 10.1007/s10506-016-9182-5.
[78] "Artificial intelligence cybersecurity challenges – Threat landscape for artificial intelligence," ENISA, Tech. Rep., 2020.
[79] B. Dimanov *et al.*, "You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods," in *Proc. European Conf. Artif. Intell.*, 2020, vol. 325, pp. 2473–2480, doi: 10.3233/FAIA200380.
[80] M. Chorás *et al.*, "Machine learning – The results are not the only thing that matters! What about security, explainability and fairness?" in *Proc. Int. Conf. Comput. Sci.,* 2020, pp. 615–628, doi: 10.1007/978-3-030-50423-6_46.
[81] T. Begley *et al.*, "Explainability for fair machine learning," 2020, arXiv: 2010.07389.