The Coming of Age of Interpretable and Explainable Machine Learning Models

P.J.G. Lisboa[a], S. Saralajew[b], A. Vellido[c,d], R. Fern´andez-Domenech[c,d], T. Villmann[e,1]

[a]Liverpool John Moores University, Liverpool – United Kingdom [b]NEC Laboratories Europe, Heidelberg – Germany
[c]Dept. of Computer Science, UPC BarcelonaTech, Barcelona – Spain [d]IDEAI-UPC Research Center, Barcelona – Spain
[e]University of Applied Sciences Mittweida, Saxon Institute for Comp. Intelligence and Machine Learning, Mittweida – Germany

Abstract

Machine-learning-based systems are now part of a wide array of real-world applications seamlessly embedded in the social realm. In the wake of this realization, strict legal regulations for these systems are currently being developed, addressing some of the risks they may pose. This is the coming of age of the concepts of interpretability and explainability in machine-learning-based data analysis, which can no longer be seen just as an academic research problem. In this paper, we discuss explainable and interpretable machine learning as *post-hoc* and *ante-hoc* strategies to address regulatory restrictions and highlight several aspects related to them, including their evaluation and assessment and the legal boundaries of application.

Keywords: XAI, Interpretable ML, Explainable ML, Transparent AI

[1] Corresponding author
Email address: thomas.villmann@hs-mittweida.de (T. Villmann)

1. Introduction

The design of *Machine Learning* (ML) models is currently dominated by the development of deep *Multi-Layer Perceptrons* (MLPs) and variants thereof, which consist of increasingly complex structures and modules under the umbrella term of Deep Learning (DL) [1, 2]. These approaches may include specific components like convolutional layers for adaptation to specific tasks like image processing and classification [3, 4]. The numerical validation of deep networks justifies the theoretical correctness [1, 2, 5, 6].

The training of such complex models requires careful adaptation of the internal model parameters (frequently accompanied by strategies to reduce numerical instabilities), to ensure robustness and to avoid overfitting [7, 8, 9]—including autoencoder learning for the pre-training of layers, dropout learning approaches, regularization techniques, and resilient network architectures, to name a few [3, 10, 11, 12].

Nevertheless, the more complex the architectures, the more difficult the interpretation or explanation of how and why a particular network prediction is obtained, or the elucidation of which components of the complex system contributed essentially to the obtained decision. The need for transparency varies with the application area. This involves proportionality between the burden imposed by explainability and interpretability, against the value this offers to the end-user. The approach we take is that the arbiter for proportionality is the applicable set of legal frameworks, hence they are the pivotal content of the paper. A unified approach for the assessment of interpretability that combines several existing axiomatic methodological frameworks is subsequently presented and its relations with the legal frameworks are discussed. An overall limitation of the accompanying review of technical methods is that it has a particular focus on tabular data and imaging data. This is because these types of data are frequently used in safety-critical contexts such as medicine and engineering, for instance, in diagnostic decision support and autonomous driving, where it becomes crucial to understand how the model generated the prediction. In other words, the model acting as a black-box system is not sufficient any longer [13].

In the European context, legal requirements have been enforced by the *General Data Protection Regulation* (GDPR) since 2018, which, as explained by Bacciu et al. [14], mandates a "right to explanation" of decisions made on citizens by "automated or artificially intelligent algorithmic systems." This is compounded by the current development of a legal framework on *Artificial Intelligence* (AI) by the European Commission of obvious impact on ML [15].

While mirroring many of the GDPR elements, it also discriminates between AI applications according to a four-level risk assessment, from "minimal-risk" to "unacceptable-risk," with "high-risk" AI applications being subject to strict obligations before being marketed, and even "limited-risk" ones being tied to specific transparency obligations. Note, though, that this legal proposal often refers to *transparency* and *trustworthiness* of AI systems, instead of explainability and interpretability, which have a role to play in delivering the higher level goals.

Regulatory restriction has made the development of tools and strategies to explain those complex models an urgent necessity [16]. As pointed out by Rudin [17], these *post-hoc* strategies might be problematic because explanations frequently are unreliable and can even be misleading. An alternative to that are interpretable models, which provide *ante-hoc* inherently the possibility for model explanations. It was recently pointed out that interpretable models should be favored for high-stakes decisions if possible, rather than "explained" black-box models [18]. For these reasons, in the review of methods of Section 3, we take as the starting point the same high-level distinction between *post-hoc* explanations and *ante-hoc* models that are explainable or interpretable by design.

This has led to a number of guidelines for algorithmic transparency and accountability, but also to a statement from the Association of Computing Machinery, made in 2017, that is aptly reviewed by Burkart and Hubert [19], that formalises five ways to gain interpretability based on three concepts: explanation generation, learning interpretable models, and surrogate model learning. In our paper, we group methods starting from the distinction between *post-hoc* and *ante-*

hoc approaches, leading to a focus on metrics to quantify transparency in a unified framework that includes several concepts that recur in current reviews.

While our proposed framework is consistent with the taxonomy of methods presented elsewhere, our review of methods can be complemented in some aspects where existing reviews are especially detailed. The survey by Guidotti et al. [20] is particularly informative about decision tress and rule-based models, which, alongside linear models, are noted as being considered "easily understandable and interpretable for humans." This paper underlines that these methods for providing explanations are effective only when they have "humanreasonable sizes." In the context of assessment of explainability, three desiderata known as the PDR framework are introduced by Burkart and Hubert [19]: predictive accuracy, descriptive accuracy, and relevancy [21]. These three aspects are contained in the framework proposed later in Section 3, and predictive accuracy, in particular, motivates the inclusion of the performance criterion in Table 1. Both Guidotti et al. [20] and Linardatos at al. [22] provide detailed and well-illustrated reviews of methods to explain DL models, including a comparison of saliency maps for samples from ImageNet.

In this position paper, we aim to make the following contributions:

- We consider in detail the state-of-the-art in regulatory and legal frameworks for the application of ML in terms of transparency, interpretability, and explainability. This is still fluid as legal statutes have, to our knowledge, been barely tested in court. Nevertheless, core principles are emerging which are discussed in the paper, along with our position regarding the direction of travel of these key developments for the future of our field.

- We develop a unified approach for the assessment of interpretability that combines several existing axiomatic frameworks. The proposed approach comprises a super-set of measurable criteria for the evaluation of standalone models, that can be readily complemented with qualitative indica-

tors from end-user experiments in real-world applications. In order to support this unified approach, we include a non-exhaustive review of the current state of eXplainable AI (XAI) and discuss it according to the regulatory and legal frameworks previously described. In order to highlight the domain-dependency of this problem, two case areas are discussed in detail: healthcare and banking.

2. The Legal Boundaries of Interpretability and Explainability for Machine Learning-based Models

At a time when ML is no longer just an academic pursuit with minor inroads on real-world applications but a commercial commodity that is proactively being sold to citizens in many guises and consequential contexts, the societal impact of this technology makes it enter a completely different realm. A realm in which ML is bound to normative regulation and law and in which model interpretability or explainability, or both, come to the fore to play a central role as tools to guarantee model accountability and acceptance, as well as trustworthiness. In Europe, this must be considered in the context of the GDPR since 2018, which, as first pointed out by Goodman and Flaxman [23] and explained by Bacciu et al. [14], mandates a "right to explanation" of decisions made on citizens on the basis of "automated or artificially intelligent algorithmic systems." Note, though, that the own interpretation of this mandate is nothing but controversial (as one might expect of the complex matching between technical systems and law) and that even its mere existence was quickly put into question by Wachter et al. [24]. A common theme in this controversy has to do with how much a requirement of explainability might harm ML effectiveness (think of DL as an extreme example) and how much it could contribute to stifling innovation. At this point, the priority is understanding how a legal text such as GDPR can ultimately be interpreted in courts that are not necessarily aware of the nuances of (semi-)automated decision-making based on AI and ML in particular. Such difficulty in bridging AI technicalities and law has been thoroughly discussed, for instance, in the socially-sensitive context of UK law and pub-

lic administration [25]. This work identifies lack of explanation in automated decision-making as the greatest legal challenge encountered, as it is "key to judicial review" and defines it in three flavors, following Burrell [26]: intentional opacity (for intellectual property protection), illiterate opacity (ML systems only interpretable by data scientists), and intrinsic opacity (systems that are *per se* non-interpretable, according to the definition defended in the current paper).

> The qualification of what the idea of explanation means specifically in legal terms (i.e., what sort of explanation would be appropriate and accepted in court) is a task that may require discussion across domains between legal scholars and data and cognitive scientists [27]. This is not just an academic flight of fancy but, in fact, the core requirement of legal interpretation and, therefore, of paramount importance. Unsurprisingly, this debate has flourished far more in legal texts than in technical publications [25, 28]. Doshi-Velez et al. [27] agreed on a definition of explanation as "human-interpretable information about the logic by which a decision-maker took a particular set of inputs and reached a particular conclusion," with obvious translation to ML-based decision-making. They also discuss the situations in which the benefits of providing an explanation outweigh the costs (which implies the notion that this is not the case in all situations) and focus on the interesting distinction between ML explanation at large and legally-operative explanation. The latter is described as best focused on elucidating the relative impact of individual data attributes on model-based decisions. Therefore, and following the distinction between interpretable and explainable ML we are posing in this paper, such idea means that, *from a legal standpoint,* it would be preferable to focus on explainable systems, whose *posthoc* nature fits the idea of the "explanation system" being independent of the ML system itself and requiring a "concept-mapping" that makes this information regarding relative input relevance amenable to human understanding. This conceptual mapping can only be made operational through the collaboration between domain experts and data scientists in the design of ML-based pipelines for decision support, as in the framework proposed by Vellido [13]. Such collab-

oration is also necessary, as the ML pipelines will not be able to provide proper explanation *ex post* unless the domain-specific explanation requisites have already become part of the ML-based pipeline design. Furthermore, we argue that if the decision-making process affects citizens and is legally liable, legal experts should also be part of the design of the explanation system tied to the ML-based pipeline, together with data scientists and domain experts. This is further sustained by the exploration of the significance of contextual and performative factors in the implementation of *retrospective* transparency in the form of explanation [29]. In this study, Felzmann et al. warn of the potential limitations of GDPR in this aspect unless "assessments of trustworthiness based on contextual factors mediate the value of transparency communications." Different stakeholders might require different types and levels of explanation in different application contexts, a detail which is not covered by the text of the law itself. The authors support "a tailored and multi-stakeholder approach to transparency for AI" with *performativity* [30] as a way to conceptualize the "link between transparency effects and contextual factors."

As already stated in Section 1, the legal frameworks and proposals often refer to transparency and trustworthiness of AI systems, instead of explainability and interpretability, but with different connotations; note also that language here can be tricky, as notions of transparency and interpretability or explainability do not necessarily equate; something that, as reported by Doshi-Velez and Kim [27], was neatly stated in a report of UK's House of Lords, back in 2017 [31].

The main risk faced by the implementation of transparency, explainability, and interpretability of ML systems in practice is the existing gap between law, regulation, and the implementation in such context-specific practice. The ability to bridge such a gap will be the true hallmark of the coming of age of these concepts in ML.

The need for interpretability and explainability increases with the liability and responsibility for decisions made by an ML model. For example, if a recommender system presents products to consumers while surfing through the

internet, there is an expectation that sales will increase by this type of advertisement. If this is not the case, the project lead might have to explain why the advertisement campaign was unsuccessful. Consequently, there might be a need to interpret or explain the recommender system to the customer because the project lead is responsible for the product. Another example where the sole responsibility for a functioning ML model could cause a need for interpretability and explainability is stock market forecasting. If the forecast is too often false, customers will lose trust in the product. However, if the company equipped the forecasting model with explaining factors so that customers can make their own decisions on whether to trust the forecast or not as a basis for an investment, this would mitigate disappointment in case of an incorrect forecast.

Doshi-Velez and Kim [32] summarized several ML-related desiderata where interpretability and explainability are used for confirmation:

- *Fairness and unbiasedness*: Avoid the discrimination of protected groups;
- *Privacy*: Protect sensitive information in the data;
- *Robustness*: Ensure the prediction is stable in terms of parameter or input variations;
- *Causality*: Guarantee that observable input-output relations will also occur in the real system;
- *Usability and trust*: Provide information on why the model made a specific prediction to improve the confidence of humans and to assist users.

Hence, beyond the required interpretability and explainability by law, there is often an intrinsic motivation behind the quest for model behavior comprehensibility. The forthcoming *Artificial Intelligence Act* (AIA) [15], though, makes us veer from desideratum to obligation in certain application areas. Its Recital 47 (linked to Article 13) states that "to address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons, a certain degree

of transparency should be required for high-risk AI systems. Users should be able to interpret the system output and use it appropriately." Annex III of the AIA lists these high-risk AI systems as biometric identification and categorization of natural persons; management and operation of critical infrastructure; education and vocational training; employment, workers management, and access to self-employment; access to and enjoyment of essential private services and public services and benefits; and law enforcement. Any ML application pertaining to any of the shortlisted areas will therefore have to abide to transparency and interpretability requirements. A tool that provides a procedure for AIA conformity assessment (called capAI) and described as "a governance tool that ensures and demonstrates that the development and operation of an AI system are trustworthy—i.e., legally compliant, ethically sound, and technically robust – and thus conform to the AIA" has recently been proposed [33]. In capAI, XAI is addressed as an element of Ethics-Based Auditing (EBA), where EBA is understood as "a governance mechanism that allows organisations to operationalize their ethical commitments and validate claims made about their AI systems."

A summary comment on different elements of current regulation that are relevant to the issues of interpretability and explainability is provided in Table A.2 in the appendix. This does not attempt to be comprehensive, but it aims to reflect the highest-level attempts by different regulatory sources.

3. Explainable and Interpretable Models in Machine Learning

As already mentioned in the introduction, the majority of currently applied ML models are based on deep MLPs, which often achieve impressive results in regression and classification problems in very different application areas. Unfortunately, most of these complex networks work as black-box algorithms such that the user is only provided with the prediction or decision of the model, but with none or very limited information on how these results were obtained. However, the benefit of ML models will be much higher for the data analysts and the experts in the application domain if they are provided with additional information

about the prediction process—even in an interactive manner as part of the learning process, such as in the *Explanatory Interactive Machine Learning* framework proposed by Teso and Kersting [34]. Such information can potentially increase the trustworthiness of the model, allowing the user to draw further conclusions and extend, in this way, the knowledge base for the problem (as clearly illustrated for specific domain scenarios in Section 4.2). In particular, several desiderata for interpretability and robust explainability of ML models can be identified as minimum requirements [18, 35, 36]:

- *Explicitness and Comprehensibility*: Is the learning approach able to represent its learned knowledge in a human-understandable fashion, and are explanations immediately understandable?
- *Faithfulness*: Does the interpretation and/or explanation truly reflect the learned model?
- *Stability*: How consistent are the explanations for similar or neighboring examples?
- *Sparsity*: Is the explanation compact, minimizing the information provided to the user?
- *Modularity*: Is the model not too complex and can be decomposed into simple sub-modules which are interpretable and can be easily explained?
- *Model inspection*: Is it possible to obtain model representations and descriptions of specific model properties?

Two main strategies in this context can be observed: XAI and *interpretable models*, which we characterize by the following definitions:

- **Explainable models**: The decision or prediction process of the model can be comprehended *post-hoc* by experts in the field using additional tools and elaborate considerations.
- **Interpretable models**: The decision or prediction process of the model can be easily comprehended (in a reasonable amount of time) by experts in

the field according to the *ante-hoc* model design and their domain knowledge.

Both strategies have to provide a qualitative understanding of the process that links the input variables (features) with the outcome or response to make the model plausible and the prediction trustable [37]. The following subsections list, non-exhaustively, some *post-hoc* and *ante-hoc* approaches, as well as an inbetween approach: Self-Explaining Neural Networks (SENN). This is followed by a description of measures to quantify interpretability and explainability.

3.1. Post-hoc Approaches: Explaining Machine Learning Models

Post-hoc approaches comprise those for black-box models for which explanations are sought to describe particular aspects of the considered model [36]. The corresponding tools generally fall into the following categories, starting with variants of sensitivity analysis, but extending to more complex methods:

- *Feature attribution with SHAP.* Feature attribution methods relate the model output to a small number of numeric or semantic input features. A landmark paper provides a unified framework to explain black-box models using *SHapley Additive exPlanations* (SHAP, [38]). Shapley values are becoming a standard method for local explanations, for instance, in medicine [39]. Originally founded on game theory as a way to determine the added value of an individual player in a coalition, Shapley values have attractive theoretical properties that translate to their application in ML. In particular, Lundberg and Lee [38] prove that the additive contributions defined by Shapley values are unique in meeting the properties of local accuracy, the requirement that inputs that are switched off must have no attributed impact, and consistency in the sense that if a model output experiences a more significant difference than another when a particular feature is omitted, then the attribution to that feature will also be more significant. The application to molecular diagnostic tests gives a critical appraisal of how the exact method and practical sub-sampling strategies compare to explain the relative importance of features for a specific patient

in this challenging setting [39]. The setting can be described as including a large number of covariates, a restricted number of observations, and a training set that is not representative of the population on which the model will be tested. Even in such setting, it was found that exact Shapley values can be used to determine the relative importance of features to the classification for a specific patient. Moreover, the feature attribution maps can be interpreted in the same way that physicians use different weightings of clinical factors to diagnose diseases for individual patients.

In contrast, Izzo et al. [40] showed that an inappropriate choice of baseline could negatively impact the explanatory power of Shapley values and possibly lead to incorrect interpretations. To avoid such defects, they present a method for choosing a baseline according to a neutrality value that is in accordance with how the model is used while decision-making. An alternative approach is to represent pre-trained black-box models with mimic models that are interpretable by design. However, it is difficult to derive them from data in a computationally efficient manner. Some advances have been made in generating nomograms for flexible models applied to tabular data [41]. This approach has also been pursued with a constructive approach to infer from a trained MLP a model with univariate and bivariate effects, in the form of partial response networks [42, 43].

- *Feature attribution with Saliency Maps* identify sparse components of the original signal that have the most influence on the model predictions, for example, *Class Activation Mappings* (CAM, [44]), occlusion maps [45], or gradient-weighted CAM [46]. It is important to note that saliency maps need to be correctly configured, or they can be misleading. Simple shifts in the inputs were used to test the so-called input invariance of several saliency maps and found that this property was not always observed and even that saliency maps could differ in models with different architectures but identical predictions for every input [47]. A popular feature attribution method that also fits this type is Local Interpretable Model-agnostic Ex-

planations (LIME), which aims to identify an interpretable model that is locally faithful to the classifier. It optimizes the trade-off between fidelity and interpretability (explainability) based on sampled instances (local) explained by simple interpretable models like linear models, which serve as local surrogates [37].

- *Adversarial and counterfactual explanations* comprise methods and approaches to reduce the opaqueness of black box models [48]. Adversarial samples demonstrate where ML models fail in a way that the investigation of those failures helps to understand model behavior and decisions [49]. Frequently, there is no specific desired outcome for an adversarial sample, but, more in general, the adversarial sample is just designed to fool the model predictions. In this sense, adversarial attacks (samples) detect model vulnerabilities, which could be exploited for malicious intent [50]. In contrast, counterfactual samples provide information about the model decision process by contrastive explanations for model decisions. More precisely, counterfactuals describe limitations of the model, such that this information can be used to influence the outcome in a directed manner; frequently a desired outcome is demanded [51]. Thus, adversarial samples should be semantically indistinguishable from original data whereas counterfactuals are designed to highlight limits of the model. However, both concepts are closely related and the generation of these samples frequently relies on similar mathematical concepts. Both concepts implicitly shed light on internal model decision processes, which can be used for the regularization and adaptation of the model as well as for explanation of the decision/prediction process [52]. Further, one can distinguish two different aims for adversarial or counterfactual investigations and explanations, particularly in the context of one-class-classifiers [53, 54, 55]: First, the perturbation of target class samples is considered to yield a prediction as non-target. We denote this as a false-negative adversarial/counterfactual approach. Second, non-target samples are modified to be detected as target by the model, which is denoted as false-positive attempt.

- *Activation maximization*, based, for example, on Generative Adversarial Networks (GAN) [56], use deep generative networks and tailored optimization methods to generate class-relevant inputs for convolutional neural networks [57]. A human user can then understand the internal represen-
360 tations assimilated by the network and the typical representations of the classes.

- *Rule extraction*, which, in this context, can be split into two categories: the first is decompositional rule extraction algorithms, which rely on the internal structure of a neural network, for instance, by interpreting the
365 activity of individual hidden nodes; the second is pedagogical rule extraction algorithms, which are model-agnostic and so apply to any black-box algorithm [58]. These models can be surprisingly effective in distilling loworder, very interpretable rules for complex models. Explanation through rule extraction is further discussed by Guidotti et al. [59]. Decision trees
370 are often used as a tool for rule extraction. An example was presented by R¨ognvaldsson et al. [60], which allows for the extraction of decision rules from deep neural networks to transfer knowledge from a reference model into an explainable equivalent [61]. This is an example of a mimic model, which seeks to reproduce the predictions of a black-box using, in this case,
375 a rule set.

- *Post-hoc metric learning* involves deriving a metric from a classifier and using it to map out the data structure [62]. Then, similarity networks are generated from which a classification of an input can be obtained by consulting its neighbors, a form of case-based reasoning (CBR). Typically,
380 the Fisher Information (FI) metric is calculated from a feedforward neural network, which induces a Riemannian metric on the space of input data. An example of the use of the FI metric in a medical CBR problem was presented by

Ortega-Martorell et al. [63]. This makes explicit the (dis)similarity metric that is implicit in all probabilistic models. The metric can be used to calculate pairwise distances along geodesics, which serve to map out the data structure in the form of a similarity network. The neighbours of a test point are the reference cases for k-NN classification with the Riemannian metric. This approach can also be implemented with Siamese networks, which have become very popular of late in the context of self-supervised learning [64]. Alternatively, the similarity maps can also be used for case-based reasoning, which is particularly relevant to medical applications [65].

3.2. Ante-hoc Approaches: Interpretable Models

Arguably, "the best explanation of a simple model is the model itself" [38] (i.e., it perfectly represents itself and is easy to understand). More formally, the propagation of information in a form that can be interpretable by the end-user with reasonable domain knowledge is clear from input through to prediction. Thus, interpretable models have to be transparent on all levels. The models surveyed in the following list belong to this category of transparent models:

- *Linear models*, such as linear regression or Linear Discriminant Analysis (LDA), in which the linear dependencies between data and prediction makes them inherently transparent (even if, as discussed by Molnar [66], they do not always create the best explanations). This concept can be adapted for classification beyond LDA through the use of appropriate link functions, as in the case of logistic regression. It should be noted that the success of logistic regression over many years is due in no small part to clever representations, for example, by careful discretization of continuous variables. This results in a linear-in-the-parameters model that is, in fact, very much non-linear. The downside of this approach is that it is at least partly subjective, as well as introducing discretization boundaries that can result in substantial variations in output for relatively small amounts of noise in the inputs, compounding the inter-observer variation that is known

to be an issue, for instance, in cytology [67]. This is a potentially strong argument for the use of ML methods that do not require discretization, provided they have appropriate levels of transparency. In this respect, this also points to the approaches discussed later involving generalized additive models.

- *Decision trees*: This rule-based system generates logical implications for model prediction (i.e., it can be taken as a rule-based model). However, interpretability becomes difficult as the decision tree grows. Geometrically, if-then rules on the original input variables generate axis-orthogonal boundaries, which will require many rules when fitting complex decision surfaces. This may be alleviated although not necessarily avoided by using other types of rule sets, such as oblique decision trees. These models have a long history including ID3, C4.5, Chi-square Automatic Interaction Detection (CHAID) [68] and Classification and Regression Trees (CART) [69], to name a few. All methods rely on a principled approach to measuring class separation either using information theory or statistical tests.

  The CART method was combined with bagging, a form of bootstrap re-sampling with replacement, to create the Random Forest algorithm [70] which is known to be very accurate but is no longer interpretable because of the large number of rules in the ensemble of decision trees, although it does provide a measure of variable importance.

- *Bayesian models*: These models show a factorization of the joint probability distribution of the data, represented in the form of a graph with variables as nodes and parent-child relationships of conditionality shown by edges. The graphs are initially derived from conditional independence maps followed by edge orientation. The resulting Directed Acyclic Graph (DAG) explicitly shows the dependencies between the input and the outcome to be predicted, which is what makes this approach interpretable. However, finding the initial skeleton of edges with the requirement of mutual independence

conditional on the rest of the graph is NP-complete. Fortunately, the PC algorithm, named after the initials of Spirtes and Gilmour [71], provides assurances of convergence to the generating graph of the data [72]. However, this method is sensitive to the order in which the nodes are tested for recursive elimination, so that care must be taken to stabilise the inferred map [73]. Recent developments include cognitive aspects of learning and knowledge representation separating detectable features and the respective reasoning for inference. This is represented by a novel network architecture denoted as the Classification-By-Components network (CBC) [74], which follows an intuitive reasoning-based decision process inspired by recognition-by-components theory from cognitive psychology.

- *Prototype methods*: These methods are based on learning of or the extraction of prototypical representations of the data set based on a dissimilarity measure [75] and a prototype assignment rule (e.g., the nearest prototype principle, k-nearest neighbors rule). By the prototypical representations and the dissimilarity measure, this paradigm naturally ensures interpretability. For classification learning, the family of *Learning Vector Quantizers* (LVQ, [76, 77]) is well-known to provide possibilities for non-standard metric usage and metric adaptation [78, 79]. The latter allows a direct evaluation of feature dependencies according to the model-inherent classification correlation analysis [80]. Unsupervised models for representation learning are the neural gas, fuzzy c-means, and self-organizing maps [81, 82, 83] or related one-class-classifier models [55]. In fact, prototype-based methods can be seen as a realization of case-based reasoning, which is a paradigm that involves solving a new problem using known solutions to similar past problems [18, 84]. Recent examples of this include the ProtoPNet method proposed by Chen et al. [85] for interpretable image classification and the method based on autoencoder architectures proposed by Li et al. [86].

- *Generalized additive models*: Recently, there has been interest in representing neural networks in the form of Generalized Additive Models (GAMs), that is to say, as a linear combination of interpretable non-linear functions involving only one or two input variables at a time [87]. GAMs are more than just explainable, as they are interpretable globally over the full range of input data and are considered a potential gold standard for interpretability [20]. Examples include Neural Additive Models (NAM) [88] and Explainable Boosting Machines[2] (EBM) [89]. They belong to a class of models that already included spline-based versions from the traditional statistical literature [90]. These models are positioned at the intersection between ML and statistics [91]. However, they have been hampered by a lack of effective ways to select the most informative components, particularly when bivariate terms are included. A proposal originally made by Friedmann [90] to interpret black box models was further developed in EBM and NAM. An alternative approach using ANOVA decompositions applied to pre-trained black box models, followed by model selection with Lasso, is used in the Partial Response Network (PRN) [42]. Results published with the NAM, EBM, SAM, and PRN show that, in performance for binary classification applied to tabular data, these methods are comparable with state-of-the-are ML including deep neural networks. Some of these methods are available for download from public domain websites.[3]

- *Evolutionary fuzzy modeling*: Fuzzy logic systems combine Boolean rules with continuous membership functions. They are capable of making accurate predictions while providing a reasonable level of interpretability [61]. This approach has been successfully applied in practical contexts, including biomarker discovery and cancer diagnosis, leading to a commercial solution for the discovery of interpretable diagnostic signatures.[4]

3.3. Self-Explaining Neural Networks

[2] https://interpret.ml/docs/ebm.html
[3] For instance, see https://interpret.ml.
[4] https://www.quartz.bio

Another strategy to generate explainable/interpretable models that is currently gaining attention are the so-called SENN [35], which, in fact, can be understood as an intermediate between explainable and interpretable models. As pointed out by Hausmann and van Lehn [92], self-explaining is a domainindependent learning strategy that generally should lead to a robust understanding of the domain knowledge. For example, the prediction model is forced to act locally as a linear model, while keeping the non-linearity for the global approach.

Another possibility to achieve self-explaining models is to demand model sparsity while disentangling the factors of variation in the data (so-called generative factors), but preserving all the relevant information for the task to be solved. Model sparsity should result in an inherent model structure such that only a relatively small subset of the latent variables are activated for any given input, whereas disentangled representation may be viewed as a concise representation of the variation in data within the model [93]. Further, additional relevance scores or metrics can be incorporated into the models to evaluate model outcomes for user inspection and interpretation [94].

In summary, and as compared to interpretable approaches, self-explaining models try to balance the trade-off between the original complex model structures and types of local linearity, sparsity, and internal disentangled data representation by using additional penalty constraints or evaluation metrics, whereas interpretable models are straightforwardly designed to be interpretable.

3.4. How to Quantify Interpretability and Explainability

There is not yet a complete consensus on how to evaluate the quality of a method for explanation and interpretation. Evaluation methods for interpretable ML include "real humans on real tasks," proposed by Doshi-Velez and Kim [32] and "AI rationalization" introduced by Ehsan et al. [95]. The quality of a given explanation needs to be evaluated in the context of its task, measuring how much the explanations facilitate and improve decision-making.

In order to compare the properties of different classes of methods for ex-

planation and interpretation, we follow the approach of application-grounded evaluation, combining the concepts reviewed earlier into a common structure starting at the highest level with the *three Cs* of interpretability (Correctness, Completeness, and Comprehensibility). In principle, application-grounded evaluations involve end-user experiments in a real-world application [96], ultimately to assess the extent to which the explanations for the model predictions can be integrated seamlessly with the reasoning model of the domain expert. We use the experience of the authors to list our understanding of the properties of different classes of methods, in best case scenarios—that is, assuming idealized explanation models of the given class. This is summarized in Table 1. Further, we develop a set of evaluation criteria that are amenable to measurement. Thereby, we make reference to the following three conceptual frameworks:

- The three axes proposed by Backhaus and Seifert [97] for radar plots, namely performance (which they call accuracy) together with slimness (model complexity) and interpretability (class typical representations). The latter characterizes compactness with the components of sparsity and grounding, defined below. These two axes closely relate to the functionally grounded evaluation method proposed by Kim et al. [98], which involves evaluation of interpretability without human experiments but relying on quantitative assessments by proxy.

- The two axioms that deep neural networks should fulfil proposed by Sundararajan et al. [99], namely sensitivity and implementation invariance (consistency). Carvalho et al. [96] considers these axioms to correspond to the foundation of Honegger [100] for an *Axiomatic Explanation Consistency Framework* grounded on human intuition. Sensitivity is the same as defined below, and implementation invariance requires that two models for which the outputs are equal for all inputs, should have identical explanations. We focus on explanations of a single model with different initialisations, where this requirement maps onto stability/consistency.

• The three desiderata by Alvarez-Melis and Jaakkola [35], namely fidelity, diversity and grounding. In particular, fidelity and diversity form components of correctness in our framework. The concepts of fidelity and diversity are included in our framework, although we add property of continuity for real-valued features, which is central to the definition in of self-explaining prediction models [35]. Grounding is directly mapped onto the first characteristic of comprehensibility.

The criteria in the last two bullet points serve to define explanation correctness in a quantifiable manner. The full set of desirable properties for explainable and interpretable models is as follows:

1. Correctness [96]: This is how much explanations or interpretations are true to the model predictions and the extent to which their values are uniquely defined for a given prediction task: Specifically, this comprises five characteristics that are listed below:

a. *Sensitivity* (Se)—Axiom 1 in the articles by Carvalho et al. [96] and Sundararajan et al. [99]:

*i***.** If there are two different predictions for two inputs that differ in only a single feature, then this feature should be present in the explanation of at least one of those predictions. The implication is that the difference between the predictions should be associated with a difference in feature values.

*ii***.** If the predictive model never depends on a particular feature value for its predictions, implying that the feature is uninformative or can be treated as noise, then its importance value should be zero.

b. *Stability/Consistency* (St)—Axiom 2 in the articles by Carvalho et al. [96] and Sundararajan et al. [99]: Similar instances should have similar explanations. In particular, the algorithm should be stable for different initializations when making predictions for the same data point.

c. *Fidelity/Faithfulness* (Fi)—Desideratum 1 in the article by Alvarez-Melis and Jaakkola [35]: How accurately the model prediction $f(x)$ can be built from the explanations φ_i, forming a direct connection between the explanations and the original model such that, as far as possible, $f(x) = g(\varphi_1,...\varphi_d)$.

d. *Diversity* (Di)—Desideratum 2 in the article by Alvarez-Melis and Jaakkola [35]: Inputs should be represented by non-overlapping concepts, for instance, with explanations φ_i and φ_j that are orthogonal according to a given measure or with prototypes, which determine a data space partition by means of Voronoi cells (nearest prototype principle) [101].

e. *Continuity* (co): We adopt the definition of Alvarez-Melis and Jaakkola [35]. This is a weakened version of Lipschitz continuity, defined by the requirement that the predictions are difference-bounded by the explanation measures, namely, given model predictions $f(x)$ and quantitative explanations $\varphi(x)$, for every data point x_0 there is an interval with $\delta > 0$ and a scalar $L \in R$ such that $\|x - x_0\| < \delta \Rightarrow \|f(x) - f(x_0)\| \leq L\|\varphi(x) - \varphi(x_0)\|$. This property confers robustness against small perturbations in the data.

2. Completeness [96]: The coverage of the explanation in terms of the number of instances explained is considered to have two states only, local *vs.* global, depending on whether a given explanation applied in a limited region of the input space or across the full range of inputs.
3. Comprehensibility [96]: This evaluates the usability of the explanations, namely, how quick and easy they are to grasp by the end-user.

a. *Grounding/Understandability* (Gr)—Desideratum 3 in the article by Alvarez-Melis and Jaakkola [35]: To which extent the explanations φ_i form class typical representations that have human-understandable interpretations.

b. *Compactness/Sparsity* (Cp) [97]: Describes the complexity of the explaining model, aiming to involve only a minimal set of explanations. In the case of saliency maps for images, this applies to pixels. In general, the aim is to improve comprehensibility by having as compact an explanation as possible.

c. *Efficiency* (Ef) [96]: Immediacy and usability of the human-understandable interpretations.

Figure 1 graphically summarizes this set of desirable properties for explainable and interpretable models.

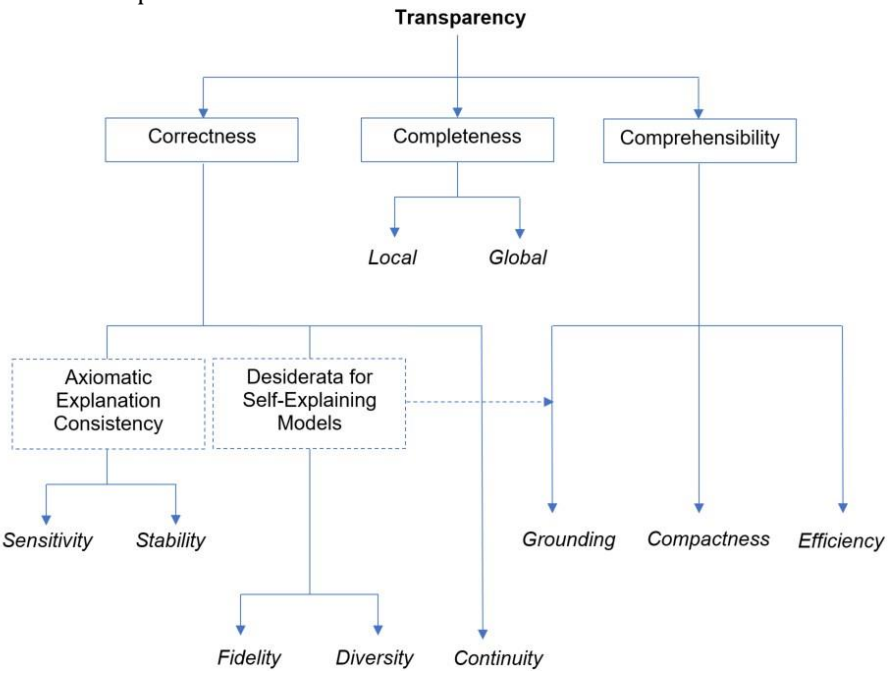

Figure 1: Graphical representation of the hierarchy of the desirable properties for explainable and interpretable/transparent ML models.

Table 1 follows the above hierarchy of desirable properties, preceded by predictive performance for the explained/interpreted model. Further, note that the six desirable properties listed in Section 3 are included in the set of properties listed above. In particular, transparency is considered to be the combination of completeness, compactness, and comprehensibility.

Model	Performance	C1					C2	C3			Comments
		Se	St	Fi	Di	Co		Gr	Cp	Ef	
		Post-hoc									
Feat. attrib. (SHAP)	-	✓	✓	✓	-	✓	local	✓	-	✓	Explanatory only; identification of class relevant inputs but not class representations.
Feat. attrib. (Saliency Maps)	-	✓	✓	✓	-	✓	local	✓	-	✓	
Adversarial and counterfactual samples	-	✓	✓	✓	-	✓	local	✓	-	✓	
Activation maximization	-	✓	✓	✓	-	✓	local	✓	-	✓	
Rule extraction	Moderate	✓	-	✓	✓	-	local	✓	✓	✓	Prediction is by application of the extracted decision tree; discontinuities at rule boundaries.
Metric learning	High	✓	✓	-	-	✓	local	-	-	✓	Prediction by nearest neighbors along geodesics; case-based reasoning.
		Ante-hoc/Intrinsic									
Linear models	Low	✓	✓	✓	-	✓	global	✓	✓	-	Assume linearity in the original inputs; does not form class representations.
Decision Trees	Low	✓	✓	✓	✓	-	local	✓	-	✓	Tree depth can be excessively large.
Bayesian models	High	✓	✓	✓	-	✓	global	-	✓	✓	Bayesian prediction from graph of joint probability distribution.
Prototype methods	High	✓	✓	✓	-	✓	local	✓	✓	✓	Extraction of multiple class prototypes.
Generalized Additive Models	High	✓	✓	✓	✓	✓	global	✓	✓	-	Addition of orthogonal nonlinear univariate and bivariate functions.
Evolutionary Fuzzy Models	High	✓	✓	✓	-	✓	local	✓	-	-	Fuzzy rule-based predictions.

Table 1: Post-hoc and ante-hoc models' desirable properties, with comments. Column descriptions: Performance categorized as *low, moderate*, or *high*. C1: Correctness with sub-categories a)–e); C2: Completeness; C3: Compactness with sub-categories a)–c).

Note that GAMs as well as prototype methods are the only methods that fulfill all of the quantitative properties. The discussion of model predictions with GAMs therefore moves beyond predictive performance and onto the form of the covariate indicator functions φ_i.

4. Interpretable and Explainable ML Models in Light of Regulation and Law

4.1. General Considerations

One of the main regulations currently affecting the use of AI and ML in particular is the European GDPR, as explained in Section 2. It does so through its stated, even if controversial, "right to explanation" on any algorithm-based decisions affecting citizens. According to Wachter et al. [24], the objections to the requirement for explanation include two aspects: ambiguity about what is meant by *meaningful information* and the feasibility of providing this for all ML models. In the case of tabular data, which are prevalent, for instance, in clinical decision support and also in risk models in the insurance and banking industries, it can be argued that meaningful information is potentially available by applying several of the methods reviewed earlier in this section. This includes any models that are interpretable by design, like those described in Section 3.2. Clearly, interpreting more complex data such as images, time series, and free-text raises significant difficulties that are still unsolved at large. Nevertheless, existing *post-hoc* methods, such as saliency maps, at least provide useful confirmatory information in a form that can also raise substantive questions if the regions being classified are incorrectly mapped [102].

But this is by no means the only possible viewpoint. Let us again emphasize the interesting point made in Doshi-Velez et al. [27] about the distinction between ML explanation as a general technical concept and legally-operative explanation, because this is the Gordian knot to be cut in order to smoothly connect technical feasibility and legal compliance. As mentioned in Section 2, legally-operative explanation benefits from elucidating the relative impact of

individual data attributes on model-based decisions. In this case, this would suggest that legal practice should rather seek *post-hoc* explainable systems that emphasize the "concept-mapping" linking relative input relevance with human understanding.

The European Commission's drafted (and soon to be implemented in full) AIA discriminates, as stated in the introduction, between AI applications according to a four-level risk assessment, with any of them even of "limited-risk" being tied to specific transparency obligations. As argued by Fink [103], and in relation to AI explainability, Article 13 in the proposal specifies that high-risk

AI systems are to be developed "to be sufficiently transparent to ensure the user's ability to interpret and use the system's output," but without including any obligations of "AI users to explain or justify the decisions they reach towards those affected by them." This leaves ML practitioners in an awkward position, given that it might seem to favour *ante-hoc* strategies and, thus, simpler models. This would be consistent with the warning by Rudin [17] that

post-hoc strategies might be problematic because explanations could be unreliable. Unfortunately, this completely ignores the fact that, as we next describe in Section 4.2, many domains cannot do without human intelligible *post-hoc* explanation. In this situation, a compromise method such as that provided by

GAMs, in the form of a linear combination of interpretable non-linear functions involving only one or two input variables at a time, as described in Section 3.2, could be both "sufficiently transparent," providing interpretability without renouncing to complexity, and easily explainable.

Regulatory requirements appear to be subject-led, where the so-called data subject is the individual person affected by a decision made by the ML model, for example, the applicant denied a request for insurance or a credit loan. There are other stakeholders in this space who may have a claim to request transparency. Another class of end-user might be an auditor, who may be retrospectively checking whether good practice has been consistently applied in the processing of the requests. In the case of the data-subject or auditor, the gold-standard will

arguably be some form of counterfactual explanation, where there is a direct link between the model and the explanation for the decision. Prototype methods, GAMs, and decision trees might be expected to meet this requirement.

A third category of stakeholder is the domain expert. This may be subject to the doctrine of the learned intermediary (used often in reference to liability in healthcare),[5] whereby it is the decision maker's responsibility to understand the algorithm well enough to know when and how to use it and so to inform users appropriately about any risks involved. It is arguably good practice in this case for the algorithm to ask the user to enter their decision prior to providing an algorithmic recommendation. This can provide a level of protection against decision makers becoming complacent in following algorithms, especially when the algorithm is often correct in its recommendation. Indeed, the onset of complacency is a potential objection against the use of decision support systems in safety-critical domains. This is an area in need of further research [104].

Finally, there is a fourth category, that of AI experts or data scientists, who may want to understand how the method uses the data, sometimes in order to detect issues in observational data, such as spurious correlations.

All in all, and summarizing the discussion of the previous paragraphs, most legal regulations in place or under development are not specific enough as to make clear overall recommendations regarding the preference of either *ante-hoc* or *post-hoc* methods in pursuit of *retrospective transparency* [29]. Furthermore and importantly, these sweeping recommendations might not apply in certain domains such as those commented next, highlighting that transparency regulatory requirements might be quite domain-specific and even, within a domain, stakeholder-dependent.

4.2. Some Example Domains for Interpretability and Explainability Realizations

[5] https://frostbrowntodd.com/the-learned-intermediary-doctrine/

It is generally accepted that interpretability is important in some domains but not in all domains. This has been related by Burkart and Huber [19] to the notion of incompleteness, meaning that the utility of the model requires more than accuracy, but demands also compliance with broader aspects of fairness and ethics [27]. Nevertheless, transparency is a common requirement in highstakes applications.

It has also been argued that we could think of "right levels" of explainability in a given domain where a combination of technical, legal, and economic aspects is used in a three-stage process, including: the definition of contextual factors and stakeholders, the operational context, the potential level of harm that the system could cause, and the legal and regulatory framework affecting the domain [105].

Even beyond domain-specificity of transparency requirements, we must be prepared to face the fact that, as mentioned in previous sections, the diverse stakeholders of the same domain might require different types and levels of explanation [106]. This is clearly discussed by Felzmann et al. [29], where, in the context of GDPR, authors support "a tailored and multi-stakeholder approach to transparency for AI."

In what follows, we illustrate these ideas using two specific but broad and sufficiently different domains: healthcare and the banking industry.

4.2.1. Healthcare

Healthcare is a domain that could be seen as *the canary in the coal mine* from the point of view of the interaction between the practical use of AI and ML and the regulation of interpretability, XAI, and transparency. This is because its obvious social relevance and its ambiguous standing in terms of the level of societal risk involved in the application of AI, as errors can have serious consequences (high-stakes decisions). An early example of the subtleties of interpretable models in healthcare is a frequently cited application for predicting pneumonia risk and 30-day re-admission by Caruana et al. [107]. The fitted model associated

asthma with a lower risk of dying from pneumonia, which is counter to clinical knowledge. However, precisely because of this known risk, asthma patients received more aggressive care, actually resulting in reduced mortality relative to the general population. It is noted by Murdoch et al. [21] that "without having been interpreted, pneumonia patients with asthma would have been deprioritized for hospitalization." This is an example of an unexpected relationship contained in observational data, which requires a contextual interpretation.

A concern recently expressed in respect of clinical medicine is the legal uncertainty that the AIA may create for medical device manufacturers. Tietjen et al. [108] argued that "a comprehensive regulatory framework has already existed for medical devices at the European level for some time in the form of the *European Medical Device Regulation*" (EMDR), also covering AI-based medical devices, but to some extent overlapping the AIA in a way in which, arguably, the European legislator has failed to integrate both regulations to avoid contradictions. A key point made by Tietjen et al. [108] is that the AIA already "stipulates that all AI medical devices that must undergo a conformity assessment procedure by a Notified Body are classified as a high-risk." With most medical devices being based on software, the EMDR already makes AI-based medical software fall into a category that requires it to undergo a conformity assessment procedure. As a result, almost all AI medical devices would be classified as "high-risk AI systems" within the meaning of the AIA.

Needless to say, unless the final European regulation on AI resolves these potential hurdles, ML developers in this application domain will be bound to ensure, amongst others such requirements, that their models fully comply with transparency and interpretability regulations regardless of the clinical medicine problem they aim to tackle. Some leeway in this context can be found in the European Commission's Q&A document for the "New rules for Artificial Intelligence," [6] where we find that "legislative action is needed to ensure a

[6] https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_1683

wellfunctioning internal market for AI systems where both benefits and risks are adequately addressed. This includes [...] AI decisions touching on important personal interests, such as in the areas of [...] healthcare." The requirements for high-risk areas include "transparency and the provision of information to users; human oversight; and robustness," all of which relate to the explainability and interpretability concepts discussed in this paper. Unfortunately, it is not specific enough as to ascertain whether *ante-hoc* interpretable models would be preferred to *post-hoc* explainability. The final form of the AIA can potentially address these matters in line with the earlier proposition that the guiding principle ought to be the "Performance of the Human-AI Team."

Another layer of complexity hides beyond this: the fact that, as previously mentioned, not all the stakeholders in a given domain have the same transparency requirements. In healthcare, stakeholders may include doctors and other medical staff, medical centers, patients, public health institutions and, potentially, private investors, and insurance companies. Doctors will need explanations that are compatible with existing guidelines and workflow, whereas patients will need explanations tailored to their lack of expertise. In turn, insurance companies might only require explanations that might affect their contractual and legal standing on a case. All these actors should be accounted for in the design of transparency strategies (and arguably involved in such design), as proposed, for instance, in the *Social Transparency* perspective developed by Ehsan et al. [109] that "incorporates the socio-organizational context into explaining AI-mediated decision-making." Creating consensus guidelines that involve multiple stakeholders on the application of AI in healthcare would be a paramount task, as those developed for clinical trials by the SPIRIT-AI and CONSORT-AI working groups [110], or even the guidelines for *reporting* AI-based diagnostic results of the STARD-AI steering group [111] and TRIPOD-ML [112].

A straightforward example of the gap between the current state of AI implementation and regulation can be found in the recent efforts by the U.S. Food and Drug Administration (FDA), together with Health Canada and the UK

Medicines & Healthcare Products Regulatory Agency to define some guiding principles for "Good Machine Learning Practice for Medical Device Development" [113]. One of the proposed guiding principles is precisely that "*Focus Is Placed on the Performance of the Human-AI Team,*" according to which "the human interpretability of the model outputs are addressed with emphasis on the performance of the Human-AI team, rather than just the performance of the model in isolation." Another guideline recommends that "*Users Are Provided Clear, Essential Information,*" having "ready access to clear, contextually relevant information that is appropriate for the intended audience." Note that guidelines such as this, issued by a trusted and relevant regulator, even if not mandatory as a law, may have a greater impact on the actual practice in a specific context such as medicine than the law itself. A manifestation of this is the currently limited number of FDA-approved AI/ML-based medical technologies [114]—the first of this in 2016 and with an AI-specific regulatory framework proposed only in 2019 [115]. In this same context, AI explainability mechanisms have been taken out of their user-oriented scope to be suggested as methods to guarantee deployment robustness in medical algorithmic audit frameworks [116].

4.2.2. Banking

The publication of the AIA draft has caused a domino effect of regulatory proposals from which banks are not exempt. Both the European Central Bank (ECB) and the European Banking Federation (EBF) have expressed their views through the publication of position papers [117, 118]. Both documents welcome the regulatory proposal, but qualify the need to clarify different points, including:

- A more specific definition of what is considered AI, in order to distinguish between the different systems and the scope of application.

- The clear identification of when the use of AI systems will be considered high risk and therefore subject to the requirements set out in this Regulation.

- Promotion of measures to support customer education and awareness of the Regulation and the use of AI.

- Clarification of the applicable requirements and competent authorities with regard to outsourcing by users of high-risk AI systems that are credit institutions.

Furthermore, the ECB document invites the Union legislator "to further reflect on the need to designate relevant competent authorities as responsible for the supervision of the conformity assessment conducted by credit institutions" in matters of "transparency and the provision of information to users, human oversight, and robustness" when AI systems are applied. Also interestingly, the EBF document underlines "the need to ensure coherence with the proposal for a revision of the Consumer Credit Directive (CCD)," given that "Recital 48 of the CCD provides for transparency and contest right as well as human intervention in case of automated decisions." This potential conflict with AIA is similar to the one with the EMDR in healthcare (software for medical devices) described in Section 4.2.1.

Conformity assessment has to do with the use of audit frameworks, a concept that also resonates with the banking domain. Beyond customers themselves, the stakeholders in the banking sector may include, among others, the bank's data scientist, usually led by a chief data officer (CDO), internal and external auditors, and several levels of regulators. Auditors can be seen in this context as guarantors of regulation compliance.

Bu¨cker et al. [119] discuss auditability and explainability in the context of one of the banking problems to which ML has historically paid more attention [120]: credit scoring. Authors argue that transparency and explainability are

inextricably linked to the stakeholder involved: analysts and risk managers are likely to be interested in the global understanding of the model provided, for instance, by feature importance methods, while auditors and regulators are likely to be more interested in global explanations—note that, in Europe, this type of transparency requirements are set by the European Banking Authority [121], reinforcing, like in the case of the healthcare sector, the already stated idea that domain-specific regulations can have stronger and more immediate effects than more general laws such as GDPR. Credit officers and customers are argued to be instead more interested in the clarification of individual credit decisions that could be obtained by methods such as SHAP or LIME. Interestingly, authors also link the latter to GDPR's requirement of methods that explain individual predictions so as to "identify an applicant's most adverse characteristics that were negatively contributing to a credit rejection by a given model."

The adaptation of explainability and interpretability to the needs of different stakeholders can also be seen from the viewpoint of scenario-based requirements elicitation, as for the specific case of fraud detection [122], another *classic* application of ML in banking. In the authors' words: "scenarios have the potential to bridge the gap between the social and operational focus with the organizational focus of information systems development." Such scenarios are seen as "narratives on the sequence of events and steps performed by a stakeholder in their daily operations." The key point here is the systematization of this process in five stages, namely: *Identify Stakeholder Settings*, *Identify Stakeholder Goals*, *Identify Stakeholder Tools Capabilities*, *Create Scenarios*, and *Use Scenarios and Identify Requirements*. Using this method, authors stress the importance of tailoring the interpretability and explainability methods to both stakeholder and scenario. For a fraud detection expert, it is concluded that explanations about

how an ML method makes a prediction should be *selective* and understandable, so as not to overwhelm the expert. For that, local explanations provided by *post-hoc* methods are encouraged. Feature importance-based explanations are also advocated.

Strategies for ML interpretability and explainability are also considered for specific scenarios and specific stakeholder by Jiang and Senge [123] using a case study on consumer lending where compliance and legal professionals are the stakeholders—as these have "strong incentives to understand AI decisions [...] largely to comply with regulatory requirements." Here, the emphasis is placed on the difference between *two cultures*: that of *technical* stakeholders (data scientists) and that of *non-technical* ones (who may still be conversant with quantitative analysis in general). Authors found that compliance and legal professionals in this area required *i)* statistically rigorous explanations that were still within their understanding; *ii)* explanations that are actionable and can be used in decision making; and *iii)* explanations that are accompanied by documentation, precise quantitative outputs and robustness tests. Methods tailored to this type of stakeholders are suggested, including feature importance lists for generalized linear models; single decision trees with features, labels, and probabilities on each split; rule-based models and, only if assisted by AI experts, SHAP summary plots and dependency plots. The importance of taking into account users' domain experience and supporting users with limited domain expertise is also emphasized by Dikmen [124], where the author explores interpretability and explainability in financial decision making using *cognitive work analysis* as a domain-centric tool and, particularly, one of its phases: *work domain analysis*, to identify goal-related constraints. These tools are recommended to be used to make data-driven *post-hoc* methods explanations amenable to end-users. This approach is exemplified using random forest classifiers with local rule-based explanations (LORE).

5. Conclusions

AI and, central to it, ML, are becoming increasingly bound by law and regulatory frameworks due to the increasing impact of their use in society at large. These go all the way from generic mandates such as the European General Data Protection Regulation or the forthcoming Artificial Intelligence Act to domain-specific recommendations and guidelines, including from the U.S. Food and Drug Administration.

In one way or another, all these regulations place great relevance on issues of algorithmic trustworthiness, transparency, interpretability, and explainability. Much of this responsibility is placed on the shoulders of the data controllers and data analysts and their ability to engage with domain-specific stakeholders, which implies that the methods required to comply with these obligations must enter a phase of maturity.

The necessary interaction between the concepts of ML interpretability and explainability, on the one hand, and legal regulation, on the other, that we have described in this paper, can be summarized as a list of general recommendations:

- Interpretability and explainability must be understood beyond their technical description to also be considered as tools for regulatory compliance.
- At this time, no standard exists for mapping the abstract transparency requirements in the text of legal regulations to the technical concepts of ML interpretability and explainability.
- The real boundaries of legal requirements will only be substantiated through jurisprudence produced in court. Therefore, public and private ML developments involving model interpretability and explainability must include risk contingency plans to account for the intrinsic uncertainty involved in any attempt to comply with legislation on this matters.
- The development of ML systems that are legally compliant in terms of

model interpretability and explainability in applications of social impact should benefit from the creation of multi-disciplinary teams that add legal and social sciences experts to the usual teams of data scientists.

- Interpretability and explainability requirements are domain-specific and, within a domain, stakeholder- and even scenario-specific; therefore, ML-based systems that are legally compliant on those terms should only be developed once these requirements have been clearly specified.

- The idea of explanation of ML-based systems requires the human-in-theloop, so as to ensure that explanations are provided in a form and at a time that human users can truly benefit from them.

Three key themes for the next stage of maturity of AI systems have been identified in primary evidence-based research on behalf of the Information Commissioner's Office in the UK involving public engagement (citizen's juries) and industry engagement (roundtables) [125]:.

i. Context is key. The expectation of explanations of AI decisions will depend on the impact of the decision, the ability to change it and the data used to inform it. This is confirmed in a further report [126], which points out

> that "the need for explainability must be considered in the context of the broader goals or intentions for the system, taking into account questions about privacy, accuracy of a system's outputs, the security of a system and how it might be exploited by malicious users if its workings are wellknown, and the extent to which making a system explainable might raise concerns about intellectual property or privacy."

ii. Challenges in explaining AI decisions extend beyond technical implementation and include "cost, commercial sensitivities and a lack of internal organisational accountability." However, it is clear that for systems impacting on individual citizens the minimum requirement remains as expressed in the

guidance that accompanies the GDPR text, that the "information provided should be sufficiently comprehensive for the data subject to understand the reasons for the decision."

iii. "There is a desire for education and awareness raising activities to better engage with the public on the benefits and risks of AI in decision-making." Linked to this important premise is the provision of information to practitioners on the relative merits and technical implementation of explainable AI systems, to which this paper contributes.

In this paper, we have first revised and discussed the current regulatory and legal frameworks for the application of AI and ML. This has been followed by a description of explainability and interpretability as *post-hoc* and *ante-hoc* strategies. The former strategies can be seen as depending on developments beyond the modeling itself (often domain-specific), whereas the latter strategies focus on interpretable models that inherently offer the possibility for model explanations. The aim of Section 3 was to show what is possible. A limitation of the methods and metrics described in this section is that they focus mostly on tabular and imaging data. Ultimately, algorithms must conform with good software development practice, including elements of verification (is the model built right?) and validation (is it the right model for the intended purpose?). Only then can ML models be integrated into workflows for routine use—often the key to success in practice. It looks increasingly as though the verification process will have to move from being primarily performance-based to becoming also founded on the internal logic of the operation of the model [127]. This might also be a step towards causal interpretability as it is demanded for forthcoming AI systems [128]. Additionally, the potential links between regulations and *ante hoc* and *post hoc* approaches have also been discussed, and we have broached, using two specific domains as examples, the issues of domain-specificity and user/stakeholderspecificity in transparency requirements. Here, we have highlighted the yet

to be resolved incompatibility between the heterogeneity of requirements this entails and the sweeping homogeneity of regulatory frameworks.

References

[1] Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (1) (2009) 1–127. doi:10.1561/2200000006.

[2] I. J. Goodfellow, Y. Bengio, A. C. Courville, Deep Learning, Adaptive computation and machine learning, MIT Press, 2016.
URL http://www.deeplearningbook.org/

[3] Y. LeCun, Y. Bengio, G. E. Hinton, Deep learning, Nat. 521 (7553) (2015) 436–444. doi:10.1038/nature14539.

[4] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, in: P. L. Bartlett, F. C. N. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States, 2012, pp. 1106–1114.

[5] G. Cybenko, Approximation by superpositions of a sigmoidal function, Math. Control. Signals Syst. 2 (4) (1989) 303–314. doi:10.1007/ BF02551274.

[6] B. Hanin, Universal function approximation by deep neural networks with bounded width and ReLU activations, Mathematics 7 (992) (2019) 1–9. doi:10.3390/math7100992.

[7] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[8] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[9] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117. doi:10.1016/j.neunet.2014.09.003.

[10] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.

[11] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.

[12] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, 2014, pp. 2672–2680.

[13] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. 32 (24) (2020) 18069–18083. doi:10.1007/s00521-019-04051-w.

[14] D. Bacciu, B. Biggio, P. Lisboa, J. D. Mart´ın, L. Oneto, A. Vellido, Societal issues in machine learning: When learning from data is not enough, in: M. Verleysen (Ed.), 27th European Symposium on Artificial Neural Networks, ESANN 2019, Bruges, Belgium, April 24-26, 2019, 2019, pp. 455–464.

[15] European Commission, Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts,

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri= CELEX:52021PC0206 (2021).

[16] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K. Mu¨ller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Vol. 11700 of Lecture Notes in Computer Science, Springer, 2019. doi: 10.1007/978-3-030-28954-6.

[17] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (5) (2019) 206–215. doi:10.1038/s42256-019-0048-x.

[18] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable machine learning: Fundamental principles and 10 grand challenges, Statistics Surveys 16 (2022) 1 – 85. doi:10.1214/21-SS133.

[19] N. Burkart, M. Hubert, A survey on the explainability of supervised machine learning, Journal of Artificial Intelligence Research 70 (2021) 245– 317.

[20] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Com-
puting Surveys (CSUR) 51 (2018) 1–42.

[21] W. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B.Yu, Definitions, methods, and applications in interpretable machine learning, Proceedings of the National Academy of Sciences 116 (2019) 22071–22080.

[22] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable AI: A review of machine learning interpretability methods, Entropy 23 (2020) 18.

[23] B. Goodman, S. Flaxman, European union regulations on algorithmic decision-making and a “right to explanation”, AI Magazine 38 (2017) 50–57. doi:10.1609/aimag.v38i3.2741.

[24] S. Wachter, B. Mittelstadt, L. Floridi, Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation, International Data Privacy Law 7 (2017) 76–99.

[25] J. Cobbe, Administrative law and the machines of government: Judicial review of automated public-sector decision-making, Legal Studies 39 (2019) 636–655.

[26] J. Burrell, How the machine "thinks": Understanding opacity in machine learning algorithms, Big Data & Society 3 (2016) 636–655. doi:10.1177/ 2053951715622512.

[27] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, et al., Accountability of AI under the law: The role of explanation, arXiv preprint arXiv:1711.01134.

[28] G. Malgieri, G. Comand´e, Why a right to legibility of automated decisionmaking exists in the general data protection regulation, International Data Privacy Law 7 (2017) 243–265. doi:10.1093/idpl/ipx019.

[29] H. Felzmann, E. F. Villaronga, C. Lutz, A. Tam`o-Larrieux, Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns, Big Data & Society 6 (1) (2019) 1–14.

[30] O. B. Albu, M. Flyverbom, Organizational transparency: Conceptual-izations, conditions, and consequences, Business & Society 58 (2) (2019) 268–297. doi:10.1177/0007650316659851.

[31] House of Lords, Select Committee on Artificial Intelligence, Report of session 2017-19, AI in the UK: Ready, Willing, and Able?, https://publications.parliament.uk/pa/ld201719/ldselect/ ldai/100/100.pdf (2018).

[32] F. Doshi-Velez, B. Kim, Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608.

[33] L. Floridi, M. Holweg, M. Taddeo, J. A. Silva, J. M¨okander, Y. Wen, capAI - a procedure for conducting conformity assessment of AI systems in line with the EU artificial intelligence act, SSRN e-library, http://dx. doi.org/10.2139/ssrn.4064091 (March 2022).

[34] S. Teso, K. Kersting, Explanatory interactive machine learning, in: 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 239–245.

[35] D. Alvarez-Melis, T. S. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montr´eal, Canada, 2018, pp. 7786–7795.

[36] A. B. Arrieta, N. D. Rodr´ıguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garc´ıa, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115. doi:10.1016/j.inffus.2019.12.012.

[37] M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, R. Rastogi (Eds.), Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[38] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N.

Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 4765–4774.

[39] J. Roder, L. Maguire, R. Georgantas, H. Roder, Explaining multivariate molecular diagnostic tests via Shapley values, BMC Medical Informatics Decis. Mak. 21 (1) (2021) 211. doi:10.1186/s12911-021-01569-9.

[40] C. Izzo, A. Lipani, R. Okhrati, F. Medda, A baseline for shapley values in MLPs: From missingness to neutrality, in: M. Verleysen (Ed.), 29th European Symposium on Artificial Neural Networks, ESANN 2021, Bruges, Belgium, October 6-8, 2021, i6doc.com, 2021, pp. 605–610.

[41] V. Van Belle, B. Van Calster, S. Van Huffel, J. A. K. Suykens, P. Lisboa, Explaining support vector machines: A color based nomogram, PLOS ONE 11 (10) (2016) 1–33. doi:10.1371/journal.pone.0164568.

[42] P. J. G. Lisboa, S. Ortega-Martorell, S. Cashman, I. Olier, The partial response network: A neural network nomogram, arXiv preprint arXiv:1908.05978.

[43] P. J. G. Lisboa, S. Ortega-Martorell, I. Olier, Explaining the neural network: A case study to model the incidence of cervical cancer, in: M. Lesot, S. M. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. BouchonMeunier, R. R. Yager (Eds.), Information Processing and Management of Uncertainty in Knowledge-Based Systems - 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15-19, 2020, Proceedings, Part I, Vol. 1237 of Communications in Computer and Information Science, Springer, 2020, pp. 585–598. doi:10.1007/978-3-030-50146-4_43.

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep` features for discriminative localization, in: 2016 IEEE Conference on

Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 2921–2929. doi:10.1109/CVPR.2016.319.

[45] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: D. J. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, Vol. 8689 of Lecture Notes in Computer Science, Springer, 2014, pp. 818–833. doi: 10.1007/978-3-319-10590-1_53.

[46] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017, IEEE Computer Society, 2017, pp. 618–626. doi:10.1109/ICCV.2017.74.

[47] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schu¨tt, S. D¨ahne, D. Erhan, B. Kim, The (un)reliability of saliency methods, in: W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Mu¨ller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer International Publishing, Cham, 2019, pp. 267–280. doi:10.1007/978-3-030-28954-6_14.

[48] T. Freiesleben, The intriguing relation between counterfactual explanations and adversarial examples, Minds and Machines 32 (2021) 77–109. doi:10.1007/s11023-021-09580-9.

[49] A. V. Looveren, J. Klaise, Interpretable counterfactual explanations guided by prototypes, in: N. Oliver, F. P´erez-Cruz, S. Kramer, J. Read, J. A. Lozano (Eds.), Machine Learning and Knowledge Discovery in Databases. Research Track - European Conference, ECML PKDD 2021, Bilbao, Spain, September 13-17, 2021, Proceedings, Part II, Vol. 12976 of

Lecture Notes in Computer Science, Springer, 2021, pp. 650–665. doi:10.1007/978-3-030-86520-7_40.

[50] M. Pawelczyk, C. Agarwal, S. Joshi, S. Upadhyay, H. Lakkaraju, Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis, arXiv (2021) 1–21doi: 10.48550/ARXIV.2106.09992.

[51] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Havard Journal of Law & and Technology 31 (2) (2018) 1–47. doi:10.2139/ ssrn.3063289.

[52] A. Artelt, F. Hinder, V. Vaquet, R. Feldhans, B. Hammer, Contrasting explanations for understanding and regularizing model adaptations, Neural Processing Lettersdoi:10.1007/s11063-022-10826-5.

[53] P. Perera, P. Oza, V. Patel, One-class classification: A survey, arXiv (2021). doi:arXiv:2101.03064Allfields.

[54] I. Stepin, J. M. Alonso, A. Catal´a, M. Pereira-Farin˜a, A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence, IEEE Access 9 (2021) 11974–12001. doi: 10.1109/ACCESS.2021.3051315.

[55] D. Staps, R. Schubert, M. Kaden, A. Lampe, W. Hermann, T. Villmann, Prototype-based one-class-classification learning using local representations, in: Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN) - Padua, IEEE Press, Los Alamitos, 2022, p. in press.

[56] Z. Zhou, H. Cai, S. Rong, Y. Song, K. Ren, W. Zhang, J. Wang, Y. Yu, Activation maximization generative adversarial nets, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018, pp. 1–24.

[57] J. Despraz, S. Gomez, H. F. Satiz´abal, C. A. Pen˜a-Reyes, Towards a better understanding of deep neural networks representations using deep generative networks, in: Proceedings of the 9th International Joint Conference on Computational Intelligence (IJCCI 2017), SCITEPRESS – Science and Technology Publications, 2017, pp. 215–222.

[58] T. Etchells, P. Lisboa, Orthogonal search-based rule extraction (OSRE) from trained neural networks: a practical and efficient approach, IEEE Transactions on Neural Networks 17 (2006) 374–384.

[59] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (5) (2019) 1–42. doi:10.1145/3236009.

[60] T. R¨ognvaldsson, T. Etchells, L. You, D. Garwicz, I. Jarman, P. Lisboa, How to find simple and accurate rules for viral protease cleavage specificities, BMC Bioinformatics 10 (2009) 149.

[61] C.-A. Pen˜a-Reyes, M. Sipper, Fuzzy CoCo: Balancing accuracy and interpretability of fuzzy models by means of coevolution, in: J. Casillas, O. Cord´on, F. Herrera, L. Magdalena (Eds.), Accuracy Improvements in Linguistic Fuzzy Modeling, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 119–146. doi:10.1007/978-3-540-37058-1_6.

[62] H. Ruiz, T. A. Etchells, I. H. Jarman, J. D. Mart´ın-Guerrero, P. J. G. Lisboa, A principled approach to network-based classification and data representation, Neurocomputing 112 (2013) 79–91. doi:10.1016/j.neucom.2012.12.050.

[63] S. Ortega-Martorell, P. Riley, I. Olier, R. G. Raidou, R. Casan˜a-Eslava, M. Rea, L. Shen, P. J. G. Lisboa, C. Palmieri, Breast cancer patient characterisation and

visualisation using deep learning and Fisher information networks, Scientific Reports 12 (2022) 14004.

[64] I. Misra, L. van der Maaten, Self-supervised learning of pretext-invariant representations, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, Computer Vision Foundation / IEEE, 2020, pp. 6706–6716. doi:10.1109/CVPR42600.2020.00674.

[65] A. Tsymbal, E. Meissner, M. Kelm, M. Kramer, Towards cloud-based image-integrated similarity search in big data, in: Proceedings of the 2014 IEEE-EMBS International Conference on Biomedicine and Health Informatics (BHI), 2014, pp. 593–596.

[66] C. Molnar, Interpretable Machine Learning, Lulu.com, 2020.

[67] T. Y. Kuzan, B. Gu¨zelbey, N. Turan Gu¨zel, B. N. Kuzan, M. S. C¸akır, C. Canbey, Analysis of intra-observer and inter-observer variability of pathologists for non-benign thyroid fine needle aspiration cytology according to bethesda system categories, Diagnostic Cytopathology 49 (7) (2021) 850–855. doi:10.1002/dc.24756.

[68] G. Kass, An exploratory technique for investigating large quantities of categorical data, Applied Statistics 29 (1980) 119–127.

[69] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[70] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[71] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction and Search, MIT Press, New York, NY, 2nd edn., 2000.

[72] M. Kalisch, P. Bu¨hlmann, Estimating high-dimensional directed acyclic graphs with the PC-algorithm, Journal of Machine Learning Research 8 (2007) 613–636.

[73] R. V. Casan˜a-Eslava, S. Ortega-Martorell, P. J. G. Lisboa, A. P. Candiota, M. Juli`a-Sap´e, J. D. Mart´ın-Guerrero, Robust conditional independence maps of single-voxel magnetic resonance spectra to elucidate associations between brain tumours and metabolites, PLoS ONE 15 (2020) e0235057.

[74] S. Saralajew, L. Holdijk, M. Rees, E. Asan, T. Villmann, Classification-by-components: Probabilistic modeling of reasoning over a set of components, in: H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 2788–2799.

[75] D. Nebel, M. Kaden, A. Villmann, T. Villmann, Types of (dis-)similarities and adaptive mixtures thereof for improved classification learning, Neurocomputing 268 (2017) 42–54. doi:10.1016/j.neucom.2016.12.091.

[76] T. Kohonen, Learning Vector Quantization, Neural Networks 1 (Supplement 1) (1988) 303.

[77] A. Sato, K. Yamada, Generalized learning vector quantization, in: D. S. Touretzky, M. C. Mozer, M. E. Hasselmo (Eds.), Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference, MIT Press, Cambridge, MA, USA, 1996, pp. 423–9.

[78] M. Biehl, B. Hammer, T. Villmann, Prototype-based models in machine learning, Wiley Interdisciplinary Reviews: Cognitive Science 7 (2) (2016) 92–111. doi:10.1002/wcs.1378.

[79] P. Schneider, M. Biehl, B. Hammer, Adaptive relevance matrices in learning vector quantization, Neural Comput. 21 (12) (2009) 3532–3561. doi:10.1162/neco.2009.11-08-908.

[80] T. Villmann, A. Bohnsack, M. Kaden, Can learning vector quantization be an alternative to SVM and deep learning? – Recent trends and advanced variants of learning vector quantization for classification learning, J. Artif. Intell. Soft Comput. Res. 7 (1) (2017) 65–81. doi: 10.1515/jaiscr-2017-0005.

[81] T. Martinetz, S. G. Berkovich, K. Schulten, 'Neural-gas' network for vector quantization and its application to time-series prediction, IEEE Trans. Neural Networks 4 (4) (1993) 558–569. doi:10.1109/72.238311.

[82] N. R. Pal, J. C. Bezdek, R. J. Hathaway, Sequential competitive learning and the fuzzy c-means clustering algorithms, Neural Networks 9 (5) (1996) 787–796. doi:10.1016/0893-6080(95)00094-1.

[83] T. Kohonen, Self-Organizing Maps, Vol. 30 of Springer Series in Information Sciences, Springer, Berlin, Heidelberg, 1995, (Second Extended Edition 1997).

[84] A. Aamodt, E. Plaza, Case-based reasoning: Foundational issues, methodological variations, and system approaches, AI Commun. 7 (1) (1994) 39– 59. doi:10.3233/AIC-1994-7104.

[85] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J. Su, *This* looks like *that*: deep learning for interpretable image recognition, in: Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, Canada, MIT Press, 2019.

[86] O. Li, H. Liu, C. Chen, C. Rudin, Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 32,

2018.

[87] C. Lee, M. Samad, I. Hofer, M. Cannesson, P. Baldi, Development and validation of an interpretable neural network for prediction of postoperative in-hospital mortality, NPJ Digital Medicine 4.

[88] R. Agarwal, N. Frosst, X. Zhang, R. Caruana, G. E. Hinton, Neural additive models: Interpretable machine learning with neural nets, in: Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2021, in press, pp. 1–23.

[89] Y. Lou, R. Caruana, J. Gehrke, G. Hooker, Accurate intelligible models with pairwise interactions, in: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 623–631.

[90] P. Ravikumar, J. Lafferty, H. Liu, L. Wasserman, Sparse additive models, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71 (5) (2009) 1009–1030.

[91] C. Br´as-Geraldes, P. A., P. Xufre, Odds ratio function estimation using a generalized additive neural network, Neural Computing & Applications 32 (2019) 3459–3474.

[92] R. Hausmann, K. van Lehn, The effect of self-explaining on robust learning, International Journal of Artificial Intelligence in Education 20 (4) (2010) 303–332. doi:10.5555/2336131.2336132.

[93] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1798–1828. doi:10.1109/tpami.2013.50.

[94] V. Bourgeais, F. Zehraoui, B. Hanczar, GraphGONet: a self-explaining

neural network encapsulating the gene ontology graph for phenotype prediction on gene expression, Bioinformatics 38 (9) (2022) 2504–2511. doi:10.1093/bioinformatics/btac147.

[95] U. Ehsan, B. Harrison, L. Chan, M. O. Riedl, Rationalization: A neural machine translation approach to generating natural language explanations, in: J. Furman, G. E. Marchant, H. Price, F. Rossi (Eds.), Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018, ACM, 2018, pp. 81–87. doi:10.1145/3278721.3278736.

[96] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, Electronics 8 (8) (2019) 1–34. doi:10.3390/electronics8080832.

[97] A. Backhaus, U. Seiffert, Classification in high-dimensional spectral data: Accuracy vs. interpretability vs. model size, Neurocomputing 131 (2014) 15–22. doi:10.1016/j.neucom.2013.09.048.

[98] B. Kim, O. Koyejo, R. Khanna, Examples are not enough, learn to criticize! Criticism for interpretability, in: D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, 2016, pp. 2280–2288.

[99] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: Proceedings of the International Conference on Machine Learning (PMLR), 2017, pp. 3319–3328.

[100] M. Honegger, Shedding light on black box machine learning algorithms: Development of an axiomatic framework to assess the quality of methods that explain individual predictions, arXiv preprint arXiv:1808.05054.

[101] S. Lazebnik, M. Raginsky, Supervised learning of quantizer codebooks by information loss minimization, IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (7) (2009) 1294–1309.

[102] J. Adebayo, J. Gilmer, M. Muelly, I. J. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montr´eal, Canada, 2018, pp. 9525–9536.

[103] M. Fink, The EU Artificial Intelligence Act and access to justice, EU Law Live.

[104] S. Rodriguez, J. O'Donovan, J. Schaffer, T. H¨ollerer, Knowledge complacency and decision support systems, in: 2019 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA), IEEE, 2019, pp. 43–51.

[105] V. Beaudouin, I. Bloch, D. Bounie, S. Cl´emen¸con, F. d'Alch´e Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, J. Parekh, Flexible and contextspecific ai explainability: A multidisciplinary approach, arXiv preprint arXiv:2003.07703.

[106] H. Suresh, S. G´omez, K. Nam, A. Satyanarayan, Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–16.

[107] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp. 1721–1730.

[108] D. Tietjen, N. von Woedtke, E. Schwind, Artificial Intelligence Act (AIA) – legal uncertainty for medical device manufacturers, TaylorWessing insight briefing (30 November 2021).

[109] U. Ehsan, Q. Liao, M. Muller, M. Riedl, J. Weisz, Expanding explainability: Towards social transparency in AI systems, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–19.

[110] S. C. Rivera, X. Liu, A. Chan, A. Denniston, M. Calvert, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension, Nature Medicine 26 (2020) 1351–1363.

[111] V. Sounderajah, H. Ashrafian, R. Aggarwal, J. D. Fauw, A. Denniston, F. Greaves, A. Karthikesalingam, D. King, X. Liu, S. Markar, M. McInnes, Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI steering group, Nature Medicine 26 (2020) 807–808.

[112] G. Collins, K. Moons, Reporting of artificial intelligence prediction models, The Lancet 393 (2019) 1577–1579.

[113] US FDA, Good machine learningpractice for medical device development, https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (October 2021).

[114] S. Benjamens, P. Dhunnoo, B. Mesk´o, The state of artificial intelligencebased FDA-approved medical devices and algorithms: an online database, NPJ Digital Medicine 3 (1) (2020) 1–8.

[115] US FDA, Proposed regulatory framework for modifications to artificial intelligence/ machine learning (AI/ML)-based software as a medical device (SaMD), Discussion paper and request for feedback (2019).

URL https://www.regulations.gov/document?D=FDA2019-N-1185-0001

[116] X. Liu, B. Glocker, M. McCradden, M. Ghassemi, A. Denniston, L. Oakden-Rayner, The medical algorithmic audit, The Lancet Digital Health 4 (2022) e384–397.

[117] C. Lagarde, Opinion of the European Central Bank of 29 december 2021 on a proposal for a regulation laying down harmonised rules on artificial intelligence, Official Journal of the European Union (2022) 2022/C–115/05.

[118] European Banking Federation, EBF position paper on the EC proposal for a regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act), Position Paper (2021) EBF–045345.

[119] M. Bu¨cker, G. Szepannek, A. Gosiewska, P. Biecek, Transparency, auditability, and explainability of machine learning models in credit scoring, Journal of the Operational Research Society 73 (1) (2022) 70–90.

[120] A. Vellido, P. J. Lisboa, J. Vaughan, Neural networks in business: a survey of applications (1992–1998), Expert Systems with Applications 17 (1) (1999) 51–70.

[121] Financial Stability Board, Artificial intelligence and machine learning in financial services – market developments and financial stability implications, https://www.fsb.org/wp-content/uploads/P011117. pdf (2017).

[122] D. Cirqueira, D. Nedbal, M. Helfert, M. Bezbradica, Scenario-based requirements elicitation for user-centric explainable AI, in: Proceedings of the 2020 International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer Cham, 2020, pp. 321–341.

[123] H. Jiang, E. Senge, On two XAI cultures: A case study of non-technical explanations in deployed AI system, arXiv preprint arXiv:2112.01016.

[124] M. Dikmen, A cognitive work analysis approach to explainable artificial intelligence in non-expert financial decision-making, Ph.D. thesis, Waterloo University (2022).

[125] ICO, Project Explain interim report, Technical Report, Information Commissioner's Office.

[126] The Royal Society, Explainable AI: the basics, The Royal Society Policy Briefing, https://royalsociety.org/-/media/policy/projects/ explainable-ai/AI-and-interpretability-policy-briefing.pdf (2019).

[127] P. J. G. Lisboa, Industrial use of safety-related artificial neural networks, HSE – Health & Safety Executive 327 (2001) 1–36.

[128] B. Sch¨olkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, Y. Bengio, Toward causal representation learning, Proc. IEEE 109 (5) (2021) 612–634. doi:10.1109/JPROC.2021.3058954.

Appendix A. Summary of regulations relevant to XAI

DocumentExtractsIETComments				
G20MinisterialStatement onTradeandDigitalEconomyandOECDArtificial IntelligencePrinciples(May, 2019)	**Annexl,Section1,Paragraph1.3:Transparencyand Explainability** "AIActorsshouldcommittotransparencyand responsibledisclosureregardingAIsystems.Tothisend,they shoulddemeaningfulinformation,appropriatetothecontext,andconsistentwiththestateofart[...]."	×	√√	Bothdocuments(G20and OECD)sharethecontentonE andT,withnoreferencetoI. Thisisarecommendationsdocument,asactorsdonothave legislativepower.
UNESCORecommendation ontheEthicsofArtificialIntelligence(23/11/2021) :	**ChapterIII,SectionIII.2:Principles:Recitals37to41: TransparencyandExplainability** "Transparencyisneces saryforrelevantnationalandinternationalliabilityregimesto workeffectively[...].Theleveloftransparencyandexplainability shouldalwaysbeappropriatetothecontextandimpact .Transparencyandexplainabilityrelatecloselyto[...]re sponsibilityandaccountabilitymeasures,aswellastothetrustworthinessofAIsystems."	×	√√	AdoptedbytheCounciland recommended,onavoluntary basis,tomemberstates(as UNESCOhasnolegislative power).Itisfairlyexplicit, evenifgeneric,intermsofthe meaningandreachoftheTand EconceptsinAI.
EuropeanGeneralDataProtectionRegulation(GDPR, 25/05/2018)	**Articles13to15** ondataaccessandcontrollerduties:"The datasubjectshallhavetherighttoobtainfromthecontroller informationabout]theexistenceofautomateddecision-making, includingprofiling,referredtoinArticle22(1)and(4)and,at leastinthosecases,meaningfulinformationaboutthelogicinvolved[...]."	×	√	Aproperlegalmandateof European scopethatfocuseson dataandspecificallyonautomateddecision-making,but notonAIinparticular.
[...] EuropeanArtificialIntelligenceAct(European Commissionversionof 21/04/2021)	**Titel:GeneralProvisions;Articles1and13:** "ThisRegulationlaysdown[...]harmonisedtransparencyrulesforAI systemsintendedtointeractwithnaturalpersons[...].HighriskAIsystemsshallbedesigned[...]toensurethattheiroperationissufficientlytransparenttoenableuserstointerprretthe system'soutput[...]. **TitleIV:Transparencyobligations forcertainAIsystems** "	××	√	Thisisaproposalforlegislative regulationatEuropeanlevel, stillunderdevelopment.Eand Iarebypassedtofocuson T, whichisonlymandatoryfor high-riskAIsystems,asspecifiedinTitleIV,Article52.

TableA.2:Regulationsandrecommendationsrelatedtotransparency,explainabilityand/orinterpretability.Fromlefttoright,columns includethefollowinginformation:First,identificationofthedocument;Second,relevantextractsfromthetexts;thirdtofifth,ticksforyes andcrossesfornoaboutwhetherthetextsexplicitlyaddressestheconceptsofinterpretability(I),explainability(E)andTransparency(T); sixth,commentsonthetext.