Harnessing Prior Knowledge for Explainable Machine Learning: An Overview

Katharina Beckh*, Sebastian Müller†, Matthias Jakobs‡, Vanessa Toborek†, Hanxiao Tan‡, Raphael Fischer‡,
Pascal Welke†, Sebastian Houben§ and Laura von Rueden*

*Fraunhofer IAIS, Germany
Email: {katharina.beckh, laura.von.rueden}@iais.fraunhofer.de
†University of Bonn, Germany
Email: {muellers, toborek, welke}@cs.uni-bonn.de
‡TU Dortmund University, Germany
Email: {matthias.jakobs, hanxiao.tan, raphael.fischer}@tu-dortmund.de
§Hochschule Bonn-Rhein-Sieg, Germany
Email: sebastian.houben@h-brs.de

***Abstract*—The application of complex machine learning models has elicited research to make them more explainable. However, most explainability methods cannot provide insight beyond the given data, requiring additional information about the context. We argue that harnessing prior knowledge improves the accessibility of explanations. We hereby present an overview of integrating prior knowledge into machine learning systems in order to improve explainability. We introduce a categorization of current research into three main categories which integrate knowledge either into the machine learning pipeline, into the explainability method or derive knowledge from explanations. To classify the papers, we build upon the existing taxonomy of informed machine learning and extend it from the perspective of explainability. We conclude with open challenges and research directions.**

***Index Terms*—Machine learning, Taxonomy, Human computer interaction, Knowledge representation**

I. Introduction

The complexity of current machine learning (ML) models prevents humans from understanding the underlying decision rule which was learned during training. Most models can only be scrutinized in terms of the correlation of input and output features, leaving their internal workings opaque. Particularly in high-stake scenarios, the resulting lack of interpretability poses a severe drawback. For example, consider applications in which an AI-supported system predicts patient sepsis or rejects a loan. In these scenarios, insight into the decision process is important to ensure safety, fairness and compliance with legislation [1]. As a consequence, recent work focuses on explainability to improve transparency and trustworthiness of machine learning models [2, 3, 4, 5, 6].

While explanation capabilities have been investigated since the development of expert systems [7], the advancements in research and the prevalent use of ML systems lead to new requirements and expectations. The computational steps a model takes may exactly describe *how* the algorithm comes up with a prediction, but the typical questions are more concerned with *why* a model makes a certain prediction, asking for causal or contrastive explanations [8]. Explanations are by nature context-sensitive as there is no explanation without a question and no question without context. The context of an explanation encompasses not only the matter that is to be explained but also the recipient of the explanation [9, 10, 11]. For example, an explanation of a clinical decision support system needs to be primarily intelligible to the clinician and not to the patient. Explainability methods do not only have to bridge a communication gap between ML systems and experts [12], but their outputs also need to be accessible to a diverse group of users and meet their respective requirements. The context-sensitive nature of the task makes it inherently difficult to develop a generalized explainability method that can be automatically deployed under all circumstances.

If the explainability method itself is not context-aware, fulfillment of this necessity is delegated to the user. Molnar et al. [13] point out pitfalls of explainability methods that require additional knowledge by the user in order to gain reliable new insights and to prevent false conclusions. These methods are part of a larger class of data-driven explainability methods that produce *feature attribution* values for a prediction. Feature attribution refers to the effect and importance of data features on the model prediction. These methods can detect a *correlation* but they do not provide an answer on *why* a feature is relevant to the model output. To answer this question, the user has to apply their own knowledge to put the results into context. However, a layperson might not have the necessary knowledge or the feature values might be intrinsically difficult to interpret, e.g., in the case of raw sensor data like audio signals. Another limitation is that these explanations are data-constrained, implying that they cannot provide insight beyond the data at hand [14].

This research has been funded by the Federal Ministry of Education and Research of Germany as part of the Competence Center Machine Learning Rhine-Ruhr ML2R (01—S18038ABC).

We claim that the integration of prior knowledge is important to overcome those limitations and to provide the user with the necessary context, thus increasing the *accessibility* of an explainability method. In fact, the integration of prior knowledge was already motivated in the 2000s in connection with support vector machines [15] and has been revisited for scientific discoveries from data and ML output [14, 16].

Recent research introduced the notion of informed machine learning (IML), which offers a comprehensive taxonomy on the integration of prior knowledge into ML [17]. IML promotes a symbiosis of data-driven neural networks with knowledge-based approaches such that the strengths of both paradigms are combined, the learning capacity of neural networks with the comprehensibility of prior knowledge. In their work, the authors focused on the classical accuracy-driven learning pipeline and identified explainability as a possible side effect of knowledge integration. In this work, we take the IML taxonomy as a formal basis for further investigation of the effect of prior knowledge on explainability.

Li et al. [3] have addressed a similar point of view and provide a distinction between data-driven and knowledge-aware explainable ML. They subdivide the knowledge-aware approaches into broad categories of general knowledge methods and knowledge-based models. However, the structure from Li et al. [3] mainly considers approaches from a method-centric standpoint while possibilities for knowledge integration are not discussed. Other surveys focus their review on modeling (prior) knowledge [18], or applications in specific domains [14]. So far, there exists no systematic overview that discerns between the different ways to integrate prior knowledge such that it benefits explainable ML.

In this paper, we present approaches that harness prior knowledge to make machine learning models more explainable. Our contributions are summarized as follows:

- We provide an overview on how the integration of prior knowledge benefits explainability in existing work.
- We identify three archetypes of different knowledge integration approaches to facilitate the application and adaptation of these methods.
- We highlight open challenges and research directions.

The framework of knowledge-driven explainable machine learning is schematically displayed in Fig. 1 and also outlines the core structure of the paper. The figure shows three ways to integrate knowledge. The first approach, shown in green, is to integrate knowledge into the machine learning pipeline. Blue exemplifies the integration of knowledge into the explainability method. The yellow arrows show how knowledge can be derived from explanations and then be integrated into the machine learning pipeline.

In Section II, we define and provide background on explainability and introduce informed machine learning with its respective taxonomy. In Section III, we investigate approaches on how prior knowledge can inform explainable machine learning. A discussion is provided in Section IV followed by open challenges and research directions. Finally, we summarize our findings in Section V.

II. Background

We give an overview on current methods to make black-box machine learning models more understandable by providing explanations. In addition, we highlight some limitations of these methods, mainly the fact that measuring their quality is a fundamentally difficult endeavor. Afterwards, we describe a recently developed formalism, how prior knowledge can be incorporated into the machine learning pipeline [17], which builds the basis for our work.

A. Explainable Machine Learning

Explanations constitute an important part of human interactions because, in a societal context, humans are interested in the motivations behind a decision [19]. Following the work by Miller [8], an explanation describes the process of abductive inference as well as the final product, i.e., the answer to a why-question. With the increasing application of machine learning models, explanations are required for multiple reasons: verification of the system, improvement of the system, learning from the system and compliance to legislation [1, 20].

We base our definition of explainable machine learning on prior work [21, 22] in which the authors draw a distinction between *interpretable* and *explainable* machine learning models. The former describes models that demonstrate an inherent transparency. The latter describes models that are incomprehensible by themselves but gain transparency through explanations created by methods dedicated to understanding how the model works [5, 21]. Note that the degree to which a model is inherently interpretable strongly depends on the model size and choice of input features. A small decision tree is self-explanatory and thus constitutes an interpretable model. However, with increasing model size, the tree becomes less interpretable, forcing the user to resort to explainability methods in order to obtain insights into the model behavior. Moreover, if input features do not correspond to human semantic concepts, the interpretability of the decision process suffers as a result, no matter how simple the model is. We differentiate the interpretability of the model at three different levels, namely at the level of the entire model, individual components or the training algorithm according to [19]. We also highlight the integration of additional knowledge as one interpretable component of the model (cf. Subsection III-A).

Currently favored models, such as neural networks and random forests, exemplify the need for explainability methods since even small versions are too complex to be interpretable in the sense as defined above. This is why they are often referred to as black-box models. In order to explain the behavior of a black-box model for individual predictions, there are three main approaches: First, attribution-based methods were developed to estimate the respective contribution of each of the input features to the prediction [23, 24, 25, 26]. In the area of computer vision, these values are often visualized as heatmaps. Some of these methods are theoretically grounded in game theory via the concept of Shapley values [27, 28], which quantify the contribution each player has to an outcome [23, 24]. One downside of these approaches is that they merely show which

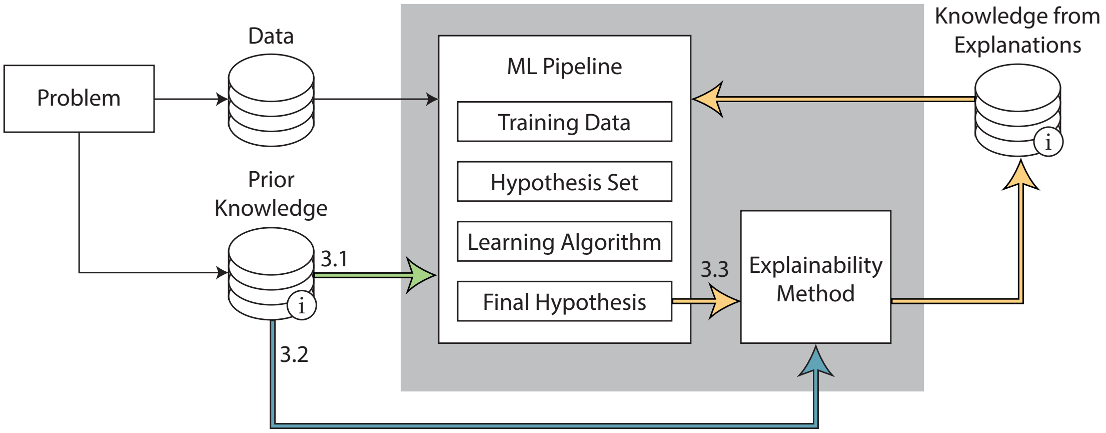

Fig. 1. An overview on how prior knowledge can be harnessed for the framework of explainable ML. Three different ways of integrating prior knowledge were identified, as depicted by colored arrows. They determine the structure of this paper, with corresponding methods being discussed in IML to increase Explainability (Subsection III-A), Informed Explainability Methods (Subsection III-B), and Deriving Knowledge from Explanations (Subsection III-C).

part of the input the model deemed important but not for which reasons [29].

The second approach consists of training an interpretable surrogate model, e.g., a decision tree or a linear model, to mimic the black-box as closely as possible, including its errors [30, 31]. As mentioned before, the interpretability of the surrogate model is itself dependent on its size. The bigger the surrogate model that is needed to accurately mimic the complex behavior, the less explainable it will be in the end and therefore miss its intended use to provide model insights.

The third approach comes from the recent push toward counterfactual explanations, following literature from social science, as pointed out by Miller [8]. Miller argues that humans tend to ask contrastive questions to gain an understanding of the underlying decision process behind an external rationale. Consider as an example a scenario where a credit scoring model is used to decide whether or not a person qualifies for a loan. A question that a customer might ask is, *”Why was my loan denied while my neighbor got their loan?”*. The customer hereby expects as an answer the most critical differences that lead to different predictions. Counterfactual explanation methods rely on this theory to construct artificial data points, which are very similar to the data point in question but lead to a different model prediction [32, 33]. One of the main challenges is the generation of plausible yet minimally changed counterfactual examples [33], that allow users to derive a possible course of action [34].

While all these approaches aim to explain the unknown inner workings of a complex model, it is often difficult to judge whether or not they are accurately following the model’s line of decision or not. Ground truth explainability data cannot be provided in most scenarios since knowledge about the decision process of the model is needed, which is exactly what explainability methods try to uncover. Some work suggests presenting users with explanations generated from explainability methods and observe how well they help users make the initial prediction [35, 36]. However, it is unclear whether or not an explanation that appears to make a model more understandable to the user is also the correct explanation. One way to remedy this conundrum of missing ground truth data might be to incorporate prior knowledge into the training and explanation process. By providing the model with information about how to navigate the path from input to prediction, it might learn a human-understandable way and become more interpretable.

B. Informed Machine Learning

The motivations for integrating prior knowledge into the machine learning pipeline can be manifold. Natural goals are to improve the model performance or to train with less data. With trustworthy artificial intelligence becoming more important [37], another purpose of IML is to ensure knowledge conformity or to improve the interpretability of a model [17].

In contrast to traditional ML that uses prior knowledge implicitly, e.g., for feature engineering or selecting hyperparameters, IML makes the integration of prior knowledge more explicit. IML can be defined as learning from a hybrid information source that consists of data and prior knowledge. Here, the prior knowledge stems from a data-independent source, is brought into a formal representation and is explicitly integrated into the machine learning pipeline.

The taxonomy of IML provides a framework for classifying its different approaches with respect to the knowledge source, the knowledge representation, and the integration stage in the learning pipeline [17]. The authors describe the spectrum of informed learning in terms of the following building blocks:

1) Knowledge source: Here, knowledge is understood as information about relations between entities in specific contexts. The *source* can be categorized into three types:

- **Scientific knowledge**: This knowledge source includes, e.g., natural sciences and engineering. The knowledge is usually formalized and empirically validated.

- **World knowledge**: This knowledge type refers to vision, linguistics or general knowledge, for example, that a tree has leaves.
- **Expert knowledge**: This knowledge type describes the intuitive knowledge acquired by experts through experience. It is informal and often implicit.

2) Representation: Knowledge can be formalized using *representations*, such as equations, simulations, rules, or graphs, but it can also be given more informally via human feedback. Von Rueden et al. [17] found that each knowledge type has common representation types. For example, scientific knowledge is often represented as algebraic equations or simulation results, whereas world knowledge is often represented as logic rules and knowledge graphs. Expert knowledge is commonly represented as human feedback or probabilistic relations.

3) Integration: The representations can be *integrated* into one of the four stages of the learning pipeline (cf. Fig. 1):

- **Training data**: In contrast to the typical way of incorporating knowledge via feature engineering, an informed approach is defined as *hybrid* by using both the original data set and an additional, separate knowledge source.
- **Hypothesis set**: The integration into the hypothesis set is accomplished through the selection of architecture and hyperparameter settings.
- **Learning algorithm**: Through a loss function and appropriate regularizer, additional knowledge can be integrated into the learning algorithm.
- **Final hypothesis**: Existing knowledge is used to compare, benchmark and post-process the output of a model.

Interpretability and explainability are only considered as a side effect in the described IML taxonomy. We will now extend this taxonomy which in turn allows us to identify separate directions of informed learning for explainable machine learning models.

III. Knowledge-driven Explainable ML

As motivated in the introduction, explanations always have to bridge a communication gap between the model and the receiver of the explanation. If the explanation fails to communicate in concepts intelligible to the receiver, the receiver will not be able to understand the explanation. Furthermore, we have noted that purely data-driven explanations put the responsibility to draw reliable conclusions fully on the user, thereby limiting the relevant user group to experts. We propose to address both shortcomings by integrating prior knowledge into the ML pipeline or the explanation, to adapt the explanation to users and contexts. Based on our literature search, we identified three main approaches to incorporate prior knowledge for improving explainable learning systems (colored arrows in Fig. 1):

1) Informed Machine Learning to increase Explainability
2) Informed Explainability Methods
3) Deriving Knowledge from Explanations

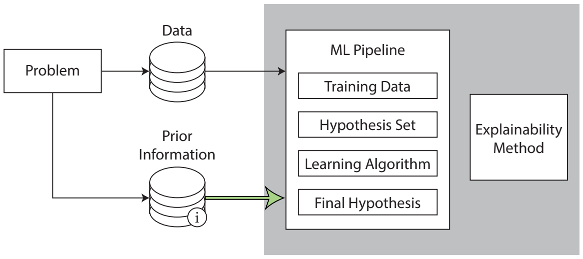

Fig. 2. In addition to the data set used in the ML setting, prior knowledge is integrated into the ML pipeline (green arrow).

The first approach (Fig. 1 green arrow, Subsection III-A) is well-captured by the IML taxonomy. It covers cases where some form of prior knowledge is available in addition to data, and which is integrated at some stage of the learning pipeline. In the works we review in this section, the integrated knowledge is used to align model components with prior knowledge. This facilitates model handling for professionals from the domain, who might not even be ML experts. In the second approach (Fig. 1 blue arrow, Subsection III-B), knowledge integration takes place at the explainability method. We consider this to be an additional component of the usual ML pipeline, due to the prevalence of post-hoc approaches. This integration approach offers the largest potential to increase accessibility of explanations for different user groups in different contexts because it allows for interactions between users and model explanations. Lastly, we found that many published works derive additional knowledge from explanation results (Fig. 1 yellow arrow, Subsection III-C). Similar to prior knowledge that is available from the start, these newly derived priors can then be re-integrated at any point of the pipeline. Approaches in this category are adapted by ML scientists and developers to debug and improve models. With the latter two approaches, we go beyond the information flow established in the IML taxonomy and thus propose an extended framework, as shown in Fig. 1.

We now discuss the related work that has been published for all approaches. The inclusion criteria for the papers were that they use an IML approach and claim improved explainability. During our investigations, we found that many published works can be neatly categorized via their knowledge integration type, i.e., training data, hypothesis set, learning algorithm, and final hypothesis. Where applicable, we structure the subsections accordingly.

A. Informed Machine Learning to Increase Explainability

In this category, additional information in the sense of IML is used in such a way that it not only increases model performance but also improves explainability. While this improvement is not always stated as an explicit objective, we argue that additional knowledge is oftentimes integrated in the sense of an individual, interpretable component, therefore increasing the interpretability of the entire ML pipeline.

a) Training Data: Two approaches in the field of recommender systems increase interpretability by integrating additional knowledge, which is understandable for humans, into the training data for their models. The general task is to predict the next item(s) a user interacts with, given an (ordered) set of user-item interactions. Wang et al. [38] and Ma et al. [39] both first create a single heterogeneous knowledge graph by connecting the user-item interaction data with multiple existing knowledge bases (MovieLens-1M + IMDb and Freebase + DBPedia respectively). Wang et al. [38] generate recommendations by extracting limited-length user-item paths from the graph and rating them for plausibility using a recurrent neural network. Ma et al. [39] compute sets of rules, where a rule is a sequence of certain types of edges, and learn a weighting of those rules. Both methods are argued to be explainable because the algorithms are forced to reason along the edges of the knowledge graph and produce a weighting that reflects the contribution of each path or rule to the decision.

b) Hypothesis Set: We found several applications that promote the interpretability of the model by using prior knowledge to inform architectural changes in the ML model. Two of them are located in the field of biology. Ma and Zhang [40] encode biological knowledge into the network in the form of factor graphs, representing either genes or gene ontology as neurons and abstracting direct influence to corresponding genes or gene ontologies as edges. They hereby give semantic meaning to all originally meaningless neurons and their connections. Such a network, constructed based on prior knowledge rather than heuristics, is easily intelligible and therefore more explainable. To learn representations of single-cell RNA-sequence data, Rybakov et al. [41] propose an interpretable autoencoder based on a regularized linear decoder. The autoencoder decomposes variations into interpretable components using prior knowledge in the form of annotated feature sets obtained from public databases. Observed covariates, such as batch or cell type, can be fed into the encoder-decoder architecture or simply weighted by a linear model and then introduced into the autoencoder. As the primary purpose of the method is to explain the components of variations, introducing prior knowledge enables more acceptable interpretations.

Prior knowledge in the form of world knowledge improves ML models for the problems of semantic image understanding [42] and conversation generation [43]. Chen et al. [42] create a pipeline that processes the visual cues of an image input as well as the background knowledge of a guide ontology. The result is a directed graphical model that constructs possible relationships between the visible objects. This ML model becomes more interpretable because all resulting relationships are verifiable through the ontology. Liu et al. [43] aim to make the process of conversation generation more transparent by integrating a factoid knowledge graph into the deep learning pipeline that is augmented with information from related text documents. While the knowledge graph provides background knowledge for an encoder-decoder model, it also makes the ML model more interpretable, because all graph traversals for knowledge selection can be retraced.

Chen et al. [44] propose a replacement for batch normalization layers, commonly found in neural network architectures, called *concept whitening* layers. The data points flowing into the layer first get decorrelated using a whitening operation and then aligned in the latent space to a fixed number of predefined concepts. As an example, the authors train a convolutional neural network (CNN) for image classification. After swapping all batch normalization layers with concept whitening layers, they use a separate data set, labeled with human-understandable concepts such as *aeroplane* and *table*, to fine-tune the alignment of the training data to these concepts. This alignment to human-interpretable concepts helps not only in debugging the training process, i.e., discovering misalignment between similar concepts, but also increases the understandability of the decision process because it allows for breaking down the decision to a mixture of known and understandable concepts.

c) Learning Algorithm: Several applications use knowledge from the medical domain to apply regularizations to the learning algorithm. The knowledge graphs used by these applications are ontologies, such as ICD-9 or SNOMED-CT, transformed into a tree structure. Choi et al. [45] and a subsequent extension proposed by Ma et al. [46] predict future diseases of a patient based on the diagnosis history of that patient. They compute a vector embedding of the ontology and derive a feature representation of the inputs by accumulating relevant parts of the ontology embedding via an attention mechanism which can be seen as an explanation capability. Jiang et al. [47] use logistic regression to infer the readmission probability of a patient after a hospital stay from their medical history. A distance measure over the ontology is included as a regularization factor in the loss function that penalizes biases toward a certain part of the ontology. Yan et al. [48] formulate a joint learning task of multi-label assignment to CT images and retrieve images similar to the input from a database. The retrieved images serve as explanations. Mutually exclusive label combinations are extracted from the ontology and this information is used to regularize the loss function of the retrieval task, this way aligning the explanations closer to the ontology.

Another two approaches in the domain of computer vision integrate additional knowledge as constraints in the ML pipeline. Donadello et al. [49] present Logical Tensor Networks, a neural network for semantic image interpretation constrained by first-order fuzzy logic. These logic rules are derived from the comprehensible WordNet ontology [50], describing part-of relations which are used to exclude classifications showing unrealistic relations like, e.g., *tail* as a part of *table*. For the problem of part localization, Zhang et al. [51] present an approach that leverages human feedback to improve a ML model. The model consists of an And-Or graph that is based on a pre-trained CNN. This graph disentangles the hierarchical relationship between semantic image parts (top level nodes) and single activated CNN units (terminal nodes). In a second step, they visualize the network's activations using up-convolutional networks and evaluate them via human feedback. This additional information improves the And-Or graph by excluding activations that do not contribute to the target semantic part.

d) Final Hypothesis: The following applications use external, human-understandable knowledge to perform plausibility checks on the results of the ML models. Doran et al. [12] make a conceptual proposition to extend an already existing, explainable model by a post-processing step that checks the explanation against a knowledge base for plausibility. A rudimentary realization of this idea is to improve the output of a multi-object detector [52]. First, the authors define a number of categories to classify the objects in the knowledge base. Subsequently, they compute a ranking that describes how closely related two categories are. Finally, they refine the output of the multi-object detector by collecting information about which categories are detected and artificially increasing the certainty score for objects that belong to related categories.

Kim et al. [22] propose to utilize an external data set of predefined concepts, for example in the form of images, to test how much the networks latent representations align with these concepts. To do so, they take the latent representation of these concepts together with representations of random samples and train a linear classifier to distinguish between concept-related and random vectors. Using the normal to the decision boundary, they can quantify how much a data point from the original training data set aligns with the concepts. This approach is similar to Chen et al. [44], in that it improves the interpretability of the model by measuring the alignment of data points to human-understandable concepts in the model. Whereas Chen et al. [44] use the external concept data set to fine-tune the network to ensure the alignment, Kim et al. [22] measure the alignment to the concepts after training the model in the conventional way.

e) Section Summary: The reviewed papers show that the integration of knowledge can improve explanation capabilities. For each stage in the pipeline, the explainability can be improved by enforcing alignment to human understandable concepts. For the training data, the data themselves are connected to semantic knowledge, e.g., in the form of knowledge graphs. In the stages of hypothesis set and learning algorithm, the model structure or representations are aligned with prior knowledge which brings the internals of a model closer to comprehensible concepts. In the final hypothesis stage, the prior knowledge enables a plausibility check for model explanations.

B. Informed Explainability Methods

The research work presented so far integrates knowledge into the learning pipeline. Moving onward, we now investigate cases in which prior knowledge helps when designing and executing explainability mechanisms, which we purposefully do not consider as a part of the pipeline (cf. Fig. 1). This is an important and unique way of integrating knowledge into ML, that extends the IML taxonomy by adding the *explainability method* as a new integration type. All methods in this section are post-hoc methods which we subdivide further into two categories: formalized priors for explanations and interactive explanations.

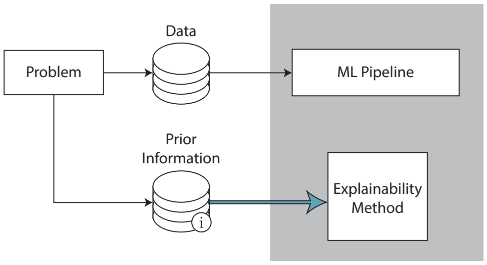

Fig. 3. Prior knowledge is integrated into the explainability method (blue arrow).

a) Formalized Priors for Explanations: Generating counterfactuals is one direction of providing example-based explanations [53]. While many approaches search counterfactuals for the original instances based on distance measures, Mothilal et al. [54] narrow down the search space via user-defined causality constraints. They formalize their priors in the form of box constraints on feasible ranges. Another work [55] attempts to assure the causal plausibility of counterfactuals by incorporating a penalty in the optimization process for infeasible values. For causal relationships that cannot be expressed with formulas, they train a variational autoencoder that generates counterfactuals and evaluates their quality based on human feedback. Both approaches promote the causal plausibility of generated counterfactuals by introducing prior knowledge into optimization in the form of algebraic constraints.

Another post-hoc explainability method is activation maximization (AM). It attempts to discover the ideal input distribution of a given class (global explanation) by optimizing the gradients of the inputs while freezing all parameters of the networks. However, AM with no priors tends to generate meaningless mosaics in high frequency, which are not human-recognizable. Two approaches [56, 57] employ ℓ_2-norm regularization to reduce noise and refine larger structures in the resulting images. Mahendran and Vedaldi [58] constrain the total variation of the explanation to prior images, thus, providing a smoother output. Moreover, Yosinski et al. [56] penalize nonsensical high-frequency pixels by applying Gaussian blur kernels to activations before each optimization step. All of the methods detailed above generate more human-understandable explanations by incorporating additional algebraic restrictions into the optimizers.

Shams et al. [59] propose a methodology which extracts conditional rules from deep neural networks and combines them with other data-driven and knowledge-driven methods. Experts are able to directly validate and calibrate the extracted rules with their domain knowledge to yield more precise and acceptable explanations.

Another approach combines *LIME* (Local Interpretable Model-Agnostic Explanations) [26] with Inductive Logic Programming [60] to obtain verbal explanations for image classification [61]. Extracting symbolic rules from images enables a different perceptual modality and more expressive explanations, such as spatial relations of image parts.

b) Interactive Explanations: The IML taxonomy considers human feedback as a valid representation type, with corresponding sources usually being world or expert knowledge. This is rooted in the success of incorporating human interaction within learning processes. As an example, the framework of coactive learning [62] allows users to correct and thus improve model predictions via direct feedback. Work has also been done on making the human interaction robust against manipulation, for example in the medical domain [63]. Other approaches use visual analytics to analyze and possibly obtain information on how to refine trained models [64, 65].

Recently, human interaction was identified to possibly benefit the explainability of ML systems [66]. This is based on the observation that, in communication between humans, personalized explanations or even explanation dialogues result in better understanding and higher acceptance [8]. Accordingly, enabling users to interact with provided explanations can potentially fulfill several explainability desiderata, i.e., requirements for explaining methods as identified by fact sheets [67].

Different works have successfully developed ML systems that provide interactive explanations for their decisions. The *What-If Tool* [68] allows its users to interactively analyze a model and to find explanations via feature importance. Krause et al. [69] also allow for the interactive exploration of model decisions via underlying feature importance. Moreover, they offer the option of tweaking feature values in order to see how it affects the predictions. Customization of explanations has also been successfully implemented in the *MUSE* (Model Understanding through Subspace Explanations) framework [70]. It allows users to interactively choose features of interest, and then explore how the model behaves in resulting subspaces. Another interactive explanation system is *Glass-Box* [71], which offers explanation dialogues via a voice-enabled virtual assistant. Schneider and Handali [72] proposed a conceptualization for methods that shall provide personalized explanations.

Besides showing the mere feasibility of interactive explainable ML, those works also evaluated the impact on users. The users seemed to obtain a better and also faster understanding of the underlying model logic [70]. By using interactive tools, users were able to improve the predictive model quality [69]. Generally, users were satisfied with the received interactive explanations, but they criticized the lack of arguability [66].

c) Section Summary: Informed explainability methods present a way to customize explanations to concrete applications and users. Formalized priors introduce desired properties into an explanation, such as causal plausibility or noise reduction. Interactive explanations are primarily implemented with user interfaces that allow for personalized explanations via the interaction between user and machine.

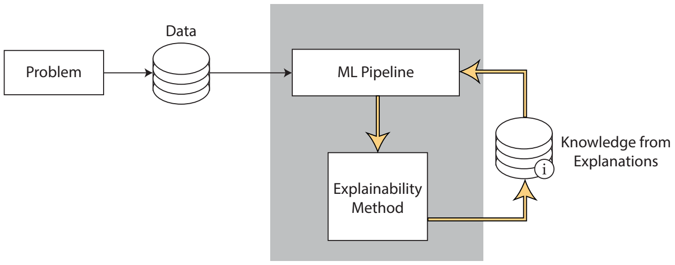

Fig. 4. From the explanations of an explainability model, knowledge is derived, formalized and subsequently incorporated into the ML pipeline (yellow arrow).

C. Deriving Knowledge from Explanations

Explainability methods often detect flaws in ML models, and as such, inform about necessary improvements. One prominent use case is to remedy the Clever Hans phenomenon [73] which refers to some models latching onto spurious correlations in the data set, instead of a known, correct relation that is present in the data and obvious to humans. As an example, the authors show an image containing a photographer's watermark in the bottom left corner on most images featuring a horse. The model picked up on the artifact and focused its decision on the watermark instead of the horse.

We have found that many approaches suggest explicit ways of formalizing and incorporating feedback on explanations that go beyond simply showing the necessity of model improvement. In essence, knowledge is derived from the explainability component of the model and is subsequently integrated into the learning pipeline (cf. Fig. 4). We structure our review according to the integration of this new knowledge into different steps in the ML pipeline.

a) Training Data: Some approaches use feedback on explanations for revising and improving the corpus of available training data in order to inform the learning system. This is usually achieved with a human-in-the-loop, who inspects visual explanations for data instances, revises them if necessary, and thus adds more training data [74]. In terms of IML, this can be understood as obtaining prior knowledge from an expert user (or from world knowledge) at the final hypothesis which is represented as human interaction. This knowledge is then used to improve the training data for the next learning iteration.

Schramowski et al. [75] and Teso and Kersting [76] generate additional training data using feedback on explanations. If an expert decides that the explanation of the model is incorrect, the respective part of the image is used to create counterexamples. Counterexamples are generated by randomizing or otherwise altering the parts an expert identified as unimportant. Each counterexample, however, retains the label of the original data point, encouraging the model to unlearn the unwanted correlation. Instead of adding new data points, the so-called *Explanatory Debugging* [77] lets the user correct mislabelled data points in an interactive way during training.

Adilova et al. [78] use expert feedback to help a model learn from unlabeled data in the context of relation extraction from text data. Examples of relations are *son-of* or *country-of-birth*. The model is first trained to extract the relations itself using distant supervision. Afterwards, tri-grams are extracted as explanations of the model behavior using a version of relevance propagation [79]. An expert is presented with the tri-grams and is asked to decide if they are representative of the relation or not. If they are not, then sentences in the data set containing the tri-gram will be filtered out and the model is retrained.

b) Learning Algorithm: The feature-based feedback from the Explanatory Debugging approach [77] is incorporated as a Bayesian prior, meaning that human feedback at the final hypothesis is transformed into a probabilistic representation, which is then integrated into the learning algorithm.

We found a line of research focusing on the integration of explanations in the learning algorithm by adding a regularizing term to a model's loss function. In addition to the training loss between the model's prediction and the ground truth label, an additional loss between the model's explanation for each prediction and a given explanation is added and thus simultaneously minimized.

To the best of our knowledge, this approach was popularized by Ross et al. [80], who gather binary annotation masks from experts, indicating parts of the input the model should not focus on. These annotations are then used to penalize the gradient of the prediction w.r.t. to specific inputs which are nonzero in the annotation mask. This forces the model to minimize the gradients at the selected locations as part of the training algorithm, which should then lead to the model being *right for the right reason* [75, 80].

Rieger et al. [81] use a decomposition-based approach to measure the importance of certain inputs to a model's decision. Using expert annotations similar to Ross et al. [80], they force certain inputs to be regularized to have zero impact on the model. They test their approach on skin cancer screening images by enforcing that colorful patches next to potentially cancerous skin lesions should be ignored since they do not inform about the type of lesion and only occur in one of the classes. In evaluating their method, they found that the model significantly outperformed the approach of Ross et al. [80] after being forced to ignore the colorful patches. A similar approach was proposed by Selvaraju et al. [82], in which the authors evaluate the effect of human intervention on visual explanations in the application domains of visual question answering and image captioning. They found, after incorporating the changed explanations, that their model was not only able to correctly highlight the images responsible for the correct answer or image caption but also outperformed other state-of-the-art question answering systems. In addition to visual feedback, Stammer et al. [29] offer the user the opportunity to provide semantic feedback in the form of relational functions. Similar to Rieger et al. [81], Erion et al. [83] use attribution priors to optimize for desired explanation qualities, such as smoothness and sparsity, using the attribution method *expected gradients*. With this method, assumptions, e.g., that neighboring pixels should have a similar effect on the output, can be integrated into the model.

Balayan et al. [84] provide a neural network-based framework that jointly makes predictions as well as associated explanations. Subsequently, the output is validated by human experts and the model is improved by adjusting its parameters through backpropagation. Using ground truth data, the authors claim that human feedback increases the prediction quality of the explanations by over 13%. For a natural language inference task, Camburu et al. [85] incorporate explanations of textual entailment, i.e., whether a premise sentence entails a hypothesis sentence, into the training process using a negative log-likelihood for classification and an explanation loss. Furthermore, a feature attribution method for text classification adds a loss with the goal to mitigate unintended bias in the text [86].

c) Section Summary: To summarize, most approaches we found use a modified loss function, which takes into account feedback on the model's explanations, to regularize the model's behavior toward the desired outcome.

IV. Discussion and Outlook

Table I gives an overview of all papers presented in Section III and advances their categorization according to the IML taxonomy by considering the knowledge representation type. We want to point out that the ”Informed Explainability Methods” in Subsection III-B constitute a new integration stage. This results in an additional column solely populated by publications from that section. The table shows a concentration of work using additional knowledge represented either as knowledge graphs or as human feedback. We did not find any work using simulation results or differential equations. Also, only two papers we identified deploy logic rules as a representation type.

We now assess the strengths and limitations of integrating prior knowledge for explainability, both from a general perspective and for each of the three presented strands separately.

a) General Considerations: We established three strands for integrating prior knowledge for explainable machine learning. They can be distinguished with regards to how prior knowledge is available and at which point it is integrated. Prior knowledge is either given as independent data or it is derived from an explainability method. In the first case, the knowledge is then integrated into either the machine learning pipeline or the explainability method. In the latter, the knowledge is only integrated into the machine learning pipeline.

We hypothesize that the prevalence of knowledge graphs and human feedback as representation types is due to their inherent intelligibility. Knowledge graphs allow for representing scientific as well as world knowledge while their benefit lies in the structured representation of world knowledge. A comprehensive review of knowledge graphs as tools for explainable machine learning is given in [18]. Human feedback, on the other hand, is the most accessible knowledge representation for expert knowledge, a knowledge type that is usually more intuitive and less formal. While this reasoning seems plausible to us, we cannot exclude the possibility of falling victim to a

TABLE I
FORMS OF KNOWLEDGE INTEGRATION; COLORS INDICATE HOW EXPLAINABILITY IS ADDRESSED OR TAKEN INTO ACCOUNT: IML IS USED TO INCREASE EXPLAINABILITY (SUBSECTION III-A), KNOWLEDGE IS USED TO ENHANCE THE EXPLAINABILITY METHOD (SUBSECTION III-B) OR EXPLAINABILITY IS USED TO DERIVE AND INTEGRATE KNOWLEDGE (SUBSECTION III-C).

	Training Data	Hypothesis Set	Learning Algorithm	Final Hypothesis	Explainability Method
Knowledge Graphs	Wang et al. (2019) [38] Ma et al. (2019) [39]	Ma and Zhang (2019) [40] Chen et al. (2012) [42] Liu et al. (2019) [43]	Choi et al. (2017) [45] Ma et al. (2018) [46] Jiang et al. (2019) [47] Yan et al. (2019) [48] Zhang et al. (2017) [51] Erion et al. (2021) [83]	Doran et al. (2017) [12] Pommellet and Lécué (2019) [52]	
Logic Rules			Donadello et al. (2017) [49]		Rabold et al. (2019) [61]
Algebraic Equations		Rybakov et al. (2020) [41]	Erion et al. (2021) [83] Rieger et al. (2020) [81]	Kim et al. (2018) [22]	Mothilal et al. (2020) [54] Mahajan et al. (2019) [55]
Probabilistic Relations		Chen et al. (2020) [44]	Erion et al. (2021) [83]		
Human Feedback	Baur et al. (2020) [74] Schramowski et al. (2020) [75] Teso and Kersting (2019) [76] Camburu et al. (2018) [85] Kulesza et al. (2015) [77]		Ross et al. (2017) [80] Rieger et al. (2020) [81] Selvaraju et al. (2019) [82] Camburu et al. (2018) [85] Liu and Avci (2019) [86] Kulesza et al. (2015) [77]	Balayan et al. (2020) [84]	Shams et al. (2021) [59] Sokol and Flach (2020) [66] Krause et al. (2016) [69] Wexler et al. (2020) [68] Lakkaraju et al. (2019) [70] Sokol and Flach (2018) [71] Schneider and Handali (2019) [72]

selection bias here. Furthermore, the low number of papers using logical expressions was surprising since logic rules are inherently expressive [87]. We found that many neuro-symbolic approaches are not informed in a sense that they first use an existing model and then try to improve it using additional knowledge. Instead, they aim to provide a concept that combines logic and connectionism (for an overview of that field we refer to [88]).

The key strength of integrating prior knowledge for explainable ML is its ability to increase accessibility of explanations. Knowledge that is already available can be utilized to give context and to address user needs. In addition, explanations can be used to inform the learning system.

Recall that some reasons for why we need explainable machine learning are verification of the system, compliance to legislation, improvement of the system, and learning from the system [20]. In the first two cases, namely verification of the system and compliance to legislation, an auditor assesses the behavior of the system. The auditor has to make an informed decision whether the system complies with all requirements without necessarily being an expert on the underlying technology. Prior knowledge can serve as a bridge to equip the auditor with the required context to probe the system in a meaningful manner and come to a sound verdict. Regarding the improvement of ML systems, current explanation methods already help developers to detect flaws like the Clever Hans phenomenon. Given the present-day need for explanation, task performance is not the only objective that needs to be considered when designing a new system. Although secondary objectives, like fairness or transparency, are elusive concepts [89], integrating prior knowledge could be a way to approach this problem (c.f. [66]). Consider this: We do not train our models with fairness or explainability in mind, yet we fault them for not demonstrating these traits inherently. In order to ensure fairness and explainability, we need to clearly capture these in the learning process. If there exists knowledge about, e.g., undesired racial bias in a given model, we can use that information for explicit regularization such that the model treats different groups equally [81].

A prerequisite for informed machine learning is the availability of prior knowledge. While not all domains have easy access to prior knowledge, a knowledge source that is not necessarily domain-specific may still be eligible for explainability. It is important to create awareness of existing knowledge bases and also to consider the different ways a knowledge source can be used. For the latter point, our work illustrates a variety of integration opportunities.

Further research is needed to capture the different effects of incorporating prior knowledge on explainability.

The absence of methods that integrate differential equations and simulations offers another starting point for future work. Both representation types are close to the field of physics, where informed machine learning is well-established [90], but the effect on explainability is less investigated. A possible direction is given by Bikmukhametov and Jäschke [91] who incorporate first principles models and investigate their effect on the explainability.

b) Informed Machine Learning to Increase Explainability: In Subsection III-A, prior knowledge is explicitly integrated into the machine learning pipeline to also improve model explainability. While we have reviewed research in Subsection III-A that is exemplary for each integration type, we can see that most approaches use the learning algorithm stage to integrate additional information. Possibly, because the integration as regularization enables using existing model architectures and is less labor-intensive.

The benefit of this approach is that it leverages existing knowledge and makes machine learning models more comprehensible by integrating the knowledge through an interpretable component into the machine learning pipeline. Choi et al. [45] demonstrate a correlation between introducing the prior knowledge source and achieving more concise visualizations of embeddings as compared to competitors.

Again, the way in which the explanation quality is measured is often not addressed or evaluated. From the papers that evaluated explainability in some regard there was no clear consensus on what type of measure is preferable. We did not find a common understanding of the degree to which a method improves interpretability or explainability. In this sense, it is not straightforward to determine whether certain knowledge representations or integration types are especially effective. Reasons for this are the dependency on the application context and the lack of formalism of interpretability.

There are many papers that can be categorized into IML which do not necessarily state it as an explicit goal to make their models more explainable. Explainability should not be treated as an afterthought but has to already be considered in the design phase.

c) Informed Explainability Methods: In Subsection III-B, prior knowledge is incorporated into the explainability method. We expanded on the informed machine learning taxonomy by introducing the explainability method as a new knowledge integration stage. A method qualifies for this category if it integrates an independent knowledge source in addition to an existing algorithm which provides explanations.

A distinction can be made between interactive approaches and formalized priors. The benefit of the interactive approach is that the user can give direct feedback on an explanation. This case can be seen as a communication module between the system and the user. For the formalized priors, knowledge is integrated once to improve the explainability component directly. In both cases, incorporating additional knowledge into the explainability method allows for the accommodation of user needs.

We emphasize that post-hoc methods should be considered with caution. Since explanations are obtained by approximations, it cannot be ensured that the explanations are faithful to the model. The integration of a comprehensible knowledge source does not change that and should not lead to a false sense of security. Especially in high-stake scenarios, inherently interpretable models should be preferred [21].

d) Deriving Knowledge from Explainable Results: In Subsection III-C, knowledge is derived from explainability methods and subsequently integrated into the ML pipeline.

The common way to improve a model is an iterative trial-and-error approach, e.g., feature engineering, in which the knowledge that is gained is rather implicit. In contrast, the reviewed methods make explicit use of explainability methods to generate insights that are formalized and then used to inform the model in the next iteration. This means that explainability is a precursor for informed machine learning. Consequently, this could give rise to an improved methodology where the formalization enables a more explicit way to elaborate on the reasoning for certain design choices. This can help to decouple seemingly arbitrary decisions made in a specific context to generate insights on a broader scale.

We found that very few methods incorporate the gained knowledge from the explanations into, for example, the model architecture, suggesting an opportunity for future work. For future research in this domain, we refer to recent work [92] which highlights data sets for explainability research in Natural Language Processing.

e) Open research directions: To conclude this section we summarize the general directions for future work.

Knowledge formalization: Transforming intuitive knowledge in the form of human feedback and fuzzy expert knowledge to formal knowledge representations. In Subsection III-A we saw a large body of work using prior knowledge to align parts of the ML pipeline with human priors. General knowledge is used to transform and enrich training data. Domain specific knowledge is used to constrain the hypothesis set and adapt the learning algorithm. These techniques require prior knowledge to be formalized, e.g. as mathematical criteria or knowledge graphs. Subsection III-C shows another possible way with the encoding of expert knowledge in annotation masks. Analogous to research dedicated to produce high quality data sets, research with the explicit goal to produce domain and task specific knowledge resources is required [93]. Subsequent indexing and systematization of these resources will be needed to make them accessible and thus encourage their use during the development of new and more interpretable systems.

Making implicit development processes explicit through formalization of applied priors to facilitate development and adaptability of future research. This is closely related to the prior point but we want to highlight its relevance to the ML community. Parameter tuning and model choice, especially in deep learning, often is an informal and under-reported process. Approaches discussed in Subsection III-C, where researchers and developers use insights obtained from explanations to improve or debug a model, can be seen as an informed search. This could be a way to give more substantial justification for model and parameter choices than to just report an improvement in accuracy.

Developing Informed Explainability Methods: This integration stage is of special interest in current times, where opaque models are the standard. Not only can it be applied to existing models but the research body of post-hoc explainability already offers a versatile set of tools. As discussed before, extra care has to be taken to ensure faithfulness of post-hoc explanations. We think that for Informed Explainability Methods this problem is exacerbated. The additional information should increase accessibility of the explanation but should not be used deceptively to make the model appear more plausible. To ensure that accessibility and plausibility do not come at

the cost of decreased faithfulness, standardized evaluation procedures are mandatory that need yet to be established.
Going beyond the notion of post-hoc methods, Informed Explainability Methods can be used to describe methods to personalize explanations to individual user needs. This could mean to provide a user with certain additional information they need in order to achieve comprehension. That would require approaches to model user knowledge as well as access to resources that describe how to adapt the present explanation accordingly. Adaptation to user needs could also mean changes to the UI, for example, based on technological literacy of the user or special needs regarding modality. The challenges here again come down to formalization of the necessary knowledge but also extend to research in Human Computer Interaction.

V. Conclusion

Explainability is an essential component to bring machine learning models to the level of being versatile and applicable. Most ML approaches are data-constrained and can only provide explanations stemming from the information in the training data. Hence, we propose to harness prior knowledge, such as logical rules or knowledge graphs, with the goal of improving explainability. In this paper, we presented three approaches to integrate prior knowledge into the ML pipeline and into the explainability component. The three approaches can be distinguished between the integration method and how prior knowledge is obtained: Prior knowledge is available independently of data or pipeline and is (1) incorporated into the ML pipeline or (2) into the explainability method. Prior knowledge can also be (3) derived from model explanations, formalized and then incorporated into the pipeline. With this, we have created a structure that serves for orientation. Further research is needed to formalize and measure to what extent knowledge integration improves explainability.

Acknowledgment

We thank Stefan Rüping, Linara Adilova, Dario Antweiler, Ann-Kathrin Oster and Katharina Peters for valuable feedback.

References

[1] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016," *Official Journal of the European Union*, 2016.

[2] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[3] X.-H. Li *et al.*, "A survey of data-driven and knowledge-aware explainable AI," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[4] A. B. Arrieta *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[5] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[6] C. Molnar, *Interpretable Machine Learning*, 2nd ed., 2022. [Online]. Available: christophm.github.io/interpretable-ml-book/

[7] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *WIREs Data Mining and Knowledge Discovery*, vol. 11, no. 1, 2021.

[8] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[9] G. Ras, M. van Gerven, and P. Haselager, "Explanation methods in deep learning: Users, values, concerns and challenges," in *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 19–36.

[10] M. Norkute, "AI explainability: Why one explanation cannot fit all," 2021, presented at ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI (HCXAI), unpublished.

[11] S. Chari, D. Gruen, O. Seneviratne, and D. McGuinness, "Directions for explainable knowledge-enabled systems," in *Knowledge Graphs for eXplainable Artificial Intelligence*. IOS Press, 2020, vol. 47, pp. 245–261.

[12] D. Doran, S. Schulz, and T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*, vol. 2071. CEUR, 2017.

[13] C. Molnar *et al.*, "General pitfalls of model-agnostic interpretation methods for machine learning models," in *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 2022, pp. 39–68.

[14] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42 200–42 216, 2020.

[15] F. Lauer and G. Bloch, "Incorporating prior knowledge in support vector machines for classification: A review," *Neurocomputing*, vol. 71, no. 7, pp. 1578–1594, 2008.

[16] A. Karpatne *et al.*, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 10, pp. 2318–2331, 2017.

[17] L. von Rueden *et al.*, "Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning system," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[18] I. Tiddi and S. Schlobach, "Knowledge graphs as tools for explainable machine learning: A survey," *Artificial Intelligence*, vol. 302, p. 103627, 2022.

[19] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, p. 31–57,

2018.

[20] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," Preprint at https://arxiv.org/abs/1708.08296, 2017.

[21] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[22] B. Kim *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)," in *Proceedings of the 35th International Conference on Machine Learning*, vol. 6. PMLR, 2018, pp. 4186–4195.

[23] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia: PMLR, 2017, pp. 5109–5118.

[24] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates, 2017, p. 4768–4777.

[25] A. Binder, G. Montavon, S. Lapuschkin, K.-R. Müller, and W. Samek, "Layer-wise relevance propagation for neural networks with local renormalization layers," in *International Conference on Artificial Neural Networks*, ser. Lecture Notes in Computer Science. Springer, 2016, pp. 63–71.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. San Francisco, California, USA: ACM, 2016, pp. 1135–1144.

[27] L. S. Shapley, *A Value for n-Person Games*. Princeton University Press, 1953, pp. 307–317.

[28] R. J. Aumann and L. S. Shapley, *Values of Non-Atomic Games*. Princeton University Press, 1974.

[29] W. Stammer, P. Schramowski, and K. Kersting, "Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3619–3629.

[30] N. Frosst and G. E. Hinton, "Distilling a neural network into a soft decision tree," in *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML*, 2017.

[31] S. Tan, R. Caruana, G. Hooker, P. Koch, and A. Gordo, "Learning global additive explanations for neural nets using model distillation," in *Machine Learning for Health (ML4H) Workshop at NeurIPS*, 2018.

[32] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2017.

[33] S. Dandl, C. Molnar, M. Binder, and B. Bischl, "Multi-objective counterfactual explanations," in *Parallel Problem Solving from Nature – PPSN XVI*. Springer, 2020, pp. 448–469.

[34] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: From counterfactual explanations to interventions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: ACM, 2021, p. 353–362.

[35] P. Hase and M. Bansal, "Evaluating explainable AI: Which algorithmic explanations help users predict model behavior?" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 2020, pp. 5540–5552.

[36] Z. Buçinca, P. Lin, K. Z. Gajos, and E. L. Glassman, "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proceedings of the 25th International Conference on Intelligent User Interfaces*, ser. IUI '20. New York, NY, USA: ACM, 2020, p. 454–464.

[37] M. Brundage *et al.*, "Toward trustworthy AI development: Mechanisms for supporting verifiable claims," Preprint at https://arxiv.org/abs/2004.07213, 2020.

[38] X. Wang, D. Wang, C. Xu, X. He, Y. Cao, and T. S. Chua, "Explainable reasoning over knowledge graphs for recommendation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5329–5336, 2019.

[39] W. Ma *et al.*, "Jointly learning explainable rules for recommendation with knowledge graph," in *The World Wide Web Conference*, ser. WWW '19. ACM, 2019, pp. 1210–1221.

[40] T. Ma and A. Zhang, "Incorporating Biological Knowledge with Factor Graph Neural Network for Interpretable Deep Learning," Preprint at https://arxiv.org/abs/1906.00537, 2019.

[41] S. Rybakov, M. Lotfollahi, F. J. Theis, and F. A. Wolf, "Learning interpretable latent autoencoder representations with annotations of feature sets," Preprint at https://www.biorxiv.org/content/10.1101/2020.12.02.401182v1, 2020.

[42] N. Chen, Q. Y. Zhou, and V. K. Prasanna, "Understanding web images by object relation network," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: ACM, 2012, pp. 291–300.

[43] Z. Liu, Z.-Y. Niu, H. Wu, and H. Wang, "Knowledge aware conversation generation with explainable reasoning over augmented graphs," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: ACL, 2019, pp. 1782–1792.

[44] Z. Chen, Y. Bei, and C. Rudin, "Concept whitening for interpretable image recognition," *Nature Machine Intelligence*, vol. 2, no. 12, pp. 772–782, 2020.

[45] E. Choi, M. T. Bahadori, L. Song, W. F. Stewart,

and J. Sun, "GRAM: Graph-based attention model for healthcare representation learning," *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 787–795, 2017.

[46] F. Ma, Q. You, H. Xiao, R. Chitta, J. Zhou, and J. Gao, "KAME: Knowledge-based attention model for diagnosis prediction in healthcare," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, ser. CIKM '18. New York, NY, USA: ACM, 2018, p. 743–752.

[47] J. Jiang, S. Hewner, and V. Chandola, "Tree-based regularization for interpretable readmission prediction," in *Proceedings of the AAAI 2019 Spring Symposium on Combining Machine Learning with Knowledge Engineering AAAI-MAKE*, ser. CEUR Workshop Proceedings, vol. 2350. CEUR, 2019.

[48] K. Yan, Y. Peng, V. Sandfort, M. Bagheri, Z. Lu, and R. M. Summers, "Holistic and comprehensive annotation of clinically significant findings on diverse CT images: Learning from radiology reports and label ontology," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June. IEEE, 2019, pp. 8515–8524.

[49] I. Donadello, L. Serafini, and A. d'Avila Garcez, "Logic tensor networks for semantic image interpretation," in *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2017, pp. 1596–1602.

[50] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[51] Q. Zhang, R. Cao, S. Zhang, M. Redmonds, Y. N. Wu, and S.-C. Zhu, "Interactively transferring CNN patterns for part localization," Preprint at https://arxiv.org/abs/1708.01783, 2017.

[52] T. Pommellet and F. Lécué, "Feeding machine learning with knowledge graphs for explainable object detection," in *Proceedings of the ISWC 2019 Satellite Tracks*, vol. 2456. CEUR, 2019, pp. 277–280.

[53] I. Stepin, J. M. Alonso, A. Catala, and M. Pereira-Fariña, "A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence," *IEEE Access*, vol. 9, pp. 11 974–12 001, 2021.

[54] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 2020, pp. 607–617.

[55] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," Preprint at https://arxiv.org/abs/1912.03277, 2019.

[56] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," in *Deep Learning Workshop, International Conference on Machine Learning*, 2015.

[57] D. Wei, B. Zhou, A. Torrabla, and W. Freeman, "Understanding intra-class knowledge inside CNN," Preprint at https://arxiv.org/abs/1507.02379, 2015.

[58] A. Mahendran and A. Vedaldi, "Visualizing deep convolutional neural networks using natural pre-images," *International Journal of Computer Vision*, vol. 120, no. 3, pp. 233–255, dec 2016.

[59] Z. Shams *et al.*, "REM: An integrative rule extraction methodology for explainable data analysis in healthcare," Preprint at https://www.biorxiv.org/content/10.1101/2021.01.22.427799v1, 2021.

[60] S. Muggleton and L. de Raedt, "Inductive logic programming: Theory and methods," *The Journal of Logic Programming*, vol. 19-20, pp. 629–679, 1994.

[61] J. Rabold, H. Deininger, M. Siebers, and U. Schmid, "Enriching visual with verbal explanations for relational concepts–combining LIME with aleph," in *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*. Springer, 2019, pp. 180–192.

[62] P. Shivaswamy and T. Joachims, "Coactive learning," *Journal of Artificial Intelligence Research*, vol. 53, pp. 1–40, 2015.

[63] P. Kieseberg, J. Schantl, P. Frühwirt, E. Weippl, and A. Holzinger, "Witnesses for the doctor in the loop," in *Brain Informatics and Health*. Springer, 2015, pp. 369–378.

[64] S. Liu, X. Wang, M. Liu, and J. Zhu, "Towards better analysis of machine learning models: A visual analytics perspective," *Visual Informatics*, vol. 1, no. 1, pp. 48–56, 2017.

[65] J. G. S. Paiva, W. R. Schwartz, H. Pedrini, and R. Minghim, "An approach to supporting incremental visual data classification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 21, no. 1, pp. 4–17, 2015.

[66] K. Sokol and P. Flach, "One explanation does not fit all," *KI - Künstliche Intelligenz*, vol. 34, no. 2, pp. 235–250, 2020.

[67] ——, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 56–67.

[68] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020.

[69] J. Krause, A. Perer, and K. Ng, "Interacting with predictions: Visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2016, pp. 5686–5697.

[70] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec, "Faithful and customizable explanations of black box models," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19. New

York, NY, USA: ACM, 2019, p. 131–138.

[71] K. Sokol and P. Flach, "Glass-box: Explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 7 2018, pp. 5868–5870.

[72] J. Schneider and J. Handali, "Personalized explanation in machine learning: A conceptualization," in *Proceedings of the 27th European Conference on Information Systems*, J. vom Brocke, S. Gregor, and O. Müller, Eds., 2019.

[73] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.

[74] T. Baur, A. Heimerl, F. Lingenfelser, J. Wagner, M. F. Valstar, B. Schuller, and E. André, "eXplainable cooperative machine learning with NOVA," *KI - Künstliche Intelligenz*, vol. 34, no. 2, pp. 143–164, 2020.

[75] P. Schramowski *et al.*, "Making deep neural networks right for the right scientific reasons by interacting with their explanations," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 476–486, 2020.

[76] S. Teso and K. Kersting, "Explanatory Interactive Machine Learning," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '19. New York, NY, USA: ACM, 2019, pp. 239–245.

[77] T. Kulesza, M. Burnett, W. K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the International Conference on Intelligent User Interfaces*. ACM, 2015, pp. 126–137.

[78] L. Adilova, S. Giesselbach, and S. Rüping, "Making efficient use of a domain experts time in relation extraction," in *Proceedings of the Workshop on Interactions between Data Mining and Natural Language Processing, co-located with ECML-PKDD*, vol. 1880. CEUR, 2017, pp. 1–16.

[79] C. dos Santos, B. Xiang, and B. Zhou, "Classifying relations by ranking with convolutional neural networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Beijing, China: ACL, 2015, pp. 626–634.

[80] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *International Joint Conference on Artificial Intelligence*, 2017.

[81] L. Rieger, C. Singh, W. Murdoch, and B. Yu, "Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 2020, pp. 8116–8126.

[82] R. R. Selvaraju *et al.*, "Taking a HINT: Leveraging explanations to make vision and language models more grounded," in *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2019, pp. 2591–2600.

[83] G. Erion, J. D. Janizek, P. Sturmfels, S. Lundberg, and S.-I. Lee, "Improving performance of deep learning models with axiomatic attribution priors and expected gradients," *Nature Machine Intelligence*, vol. 3, pp. 620–631, 2021.

[84] V. Balayan, P. Saleiro, C. Belém, L. Krippahl, and P. Bizarro, "Teaching the machine to explain itself using domain knowledge," in *HAMLETS Workshop 2020, NeurIPS*, 2020.

[85] O. M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom, "e-SNLI: Natural language inference with natural language explanations," in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, 2018, pp. 9539–9549.

[86] F. Liu and B. Avci, "Incorporating priors with feature attribution on text classification," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: ACL, Jul. 2019, pp. 6274–6283.

[87] S. Muggleton, U. Schmid, C. Zeller, A. Tamaddoni-Nezhad, and T. Besold, "Ultra-strong machine learning: comprehensibility of programs learned with ILP," *Machine Learning*, vol. 107, no. 7, pp. 1119–1140, 2018.

[88] Z. Bouraoui *et al.*, "From shallow to deep interactions between knowledge representation, reasoning and machine learning," Preprint at https://arxiv.org/abs/1912.06612, Dec 2019.

[89] M. Krishnan, "Against interpretability: A critical examination of the interpretability problem in machine learning," *Philosophy and Technology*, vol. 33, no. 3, pp. 487–502, 2020.

[90] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[91] T. Bikmukhametov and J. Jäschke, "Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models," *Computers & Chemical Engineering*, vol. 138, p. 106834, 2020.

[92] S. Wiegreffe and A. Marasović, "Teach me to explain: A review of datasets for explainable natural language processing," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds., vol. 1, 2021.

[93] M. Jakobs, H. Kotthaus, I. Röder, and M. Baritz, "SancScreen: Towards a real-world dataset for evaluating explainability methods," *LWDA*, 2022, in press.