LEVERAGING EXPLANATIONS IN INTERACTIVE MACHINE LEARNING: AN OVERVIEW

Stefano Teso
University of Trento
Trento, Italy
stefano.teso@unitn.it

Öznur Alkan
Optum
Dublin, Ireland
oznur.alkan@optum.com

Wolfang Stammer
Technical University Darmstadt
Darmstadt, Germany
wolfgang.stammer@cs.tu-darmstadt.de

Elizabeth Daly
IBM Research
Dublin, Ireland
elizabeth.daly@ie.ibm.com

August 1, 2022

ABSTRACT

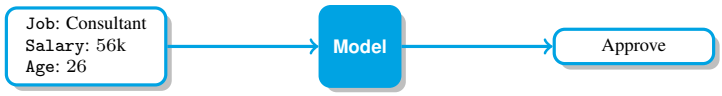
Explanations have gained an increasing level of interest in the AI and Machine Learning (ML) communities in order to improve model transparency and allow users to form a mental model of a trained ML model. However, explanations can go beyond this one way communication as a mechanism to elicit user control, because once users understand, they can then provide feedback. The goal of this paper is to present an overview of research where explanations are combined with interactive capabilities as a mean to learn new models from scratch and to edit and debug existing ones. To this end, we draw a conceptual map of the state-of-the-art, grouping relevant approaches based on their intended purpose and on how they structure the interaction, highlighting similarities and differences between them. We also discuss open research issues and outline possible directions forward, with the hope of spurring further research on this blooming research topic.

Keywords Human-in-the-Loop · Explainable AI · Interactive Machine Learning · Debugging · Model Editing

1 Introduction

The fields of eXplainable Artificial Intelligence (XAI) and Interactive Machine Learning (IML) have traditionally been explored separately. On the one hand, XAI aims at making AI and Machine Learning (ML) systems more transparent and understandable, chiefly by equipping them with algorithms for explaining their own decisions [66, 125]. Such explanations are instrumental for enabling stakeholders to inspect the system's knowledge and reasoning patterns, however stakeholders only participate as *passive observers* and have no control over the system or its behavior. On the other hand, IML focuses primarily on communication between machines and humans, and it is specifically concerned with eliciting and incorporating human feedback into the training process via intelligent user interfaces [53, 10, 109, 176, 71, 173]. IML covers a broad range of techniques for in-the-loop interaction between humans and machines, however, most research *does not explicitly consider explanations*.

Recently, a number of works have sought integrating techniques from XAI within the IML loop. The core observation behind this line of research is that, *interacting through explanations* is an elegant and human-centric solution to the problem of acquiring rich human feedback, and therefore leads to higher-quality AI and ML systems, in a manner that is effective and transparent for both users and machines. In order to accomplish this vision, these works leverage either *machine explanations* obtained using techniques from XAI, *human explanations* provided as feedback by sufficiently expert annotators, or both, to define and implement a suitable interaction protocol.



Input Attribution	Example Attribution	Rule	Concept Attribution
Importance of Job: +0.2 Salary: +0.6 Age: −0.3	Similarity to Job: Consultant Salary: 53k Age: 30 Approve	Approve if (Salary > 50k AND Age > 24) OR (Age > 45 AND Job = Consultant)	Importance of High Earner: +0.6 Professional: +0.9 Sporty: −0.2 ...

Figure 1: Illustration of how various kinds of explanations support human understanding in the context of loan requests. From left to right: input attributions and example attributions (discussed in Section 3), rules (discussed in Section 4), and concept attributions (discussed in Section 5).

These two types of explanations play different roles. By observing the machine's explanations, users have the opportunity of building a better *understanding* of the machine's overall logic, which not only facilitates trust calibration [10], but also supports and amplifies our natural capacity of providing appropriate feedback [92]. Machine explanations are also key for identifying imperfections and bugs affecting ML models, such as reliance on confounded features that are not causally related with the desired outcome [97, 60, 137]. At the same time, human explanations are a very rich source of supervision for models [31] and are also very natural for us to provide. In fact, explanations tap directly into our innate learning and teaching abilities [103, 106] and are often spontaneously provided by human annotators when given the chance [157]. Machine explanations and human feedback can also be combined to build interactive *model editing* and *debugging* facilities, because – once aware of limitations or bugs in the model – users can indicate effective improvements and supply corrective feedback [92, 160].

The goal of this paper is to provide a general overview and perspective of research on leveraging explanations in IML. Our contribution is two-fold. First, we survey a number of relevant approaches, grouping them based on their intended purpose and how they structure the interaction. Our second contribution is a discussion of open research issues, on both the algorithmic and human sides of the problem, and possible directions forward. This paper is not meant as an exhaustive commentary of all related work on explainability or on interactivity. Rather, we aim to offer a conceptual guide for practitioners to understand core principles and key approaches in this flourishing research area. In contrast to existing overviews on using explanations to guide learning algorithms [69], debug ML models [98], and transparently recommend products [194], we specifically focus on *human-in-the-loop* scenarios and consider a broader range of applications, highlighting the variety of mechanisms and protocols for producing and eliciting explanations in IML.

Outline: This paper is structured as follows. In Section 2 we discuss a general recipe for integrating explanations into interactive ML, recall core principles for designing successful interaction strategies, and introduce a classification of existing approaches. Then we discuss key approaches in more detail, organizing them based on their intended purpose and the kind of machine explanations they leverage. Specifically, we survey methods for debugging ML models using saliency maps and other local explanations in Section 3, for editing models using global explanations (e.g., rules) in Section 4, and for learning and debugging ML models with concept-level explanations in Section 5. Finally, we outline remaining open problems in Section 6 and related topics in Section 7.

2 Explanations in Interactive Machine Learning

Interactive Machine Learning (IML) stands for the design and the implementation of algorithms and user interfaces in order to engage users to actually build ML models. This stands in contrast to standard procedures, in which building a model is a fully automated process and domain experts have little control beyond data preparation [176]. Research in IML explores ways to learn and manipulate models through an intuitive human-computer interface [109] and encompasses a variety of learning and interaction strategies. Perhaps the most well-known IML framework is active learning [141, 74], which tackles learning high-performance predictors in settings in which supervision is expensive. To this end, active learning algorithms interleave acquiring labels of carefully selected unlabeled instances from an annotator and model updates. As another brief example, consider a recommender solution in a domain like movies, videos, or music, where the user can provide explicit feedback through rating the recommended items [71]. These ratings can then be infused to the recommendation model's decision making process so as to tailor the recommendations towards end users interests.

In IML, users are encouraged to shape the decision making process of the model, so it is important for users to build a correct mental model of the "intelligent agent", which will then allow them to seamlessly interact with it. Conversely, since user feedback and involvement are so central, uninformed feedback risks wasting annotation effort and ultimately compromising model quality. Not all approaches to IML are equally concerned with user understanding. For instance, in active learning the machine is essentially a black-box, with no information being disclosed about what knowledge it has acquired and what effect feedback has on it [160]. Strategies for interactive customization of ML models, like the one proposed by Fails and Olsen Jr [53], are less opaque, in that users can explore the impact of their changes and tune their feedback accordingly. Yet, the model's logic can only be (approximately) reconstructed from changes in behavior, making it hard to *anticipate* what information should be provided to guide the model in a desirable direction [92]. Following Kulesza et al. [92], we argue that proper interaction requires transparency and an understanding of the underlying model's logic. And it is exactly here that *explanations* can be used to facilitate this process.

2.1 A General Approach for Leveraging Explanations in Interaction

In order to appreciate the potential roles played by explanations, it is instructive to look at *explanatory debugging* [91, 92], the first framework to explicitly leverage them in IML, and specifically at EluciDebug [92], its proof-of-concept implementation.

EluciDebug is tailored for interactive customization of Naïve Bayes classifiers in the context of email categorization. To this end, it presents users with explanations that illustrate the relative contributions of the model's prior and likelihood towards the class probabilities. In particular, its explanations convey what *words* the model uses to distinguish work emails from personal emails and how much they impact its decisions. In a second step, the user has the option of increasing or decreasing the relevance of certain input variables toward the choice of certain classes by directly adjusting the model weights. Continuing with our email example, the user is free to specify relevant words that the system is currently ignoring and irrelevant words that the system is wrongly relying on. This very precise form of feedback contrasts with traditional label-based strategies, in which the user might, e.g., tell the system that a message from her colleague about baseball is a personal (rather than work-related) communication, but has no direct control over what inputs the system decides to use. The responsibility of choosing what examples (e.g., wrongly classified emails) to provide feedback on is left to the user, and the interaction continues until she is happy with the system's behavior. EluciDebug was shown to help users to better understand the system's logic and to more quickly tailor it toward their needs [92].

EluciDebug highlights how explanations contribute to both *understanding* and *control*, two key elements that will reoccur in all approaches we survey. We briefly unpack them in the following.

Understanding. By observing the machine's explanations, users get the opportunity of building a better *understanding* of the machine's overall logic. This is instrumental in uncovering limitations and flaws in the model [167]. As a brief example, consider a recommender solution in a domain like movies, videos, or music, where the users are presented with explanations in the form of a list of features that are found to be most relevant to the users' previous choices [162]. Upon observing this information, users can see the assumptions the underlying recommender has made for their interests and preferences. It might be the case that the model made incorrect assumptions for the users' preferences possibly due to some changes of interests which is not explicitly available in the data. In such a scenario, explanations provide a perfect ground for understanding the underlying model's behavior.

Explanations are also instrumental for identifying models that rely on confounds in the training data, such as watermarks in images, that happen to correlate with the desired outcome but that are not causal for it [97, 60]. Despite achieving high accuracy during training, these models generalize poorly to real-world data where the confound is absent. Such buggy behavior can affect high-stakes applications like COVID-19 diagnosis [46] and scientific analysis [137], and cannot be easily identified using standard evaluation procedures without explanations. Ideally, the users would develop a structural mental model that gives them a deep understanding of how the model operates, however a functional understanding is often enough for them to be able to interact [152]. The ability of disclosing issues with the model, in turn, facilitates trust calibration [10]. This is especially true in interactive settings as here the user can witness how the model evolves over time, another important factor that contributes to trust [177, 175].

Control. Understanding supports and amplifies our natural capacity of providing appropriate feedback [92]: once bugs and limitations are identified, interaction with the model enables end-users to modify the algorithm in order to correct those flaws. Bi-directional communication between users and machines together with transparency enables *directability*, that is, the ability to rapidly assert control or influence when things go astray [75]. Clearly, control is fundamental if one plans to take actions based on a model's prediction, or to deploy a new model in the first place [127].

Goal	Explanations	Feedback	Incorporation	Method
Learning	Local, CA	Adjust Feature Association	Update model w/ auxiliary loss	Lage and Doshi-Velez [94]
		Adjust Encodings	Update model w/ auxiliary loss	Stammer et al. [155]
Debugging	Local, IA	Adjust Parameters	Update model w/ improved parameters	EluciDebug [92]
		Adjust Attributions	Update data	CAIPI [160]
			Update model w/ auxiliary loss	RRR [137, 130]
			Update model w/ auxiliary loss	Teso [159]
	Local, EA	Additional Features	Update model w/ additional classifiers	ALICE [100]
		Adjust Attributes	Update data	Biswas and Parikh [25]
		Example Similarity	Update data	HILDIF [195]
		Counter Examples	Update data	CINCER [161]
	Local, CA	Adjust Attributions	Update model w/ auxiliary loss	RRC [154]
			Update model w/ auxiliary loss	Bontempelli et al. [27]
			Update model w/ auxiliary loss	ProtoPDebug [28]
		Sample Pairing	Update model w/ auxiliary loss	Shao et al. [144]
	Global, Rules	Adjust Attributions	Update model w/ hard constraint	FIND [99]
			Update model w/ auxiliary loss	REMOTE [185]
		Counter Examples	Update data	XGL [122]
Editing	Global, Rules	Rule Editing	Update data	FROTE [8]
			Post-processing	Overlay [41]
			Post-processing	XIML [68]
		Adjust Feature Association	Update model w/ auxiliary loss	Antognini et al. [13]
			Post-processing	Alkan et al. [6, 7]

Table 1: Table of methods covered in this overview. We here differentiate the various methods in their algorithmic goals (Goal), the type of explanations they consume (Explanations), the type of feedback provided by the user (Feedback), and, finally, the strategy of incorporating the explanatory user feedback (Incorporation). *Abbreviations*: CA = concept attribution, EA = example attribution, IA = input attribution.

At the same time, directability also contributes to trust allocation [75]. The increased level of control can also help to achieve significant gains in the end user's satisfaction.

Human feedback can come in many forms, and one of these forms is explanations, either from scratch or by using the machine's explanations as a starting point [160]. This type of supervision is very informative: a handful of explanations are oftentimes worth many labels, substantially reducing the sample complexity of *learning* (well-behaved) models [31]. Importantly, it is also very natural for human to provide as explanations lie at the heart of human communication and tap directly into our innate learning and teaching abilities [103, 106]. In fact, Stumpf et al. [157] showed that when given the chance to provide free-form annotations, users had no difficulty providing generous amounts of feedback.

Principles. To ground the exchange of explanations between an end user and a model, Kulesza et al. [91] presented a set of key principles around *explainability* and *correctability*. Although the principles are discussed in the context of explanatory debugging, they apply to all the approaches that are presented in this paper. These include: (1) Being iterative, so as to enable end-users to build a reasonable and informed model of the machine's behavior. (2) Presenting sound, faithful explanations that do not over-simplify the model's reasoning process. (3) Providing as complete a picture of the model as possible, without omitting elements that play an important role in its decision process. (4) Avoiding to overwhelming the user, as this complicates understanding and feedback construction. (5) Ensuring that explanations are actionable, making them engaging for users to attend to, thus encouraging understanding while enabling users to adjust them to their expertise. (6) Making user changes easily reversible. (7) Always honoring feedback, because when feedback is disregarded users may stop bothering to interact with the system. (8) Making sure to effectively communicate what effects feedback has on the model. Clearly there is a tension behind these principles, but they nonetheless are useful in guiding the design of explanation-based interaction protocols. As we will discuss in Section 6, although existing approaches attempt to satisfy one or more of these desiderata, no *general* method yet exists that satisfies all of them.

2.2 Dimensions of Explanations in Interactive Machine Learning

Identifying *interaction* and *explainability* as two key capabilities of a well performing *and* trust-worthy ML system, motivates us to layout this overview on leveraging explanations in interactive ML. The methods we survey tackle different applications using a wide variety of strategies. In order to identify common themes and highlight differences, we organize them along four dimensions:

Algorithmic goal: We identify three high-level scenarios. One is that of using explanation-based feedback, optionally accompanied by other forms of supervision, to *learn* an ML model from scratch. Here, the machine is typically in charge of asking appropriate questions, feedback may be imperfect, and the model is updated incrementally as feedback is received. Another scenario is model *editing*, in which domain experts are in charge of inspecting the internals of a (partially) trained model (either directly if the model is white-box or indirectly through its explanations) and can manipulate them to improve and expand the model. Here feedback is typically assumed high-quality and used to constrain the model's behavior. The last scenario is *debugging*, where the focus is on fixing issues of misbehaving (typically black-box) models and the machine's explanations are used to both spot bugs and elicit corrective feedback. Naturally, there is some overlap between goals. Still, we opt to keep them separate as they are often tackled using different algorithmic and interaction strategies.

Type of machine explanations: The approaches we survey integrate four kinds of machine explanations: *input attributions* (IAs), *example attributions* (EAs), *rules*, and *concept attributions* (CAs). IAs identify those input variables that are responsible for a given decision, and as such they are *local* in nature. EAs and CAs are also local, but justify decisions in terms of relevant training examples and high-level concepts, respectively. At the other end of the spectrum, rules are *global* explanations in that they aim to summarize, in an interpretable manner, the logic of a whole model. These four types of explanations are illustrated in Fig. 1 and described in more detail in the next sections

Type of human feedback and incorporation strategy: Algorithmic goal and choice of machine explanations act as a determiner for the types of interactions that can happen, in turn affecting two other important dimensions, namely the *type of feedback* that can be collected and the way the machine can *consume this feedback* [116]. Feedback ranges from updated parameter values, as in EluciDebug, to additional data points, to gold standard explanations supplied by domain experts. Incorporation strategies go hand-in-hand, and range from updating the model's parameters as instructed to (incrementally) retraining the model, perhaps including additional loss terms to incorporate explanatory feedback. All details are given in the following sections.

The methods we survey are listed in Table 1. Notice that the two most critical dimensions, namely algorithmic goal and type of machine explanations, are tightly correlated: learning approaches tend to rely on local explanations and editing approaches on rules, while debugging approaches employ both. For this reason, we chose to structure the next three sections by explanation type. One final remark before proceeding. Some of the approaches we cover rely on choosing specific instances or examples to be presented to the annotator. Among them, some rely on *machine-initiated* interaction, in the sense that they leave this choice to the machine (for instance, methods grounded on active learning tend to pick specific instances that the model is uncertain about [141]), while others rely on *human-initiated* interaction and expect the user to pick instances of interest from a (larger) set of options. We do not group approaches based on this distinction, so as to keep our categorization manageable. The specific type of interaction used will be made clear in the following sections on a per-method basis.

3 Interacting via Local Explanations

In this section, we discuss IML approaches that rely on *local explanations* to carry out interactive model debugging. Despite sharing some aspects with EluciDebug [92], these approaches exploit modern XAI techniques to support state-of-the-art ML models and explore alternative interaction protocols. Before reviewing them, we briefly summarize those types of local explanations that they build on.

3.1 Input Attributions and Example Attributions

Given a classifier and a target decision, local explanations identify a subset of "explanatory variables" that are most responsible for the observed outcome. Different types of local explanations differ in what variables they consider and in how they define responsibility.

Input attributions, also known as saliency maps, convey information about relevant vs. irrelevant input variables. For instance, in loan request approval an input attribution might report the relative importance of variables like `Job`, `Salary` and `Age` of the applicant, as illustrated in Fig. 1, and in image tagging that of subsets of pixels, as shown in Fig. 2 (left). A variety of attribution algorithms have been developed. Gradient-based approaches like Input Gradients (IGs) [16, 146],

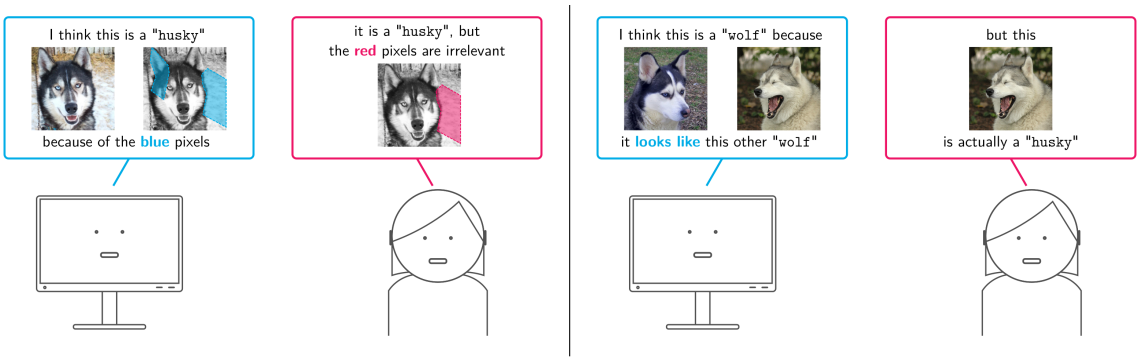


Figure 2: Illustration of two explanation strategies based on local explanations. **Left**: The machine explains its predictions by highlighting relevant *input variables* – in this case, relevant pixels – and the user replies with an improved attribution map. **Right**: The machine justifies its predictions in terms of training examples that support them, and the user either corrects the associated label. In both cases, the data and model are aligned to the user's feedback.

GradCAM [138], and Integrated Gradients [158] construct a saliency map by measuring how sensitive the score or probability assigned by the classifer to the decision is to perturbations of individual input variables. This information is read off from the gradient of the model's output with respect to its input, and as such it is only meaningful if the model is differentiable and the inputs are continuous (e.g., images). Sampling-based approaches like LIME [127] and SHAP [156, 104] capture similar information [58], but they rely on sampling techniques that are applicable also to non-differentiable models and to categorical and tabular data. For instance, LIME distills an interpretable surrogate model using random samples labeled by the black-box predictor and then extracts input relevance information from the surrogate. Despite their wide applicability, these approaches tend to be more computationally expensive than gradient-based alternatives [48] and, due to variance inherent in the sampling step, in some cases their explanations may portray an imprecise picture of the model's decision making process [193, 159]. One last group of approaches focus on identifying the smallest subset of input variables whose value, once fixed, ensures that the model's output remains the same regardless of the value taken by the remaining variables [145, 172, 32], and typically come with faithfulness guarantees. Although principled, these approaches however have not yet been used in explanatory interaction.

Example attributions, on the other hand, explain a target decisions in terms of those training examples that most contributed to it, and are especially natural in settings, like medical diagnosis, in which domain experts are trained to perform case-based reasoning [24, 86, 36]. For instance, in Fig. 1 a loan application is approved by the machine because it is similar to an previously approved application and in Fig. 2 (right) a mislabeled training image fools the model into mispredicting a `husky` dog as a `wolf`. For same classes of models, like nearest neighbor classifiers and prototype-based predictors, example relevance can be easily obtained. For all other models, it can in principle be evaluated by removing the example under examination from the training set and checking how this changes the models' prediction upon retraining. This naïve solution however scales poorly, especially for larger models for which retraining is time consuming. A more convenient alternative are Influence Functions (IFs), which offer an efficient strategy for approximating example relevance without retraining [88], yielding a substantial speed-up. Evaluating IFs is however non-trivial, as it involves inverting the Hessian of the model, and can be sensitive to factors such as model complexity [21] and noise [161]. This has prompted researchers to develop more scalable and robust algorithms for IFs [67] as well as alternative strategies to approximate example relevance [187, 85].

3.2 Interacting via Input Attributions

One line of work on integrating input attributions and interaction is *eXplanatory Interactive Learning* (XIL) [160, 137]. In the simplest case, XIL follows the standard active learning loop [141], except that whenever the machine queries the label of a query instance, it also presents a *prediction* for that instance and a *local explanation* for the prediction. Contrary to EluciDebug, which is designed for interpretable classifiers, in XIL the model is usually a black-box, e.g., a support vector machine or a deep neural network, and explanations are extracted using input attribution methods like LIME [160] or GradCAM [137]. At this point, the annotator supplies a label for the query instance – as in regular active learning – and, optionally, corrective feedback on the explanation. The user can, for instance, indicate what input variables the machine is wrongly relying on for making its prediction. Consider Fig. 2 (left): here the query instance depicts a `husky` dog, but the model wrongly classifies it as a `wolf` based on the presence of snow in the background (in **blue**), which happens to correlate with wolf images in the training data. To correct for this, the user indicates that the snow should not be used for prediction (in **red**).

XIL then aligns the model based on this corrective feedback. CAIPI [160], the original implementation of XIL, achieves this using data augmentation. Specifically, CAIPI makes a few copies of the target instance (e.g., the husky image in Fig. 2) and then *randomizes* the input variables indicated as irrelevant by the user while leaving the label unchanged, yielding a small set of synthetic examples. These are then added to the training set and the model is retrained. Essentially, this teaches the model to classify the image correctly *without relying on the randomized variables*. Data augmentation proved effective at debugging shallow models for text classification and other tasks [160, 150] and at reducing labeling effort [150], at the cost of requiring extra space to store the synthetic examples.

A more refined version of CAIPI [137] solves this issue by introducing two improvements: LIME is replaced with GradCAM [138], thus avoiding sampling altogether, and the model is aligned using a generalization of the *right for the right reasons* (RRR) [130] modified to work with GradCAM. Essentially, the RRR loss penalizes the model proportionally to the relevance that its explanations assign to inputs that have been indicated as irrelevant by the user. Combining it with a regular loss for classification (e.g., the categorical cross-entropy loss) yields an end-to-end differentiable training pipeline that can be optimized using regular back-propagation. This in turn leads to shorter training times, especially for larger models, without the need for extra storage space. This approach was empirically shown to successfully avoid Clever Hans behavior in deep neural networks used for hyperspectral analysis of plant phenotyping data [137].

Several alternatives to the RRR loss have been developed. The Contextual Decomposition Explanation Penalization (CDEP) [129] follows the same recipe, but it builds on Contextual Decomposition [147], an attribution technique that takes relationships between input variables into account. The approach of Yao et al. [185] enables annotators to dynamically explore the space of feature interactions and fine-tune the model accordingly. The Right for Better Reasons (RBR) loss [143] improves on gradient-based attributions by replacing input gradients with their influence function [88], and it was shown to be more robust to changes in the model and speed up convergence. Human Importance-aware Network Tuning (HINT) [139] takes a different route in that it rewards the model for activating on inputs deemed *relevant* by the annotator. Finally, Teso [159] introduced a ranking loss designed to work with partial and possibly sub-optimal explanation corrections. These methods have recently been compared in the context of XIL by Friedrich et al. [56]. There, the authors introduce a set of benchmarking metrics and tasks and conclude that the "no free lunch" theorem [178] also holds for XIL, i.e., no method exceeds on all evaluations.

ALICE [100] also augments active learning, but it relies on contrastive explanations. In each interaction round, the machine selects a handful of class *pairs*, focusing on classes that the model cannot discriminate well, and for each of them asks an annotator to provide a natural language description of what features enable them to distinguish between the two classes. It then uses semantic parsing to convert the feedback into rules and integrates it by "morphing" the model architecture accordingly. Parkash and Parikh [119], Biswas and Parikh [25], on the other hand, enable users to specify what attributes make an instance a negative, and use them to acquire negative examples using pre-trained attribute classifiers.

FIND is an alternative approach for interactively debugging models for natural language processing tasks [99]. What sets it apart is that interaction is framed in terms of *sets* of local explanations. FIND builds on the observation that, by construction, local explanations fail to capture how the model behaves in regions far away from the instances for which the user receives explanations. This, in turn, complicates acquiring high-quality supervision and allocating trust [182, 122]. FIND addresses this issue by extracting those words and n-grams that best characterize each latent feature acquired by the model, and then visualizing the relationship between words and features using a word cloud. The characteristic words are obtained using layer-wise relevance propagation [15], a technique akin to input gradients, to all examples in the training set. Based on this information, the user can turn off those latent features that are not relevant for the predictive task. For instance, if a model has learned a latent feature that strongly correlates with a polar word like "`love`" and uses it to categorize documents into non-polar classes such as spam and non-spam, FIND allows to instruct the model to no longer rely on this feature. This is achieved by introducing a hard constraint directly into the prediction process. Other strategies for overcoming the limits of explanation locality are discussed in Section 4.

3.3 Interacting via Example Attributions

Example attributions also have a role to play in interactive debugging. Existing strategies aim to uncover and correct cases where a model relies on "bad" training examples for its predictions, but target different types bugs and elicit different types of feedback. HILDIF [195] uses a fast approximation of influence functions [67] to identify examples in support of a target prediction and and then asks a human-in-the-loop to verify whether the their *level of influence* is justified. It then calibrates the influence of these examples on the model via data augmentation. Specifically, HILDIF tackles NLP tasks and it augments those training examples that are most relevant, as determined by the user, by replacing words by synonyms, effectively boosting their relative influence compared to the others. In this sense, the general idea is reminiscent of XIL, although viewed from the lens of example influence rather than attribute relevance.

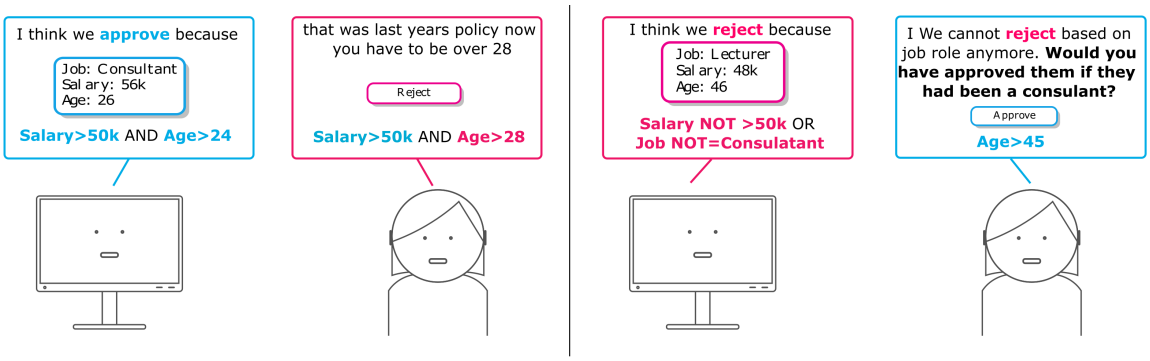


Figure 3: Illustration of rule-based explanation feedback. **Left**: Rule-based explanations shown to the user to support the prediction. **Right**: The input instance does not satisfy any rules to support approve, so the clauses the input instance violates can be shown so the user can validate whether those reasons should be upheld.

A different kind of bug occurs when a model relies on mislabeled examples, which are frequently encounteded in applications, especially when interacting with (even expert) human annotators [55]. CINCER [161] offers a direct way to deal with this issue in sequential learning tasks. To this end, in CINCER the machine monitors incoming examples and asks a user to double-check and re-label those examples that look suspicious [190, 26]. A major issue is that the machine's *skepticism* may be wrong due to – again – presence of mislabeled training examples. This begs the question: is the machine skeptical for the right reasons? CINCER solves this issue by identifying those training examples that most support the model's skepticism using IFs, giving the option of *relabeling* the suspicious incoming example, the influential examples, or both. This process is complementary to HILDIF, as it enables stakeholders to gradually improve the quality of the training data itself, and – as a beneficial side-effect – also to calibrate its influence on the model.

3.4 Benefits and Limitations

Some of the model alignment strategies employed by the approaches overviewed in this section were born for *offline* alignment task [130, 61, 139, 69]. A major issue of this setting is that it is not clear where the ground-truth supervision comes from, as most existing data sets do not come with dense (e.g., pixel- or word-level) relevance annotations. In interactive the settings we consider, this information is naturally provided by a human annotator. Another important advantage is that in interactive settings users are only required to supply feedback tailored for those buggy behaviors exhibited by the model, dramatically reducing the annotation burden. The benefits and potential drawbacks of explanatory interaction are studied in detail by Ghai et al. [62].

4 Interacting via Global Explanations

All explanations discussed so far are *local*, in the sense that they explain a given decision $f(\mathbf{x}) = y$ of a specific input $\mathbf{x}$. Local explanations enable the user to build a mental model of individual predictions, however bringing together the information gleaned by examining a number of individual predictions may prove challenging to get an overall understanding of a model. Lifting this restriction immediately gives *regional* explanations, which are guaranteed to be valid in a larger, well-defined region around $\mathbf{x}$. For instance, the explanations output by decision trees are regional in this sense. Global explanations aim to describe the behavior of f across all of its domain [66] providing an approximate overview of the model [40, 49]. One approach is to train a directly interpretable model such as a decision tree or a rule set, using the same training data optimised for the interpretable model to behave similarly to the original model. This provides a surrogate white-box model of f which can then be used as an approximate global map of f [66, 105]. Other approaches start with local or regional explanations and merge them to provide a global explainer [105, 142].

4.1 Rule-based Explanations

An example rule-based explanation on the loan request approval could be "`Approve` if `Salary` $> 50k$ `AND Age` > 24", as shown in Fig. 1 and Fig. 3. Rule-based explanations decompose the models predictions into simple atomic elements either via decision trees, rule lists or rule sets. While decision trees are similar to rule lists and sets in terms of how they logically represent the decision processes, it is not always the case that decision trees are interpretable from a user perspective, particularly when the number of nodes are large [116]. As a result, for explanations to facilitate user interaction, the size and complexity of the number of rules and clauses need to be considered.

LORE [65] is an example of a local rule-based explainer, which builds an interpretable predictor by first generating a balanced set of neighbour instances of a given data instance through an ad-hoc genetic algorithm, and then building a decision tree classifier out of this set. From this classifier, a local explanation is then extracted, where the local explanation is a pair of logical rules, corresponding to the path in the tree and a set of counterfactual rules, explaining which conditions need to be changed for the data instance to be assigned the inverted class label by the underlying ML model. Both LIME [127], and Anchors [128] are other examples which uses a similar intuition such that, even if the decision boundary for the black box ML model can be arbitrarily complex over the whole data space for a given data instance, there is a high chance that the decision boundary is clear and simple in the neighborhood of a data point which can be captured by an interpretable model. Nanfack et al. learn a global decision rules explainer using a co-learning approach to distill the knowledge of a blackbox model into the decision rules [115].

Rule-based surrogates have the advantage of being interpretable however, in order to achieve coverage, the model must add rules to cover increasingly narrow slices which can in turn negatively impact interpretability. BRCG [42] and GLocalX [142] trade-off fidelity with interpretability along with compactness of the rules to produce a rule-based representation of the model, which makes them more appropriate as the basis for user feedback.

4.2 Interacting via Rule-based Explanations

Rule-based explanations are logical representations of a machine learning model's decision making process which have the benefit of being interpretable to the user. This logical representation can then be modified by end users or domain experts, and these edits bring the advantage that the expected behaviour is more predictable to the end user. Another key advantage of rule-based surrogate models is that, they provide a global representation of the underlying model rather than local explanations which make it more difficult for a user to build up an overall mental model of the system. Additionally, if the local explanations presented to the user to learn and understand the machine learning model are not well distributed, they may create a biased view of the model, leading users to trust a model that for some regions has inaccurate logic.

Lakkaraju et al. [96] produce decision sets where the decision set effectively becomes the predictive model. Popordanoska et al. supports user feedback by generating a rule-based system from data that the user may modify which is then used as a rule-based executable model [122].

Daly et al. [41] present an algorithm which brings together a rule-based surrogate derived by Dash et al. [42] and an underlying machine learning model to support user provided modifications to existing rules in order to incorporate user feedback in a post-processing approach. Similar to CAIPI [160, 137] predictions and explanations are presented to experts, however the explanations are rule-based. The user can then specify changes to the labels and rules by adjusting the clauses of the explanation. For example, the user can modify a value such as change $\texttt{Age} = 26$ to $\texttt{Age} = 30$, remove a clause such as $\texttt{Gender} = \texttt{male}$ or add a clause such as an additional feature requirement such as $\texttt{Employed} = \texttt{true}$. The modified labels and clauses are stored together with transformations that map between original and feedback rules as an Overlay or post-processing approach to the existing model. In a counterfactual style approach, the transformation function is applied to relevant new instances presented to the system in order to understand if the user adjustment influences the final prediction. One advantage of the combined approach is that the rule-based surrogate does not need to encode the full complexity of the model as the underlying prediction still comes from the ML model. The rule-based explanations are used as the source of feedback to provide corrections to key variables. Results showed this approach can support updates and edits to an ML model without retraining as a 'band-aid', however once the intended behaviour diverges too much from the underlying model, results deteriorate. Alkan et al. similarly use rules as the unit for feedback where the input training data is pre-processed in order to produce an ML model that aligns with the user provided feedback rules [8]. The FROTE algorithm generates synthetic instances that reflect both the feedback rules as well as the existing data and has the advantage of encoding the user feedback into the model. As with the previous solution, the advantage is the rules do not need to reflect the entire model, but can focus on the regions where feedback or correction is needed. The results showed that the solution can support modifications to the ML model even when they diverge quite significantly from that of the existing model. Ratner et al. [126] combines data-sources to produce weak labels and one source of labels they consider are user provided rules or labelling functions which can then be tuned. A recent work [68] explored an explanation-driven interactive machine learning (XIML) system with the Tic-Tac-Toe game as a use case to understand how an XIML mechanism effects users' satisfaction with the underlying ML solution. A rule based explainer, BRCG [42], is used for generating explanations, and authors experimented on different modalities to support user feedback through visual or rule-based corrections. They conducted a user study in order to test the effect of allowing users to interact with the rule-based system through creating new rules or editing or removing existing rules. The results of the user study demonstrated that allowing interactivity within the designed XIML system leads to increased satisfaction for the end users.

4.3 Benefits and limitations

An important advantage of leveraging rules to enable user feedback is that, editing a clause can impact many different data points which can aid in reducing the cognitive load for the end user and making the changes more predictable. When modifying a Boolean clause, the feedback and the intended consequences are clearer in comparison to alternative techniques such as re-weighting feature importance or a training data point. However, one challenge with rule based methods is that, feedback can be conflicting in nature, therefore some form of conflict resolution is needed to be implemented in order to allow experts to resolve collisions [8]. Additionally, rules can be probabilistic and eliciting such probabilities from experts can be difficult.

An additional challenge is that, most of the existing rule-based solutions only build upon original features in the data. However, recent work has started to consider this direction for example [5] produces a symbolic equation as a white-box model and [134] supports editing concept based rules for image classification. In the next section, let us therefore continue with a branch of work that potentially allows for interactions on a local and global level via concept-based explanations.

5 Interacting Using Concept-Based Explanations

An advantage of input attributions is that they can be extracted from any black-box model without the need for retraining, thus leaving performance untouched. Critics of these approaches, however, have raised a number of important issues [131]. Perhaps the most fundamental ones, particularly from an interaction perspective, are the potential lack of faithfulness of post-hoc explanations and that input-based explanations are insufficient to accurately capture the reasons behind a model's decision, particularly when these are abstract [154, 168]. Consider an input attribution highlighting a red sports car: does the model's prediction depend on the fact that a car is present, that it is a sports car, or that it is red? This lack of precision can severely complicate interaction and revision.

A possible solution to this issue is to make use of white-box models, which – by design – admit inspecting their whole reasoning process. Models in this category include, e.g., shallow decision trees [12, 131] and sparse linear models [164] based on human-understandable features. These models, however, do not generally support representation learning and struggle with sub-symbolic data.

A more recent solution are concept-based models (CBMs), which combine ideas from white and black-box models to achieve partial, selective interpretability. Since it is difficult – and impractical – to make every step in a decision process fully understandable, CBMs generally break down this task into two levels of processing: a bottom level, where one module (typically black-box) is used for extracting higher-level concept vectors $c_j(\mathbf{x})$, with $j = 1, \ldots, k$, from raw inputs, and a more transparent, top level module in which a decision $y = f(c_1(\mathbf{x}), \ldots, c_k(\mathbf{x}))$ is made *based on the concepts alone*. Most often, the top layer prediction is obtained by performing a weighted aggregation of the extracted concepts. Fig. 1 provides an example of such concept vectors extracted from the raw data in the context of loan requests. Such concept vectors are often of binary form, e.g. a person applying for a loan is either considered a professional or not. The explanation finally corresponds to importance values on these concept vectors.

CBMs combine two key properties. First, the prediction is (roughly) independent from the inputs given the concept vectors. Second, the concepts are chosen to be as human understandable as possible, either by designing them manually or through concept learning, potentially aided by additional concept-level supervision. Taken together, these properties make it possible to faithfully explain a CBMs predictions based on the concept representation alone, thus facilitating interpretability without giving up on representation learning. Another useful feature of CBMs is that they allow for *test-time interventions* to introspect and revise a model's decision based on the individual concept activations [89].

Research on CBMs has explored different representations for the higher-level concepts, including (i) autoencoder-based concepts obtained in an unsupervised fashion and possibly constrained to be independent and dissimilar from each other [9]. (ii) prototype representations that encode concrete training examples or parts thereof [36, 70, 133, 118, 19], (iii) concepts that are explicitly aligned to concept-level supervision provided upfront [89, 38], (iv) white-box concepts obtained by interactively eliciting feature-level dependencies from domain experts [94].

The idea and potential of symbolic, concept-based representations is also found in neuro-symbolic models, although this branch of research was developed from a different standpoint than interpretability alone. Specifically, neuro-symbolic models have recently gained increased interest in the AI community [57, 43, 188, 171] due to their advantages in terms of performance, flexibility and interpretability. The key idea is that these approaches combine handling of sub-symbolic representations with human-understandable, symbolic latent representations. Although it is still an open debate on whether neuro-symbolic approaches are ultimately preferable over purely subsymbolic or symbolic approaches, several recent works have focused on the improvements and richness of neuro-symbolic explanations in the context of understandability and the possibilities of interaction that go beyond the approaches of the previous sections.

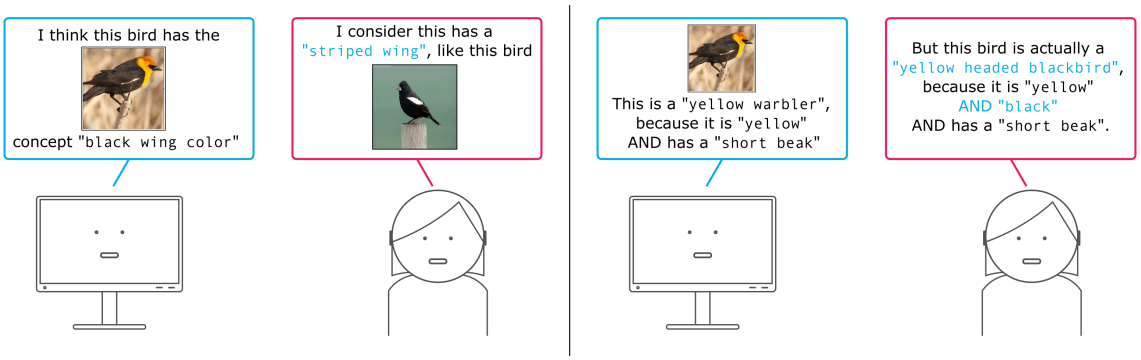


Figure 4: Illustration of providing feedback on concept learning and concept aggregation strategies. **Left**: In concept learning a model learns a set of basic concepts that are present in the dataset. Hereby it learns to grounding specific features of the dataset on symbolic concept labels. **Right**: Given a set of known basic concepts a model could falsely aggregate the concept activations leading to a false final prediction. Users can provide feedback on the concept explanations.

Notably, the distinction between CBMs and neuro-symbolic models can be quite fuzzy, with CBMs possibly considered as one category of neuro-symbolic AI [135, 186].

Although research on concept-level and neuro-symbolic explainability is a flourishing branch of research, it remains quite recent and only selected works have incorporated explanations in an interactive setting. In this section we wish to provide details on these works, but also mention noteworthy works that show potential for leveraging explanations in human machine interactions.

5.1 Interacting with Concept-based Models

Like all machine learning models, CBMs are also prone to erroneous behaviour and may require revision and debugging by human experts [17]. The special structure of CBMs poses challenges that are specific to this setting. A major issue that arises in the context of CBMs is that not only the weights used in the aggregation of the concept activations can be faulty and require adjustment, but the concepts themselves – depending on how they are defined or acquired – can be insufficient, incorrect, or uninterpretable [94, 81].

These two steps have mostly been tackled separately. Fig. 4 gives a brief sketch of this where a human user can guide a model in learning the basic set of concepts from raw images (left), but also provide feedback on the concept aggregations, e.g. in case relevant concepts are ignored for the final class prediction (right).

Several works have tackled interactively learning concepts. For instance, Lage and Doshi-Velez [94] focus on human-machine concept alignment, and propose an interactive loop in which a human expert directly guides concept learning. Specifically, the machine elicits potential dependencies between inputs and concepts by asking questions like, e.g., "does the input variable `lorazepam` contribute to the concept `depression`?". This is reminiscent of FIND [99], but the queried dependencies are chosen by the machine so to be as informative as possible. In their recent work, Stammer et al. [155] propose to use prototype representations for interactively learning concepts and thereby grounding features of image data to discrete (symbolic) concept representations. Via these introspectable encodings a human user can guide the concept learning by directly giving feedback on the prototype activations or by providing paired samples that possess specific concepts which the model should learn. The sample pairing feedback is reminiscent of Shao et al. [144] who proposed debiasing generative models via weak supervision. Bontempelli et al. [28] on the other hand propose debugging part-prototype networks via interactive supervision on the learned part-prototypes using a simple interface in which the user can determine signal that concept is valid or not valid for a particular precision using a single click. They also remark that this kind of concept-level feedback on concepts generalizes is very rich, in that it naturally generalizes to all instances that feature those concepts, facilitating debugging from few interaction rounds.

Other works focus on the concept aggregation step. For instance, Teso [159] applied explanatory interactive learning to self-explainable neural networks, enabling end-users to fine-tune the aggregation weights [9]. An interesting connection to causal learning is made by Bahadori and Heckerman [17], who present an alternative and principled approach for debiasing CBMs from confounders. These two strategies, however, assume the concepts to be given and fixed, which is often not the case.

Finally, Bontempelli et al. [27] outline a unifying framework for debugging CBMs that clearly distinguishes between bugs in the concept set and bugs in the aggregation step, and advocate a multi-step procedure that encompasses determining where the source of the error lies and providing corrective supervision accordingly.

5.2 Interacting with Neuro-symbolic Models

Neuro-symbolic models, similar to CBMs, also support communicating with users in terms of symbolic, higher-level concepts as well as more general forms of constraints and knowledge. There are different ways in which these symbols can be presented and the interaction can be structured.

Ciravegna et al. [39], propose to integrate deep learning with expressive human-driven first-order logic (FOL) explanations. Specifically, a neural network maps to a symbolic concept space and is followed by an additional network that learns to create FOL explanations from the latent concept space. Thanks to this direct translation into FOL explanations, it is in principle easy to integrate prior knowledge from expert users as constraints onto the model.

On the other hand Stammer et al. [154] focus more on human-in-the-loop learning. With their work, the authors show that receiving explanations on the level of the raw input – as done by standard approaches presented in the previous sections – can be insufficient for removing Clever Hans behavior, and show how this problem can be solved by integrating and interacting with rich, neuro-symbolic explanations. Specifically, the authors incorporate a right for the right reason loss [130] on a neural reasoning module which receives multi-object concept activations as input. The approach allows for concept-level explanations, and user feedback is cast in terms of relational logic statements.

The benefits of symbolic explanations and feedback naturally extend from supervised learning and computer vision domains, as in the previous works, to settings like planning [34] and, most recently, reinforcement learning (RL). In particular, in the context of deep RL, Guan et al. [64] provide coarse symbolic feedback in the form of object-centric image regions to accompany binary feedback on an agent's proposed actions. Another interesting use of symbolic explanations in RL is that of Zha et al. [191], in which an RL agent learns to better understand human demonstrations by grounding these in human-aligned symbolic representations.

5.3 Benefits and Limitations

The common underlying motivation of all these approaches is the potential of symbolic communication to improve both precision and bandwidth of explanatory interaction between (partially sub-symbolic) AI models and human stakeholders. A recent paper by Kambhampati et al. [81] provides an excellent motivation on the importance of this property as well as important remaining challenges. The main challenge of concept-based and neuro-symbolic models lies in identifying a set of basic concepts or symbols [186] and grounding a model's latent representations on these. Though recent works, e.g. Lage and Doshi-Velez [94] and Stammer et al. [155], have started to tackle this it remains an important issue to solve particularly for real world data.

6 Open Problems

Despite recent progress on integrating explanations and interaction, many unresolved problems remain. In this section, we outline a selection of urgent open issues and, wherever possible, suggest possible directions forward, with the hope of spurring further research on this important topic.

6.1 Handling Human Factors

Machine explanations are only effective insomuch as they are understood by the person at the receiving end. Simply ensuring algorithmic *transparency*, which is perhaps the major focus in current XAI research, gives few guarantees, because understanding strongly depends on human factors such as mental skills and familiarity with the application domain [152, 101]. As a matter of fact, factual but ill-designed machine guidance can actively harm understanding [4].

Perhaps the most critical element for successful explanation-based interaction requires is that the user and the machine to agree on the *semantics* of the explanations they exchange. This is however problematic, partly because conveying this information to users is non-trivial, and partly because said semantics are often unclear or brittle to begin with, as discussed in Sections 6.2 and 6.3. The literature does provide some guidance on what classes of explanations may be better suited for IML. Existing studies suggest that people find it easier to handle explanations that express concrete cases [86] and that have a counterfactual flavour [170], and that breaking larger computations into modules facilitates simulation [95], but more work is needed to implement and evaluate these suggestions in the context of explanation-based IML. Moreover, settings in which users have also to *manipulate* or *correct* the machine's explanations,

like interactive debugging, impose additional requirements. Another key issue is that of cognitive biases. For instance, human subjects tend to pay more attention to affirmative aspects of counterfactuals [29], while AIs have no such bias. Coping with these human factors requires to design appropriate interaction and incorporation strategies, and it is necessary for correct and robust operation of explanation-based IML.

We also remark that different stakeholders inevitably need different kinds of explanations [110, 101]. A promising direction of research is to enable users to customize the machine's explanations to their needs [152, 54]. Challenges on this path include developing strategies for eliciting the necessary feedback and assisting users in exploring the space of alternatives.

6.2 Semantics and Faithfulness

Not all kinds of machine explanations are equally intelligible and not all XAI algorithms are equally reliable. For instance, some gradient-based attribution techniques fail to satisfy intuitive properties (like implementation invariance [158]) or ignore information at the top layers of neural networks [2], while sampling-based alternatives may suffer from high variance [193, 159]. A number of other issues have been identified in the literature [79, 87, 3, 149, 93]. The semantics of transparent models is also not always well-defined. For instance, the coefficients of linear models are often viewed as capturing feature importance in an additive manner [127], but this interpretation is only valid as long as the input features are independent, which is seldom the case in practice. Decision tree-based explanations have also received substantial scrutiny [80].

Another critical element is faithfulness. The reason is that bugs identified by unfaithful explanations may be artifacts in the explanation rather than actual issues with the model, meaning that asking users to correct them is not only deceptive, but also uninformative for the machine and ultimately wasteful [160]. The *distribution* of machine explanations is equally important for ensuring faithfulness: individually faithful local explanations that fail to cover the whole range of machine behaviors may end up conveying an incomplete [99] and deceptively optimistic [122] picture of the model's logic to stakeholders and annotators.

Still, some degree of unfaithfulness is inevitable, for both computational and cognitive reasons. On the one hand, interaction should not overwhelm the user [92]. This entails presenting a necessarily simplified picture of the (potentially complex) inference process carried out by the machine. On the other hand, extracting faithful explanations often comes at a substantial computational cost [48], which is especially problematic in interactive settings where excessive repeated delays in the interaction can estrange the end-user. Alas, more light-weight XAI strategies tend to rely on approximations and leverage less well-defined explanation classes.

6.3 Abstraction and Explanation Requirements

Many attribution methods are restricted to measuring relevance of individual input variables or training examples. In stark contrast, explanations used in human–human communication convey information at a more abstract, conceptual level, and as such enjoy improved expressive power. An important open research question is how to enable machines to communicate using such higher-level concepts, especially in the context of the approaches discussed in Section 5.

This immediately yields the issue of obtaining a relevant, user-understandable symbolic concept space [81, 132]. This is highly non-trivial. In many cases it might not be obvious what the relevant higher-level concepts should be, and more generally it is not clear what properties should be enforced on these concepts – when learned from data – so to encourage interpretability. Existing candidates include similarity to concrete examples [36], usage of generative models [83], and enforcing disentanglement among concepts [136, 155]. One critical challenge is that imperfections in the learned concepts may compromise the predictive power of a model as well as the the semantics of explanations while being hard to spot [117, 76, 90, 107, 108], calling for the development of concept-level debugging strategies [27]. Additionally, assuming a basic set of concepts has been identified, it seems likely that this set will not be sufficient and should allow for expanding [81].

On the broader topic of explanation requirements, Liao and Varshney [101] discuss many different aspects that XAI brings, such as the *diverse explainability needs* of stakeholders due to the no "one-fits-all" solutions from the collection of XAI algorithms, and *pitfalls of XAI* in the sense that there can be a gap between algorithmic explanations that several XAI works provide and the actionable understanding that these solutions can facilitate. One important statement that is highlighted in [101] is that, closing the gap between *technicality of XAI* and the *user's engagement with the explanations* requires considering possible user interactions with XAI, and operationalizing human-centered perspectives in XAI algorithms.

This latter point also requires developing evaluation methods that better consider the actual user needs in the downstream usage contexts. Tis further raises the question of whether "good" explanations actually exist and how one can quantify

these. One interesting direction forward is to consider explanation approaches in which a user can further query an initial model's explanation similar to how humans provide additional (detailed) queries in case the initial explanation is confusing or insufficient.

6.4 Modulating and Manipulating Trust

In light of the ethical concern of deploying ML models in more real-world applications, the field of *trustworthy ML* has grown, which studies and pursues desirable qualities such as fairness, explainability, transparency, privacy and robustness [165]. As discussed by Liao and Varshney [101], explainability has moved beyond providing details to comprehend the ML models being developed, and it has rather become an essential requirement for people to trust and adopt AI and ML solutions.

However, the relationship between (high-quality) explanations and trust is not straightforward. One reason is that explanations are not the only contributing factor [174]. However, while user studies support the idea that explanations enable stakeholders to *reject* trust in misbehaving models, the oft stated claim that explanations help to *establish* trust into deserving models enjoys less empirical support. This is related to a trend, observed in some user studies, that participants may put too much trust in AI models to begin with [137], thus making the effects of additional explanations less obvious. Understanding also plays a role. Failure to understand an explanation may drive users to immediately and unjustifiably distrust the system [78] and, rather surprisingly, in some cases the mere fact of being exposed to the machine's internal reasoning may induce a loss of trust [78]. More generally, the link between interaction and trust is under-explored.

Another important issue that explanations can be intentionally manipulated by malicious parties so to persuade stakeholders into trusting unfair or buggy systems [50, 72, 11]. This is a serious concern for the entire enterprise of explainable AI and therefore for explanatory interaction. Despite initial efforts on making explanations robust against such attacks, more work is still needed to guarantee safety for models used in practice.

6.5 Annotation Effort and Quality

Explanatory supervision comes with its own set of challenges. First and foremost, just like other kinds of annotations, corrections elicited from human supervisors are bound to be affected by noise, particularly when interacting with non-experts [55]. Since explanations potentially convey much more information than simple labels, the impact of noise is manifold. One option to facilitate ensuring high-quality supervision is that of providing human annotators with guidance, cf. [30], but this cannot entirely prevent annotation noise. In order to cope with this, it is critical to develop learning algorithms that are robust to noisy explanatory supervision. An alternative is to leverage interactive strategies for enabling users to identify and rectify mislabeled examples identified by the machine [190, 161].

Some forms of explanatory supervision – for instance, pixel-level relevance annotations – require higher effort on the annotator's end. This extra cost is often justified by the larger impact that explanatory supervision has on the model compared to pure label information, but in most practical application effort is constrained by a tight budget and must be kept under control. Analogously to what is normally done in interactive learning [141], doing so involves developing querying strategies that only eliciting explanatory supervision when strictly necessary (*e.g.*, when label information is not enough) and to efficiently identify the most informative queries. To the best of our knowledge, this problem space is completely unexplored.

6.6 Benchmarking and Evaluation

Evaluating and benchmarking current and novel approaches that integrate explanations and interaction is particularly challenging. There are several reasons for this. One reason is the effort in providing extensive user studies, where many recent studies tend to focus on simulated user interactions for evaluations. The difficulties for this are not just the participant organization and proper study design itself (which can lead to many pitfalls if not properly devised), but also from the engineering perspective a swift user feedback integration is not immediate in all studies. Thus, in many cases additional engineering and an extensive user study remain necessary before real-world deployment.

Further reasons are the individual use case of a method, but also the possibly high variance in user's feedback which make it challenging to assess a methods properties with one task and study alone. A very important branch for future research is thus to develop more standardized benchmarking tasks and metrics for evaluating such methods, where Friedrich et al. [56] provide an initial set of important evaluation metrics and tasks for future research on XIL.

7 Related Topics

Research on integrating interaction and explanations builds on and is related to a number of different topics. The main source of inspiration is the vast body of knowledge on explainable AI, which has been summarized in several overviews and surveys [66, 125, 101, 63, 114, 1, 33, 153, 23]. A major difference to this literature is that in XAI the communication between machine and user stops after receiving the machine's explanations.

Recommender systems are another area where explanations have found wide applicability, with a large number of approaches being proposed and applied in real-world systems in recent years. Compatibly with our discussion, explanations has been shown to improve transparency [148, 163] and trustworthiness [192] of recommendations, to help users make better decisions, and to enable users to provide feedback to the system by correcting any incorrect assumptions that the recommender has made for their interests [6]. With critiquing-based recommenders, users can critique the presented suggestions and provide their preferences [13, 179, 37]. Here we focus on parallel advancements made in interactive ML, and refer the reader to Zhang et al. [194] for a comprehensive review of explainable recommender systems.

The idea of using explanations as supervision in ML can be traced back to explanation-based learning [112, 47], where the goal is to extract logical concepts by generalizing from symbolic explanations, although the overall framework is not restricted to purely logical data.

Another major source of inspiration are approaches for *offline* explanation-based debugging, a topic has recently received substantial attention, especially in natural language processing community [98]. This topic encompasses, for instance, learning from annotator rationales [189], i.e., from snippets of text appearning in a target sentence that support a particular decision, and that effectively the same role as input attributions. Recent works have extended this setup to complex natural language inference tasks [31]. Another closely related topic is learning from feature-level supervision, which has been used to complement standard label supervision for learning higher quality predictors [45, 124, 123, 51, 52, 14, 151, 140]. These earlier approaches assume (typically complete) supervision about attribute-level relevance to be provided in advance, independently from the model's explanations, and focus on shallow models. More generally, human explanations can be interpreted as encoding prior knowledge and thus used guide the learning process towards better aligned models faster [69]. This perspective overlaps with and represents a special case of informed ML [169, 22].

Two other areas where explanations have been used to inform the learning process are explanation-based distillation and regularization. The aim of distillation is to compress a larger model into a smaller one for the purpose of improving efficiency and storage costs. It was shown that paying attention to the explanations of the source model can tremendously (and provably) improve the sample complexity of distillation [111], hinting at the potential of explanations as a rich source of supervision. Regularization instead aims at encouraging models to produce more simulatable [180, 181] or faithful explanations [121] by introducing an additional penalty into the learning process. Here, however, no explanatory supervision is involved.

Two types of explanations that we have not delved into are counterfactuals and attention. Counterfactuals identify changes necessary for achieving alternative and possibly more desirable outcomes [170], for instance what should be changed in a loan application in order for it to be approved. They have become popular in XAI as a mean to help stakeholders to form actionable plans and control the system's behavior [102, 82], but have recently shown promise as a mean to design novel interaction strategies [84, 183, 44]. Attention mechanisms [18, 166] also offer insight into the decision process of neural networks and, although their interpretation is somewhat controversial [20], they are a viable alternative to gradient-based attributions for integrating explanatory feedback into the model in an end-to-end manner [113, 73].

Finally, another important aspect that we had to omit is causality [120], which is perhaps the most solid foundation for imbuing cause-effect relationships within ML models [35, 184], and conversely the ability to identify causal relationships between inputs and predictions constitutes a fundamental step towards explaining model predictions [77, 59]. Work on causality in interactive ML is however sparse at best.

8 Conclusion

This overview provides a conceptual guide on current research on integrating explanations into interactive machine learning for the purpose of establishing a rich bi-directional communication loop between machine learning models and human stakeholders in a way that is beneficial to all parties involved. Explanations make it possible for users to better understand the machine's behavior, spot possible limitations and bugs in its reasoning patterns, establish control over the machine, and modulate trust. At the same time, the machine obtains high-quality, informed feedback

in exchange. We categorized existing approaches along four dimensions, namely algorithmic goal, type of machine explanations involved, human feedback received, and incorporation strategy, facilitating the identification of links between different approaches as well as respective strengths and limitations. In addition, we identified a number of open problems impacting the human and machine sides of explanatory interaction and highlighted noteworthy paths of future, with the goal of spurring further research into this novel and promising approach, helping to bridge the gap towards human-centric machine learning and AI.

References

[1] A. Adadi and M. Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[2] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *International Conference on Neural Information Processing Systems*, pages 9525–9536, 2018.

[3] J. Adebayo, M. Muelly, I. Liccardi, and B. Kim. Debugging tests for model explanations. In *Conference on Neural Information Processing Systems*, 2020.

[4] L. Ai, S. H. Muggleton, C. Hocquette, M. Gromowski, and U. Schmid. Beneficial and harmful explanatory machine learning. *Machine Learning*, 110(4):695–721, 2021.

[5] A. M. Alaa and M. van der Schaar. Demystifying black-box models with symbolic metamodels. *Advances in Neural Information Processing Systems*, 32, 2019.

[6] Ö. Alkan, E. M. Daly, A. Botea, A. N. Valente, and P. Pedemonte. Where can my career take me? harnessing dialogue for interactive career goal recommendations. In *International Conference on Intelligent User Interfaces*, pages 603–613, 2019.

[7] O. Alkan, M. Mattetti, E. M. Daly, A. Botea, I. Vejsbjerg, and B. Knijnenburg. Irf: A framework for enabling users to interact with recommenders through dialogue. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1), 2021.

[8] O. Alkan, D. Wei, M. Mattetti, R. Nair, E. Daly, and D. Saha. Frote: Feedback rule-driven oversampling for editing models. In D. Marculescu, Y. Chi, and C. Wu, editors, *Proceedings of Machine Learning and Systems*, volume 4, pages 276–301, 2022.

[9] D. Alvarez-Melis and T. S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *International Conference on Neural Information Processing Systems*, pages 7786–7795, 2018.

[10] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *Ai Magazine*, 35(4):105–120, 2014.

[11] C. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Müller, and P. Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pages 314–323, 2020.

[12] E. Angelino, N. Larus-Stone, D. Alabi, M. I. Seltzer, and C. Rudin. Learning certifiably optimal rule lists for categorical data. *J. Mach. Learn. Res.*, 18:234:1–234:78, 2017.

[13] D. Antognini, C. Musat, and B. Faltings. Interacting with explanations through critiquing. In *International Joint Conference on Artificial Intelligence*, pages 515–521, 2021.

[14] J. Attenberg, P. Melville, and F. Provost. A unified approach to active dual supervision for labeling features and examples. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 40–55. Springer, 2010.

[15] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10, 2015.

[16] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

[17] M. T. Bahadori and D. Heckerman. Debiasing concept-based explanations with causal analysis. In *International Conference on Learning Representations*, 2021.

[18] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*, 2015.

[19] A. J. Barnett, F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin. Iaia-bl: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *arXiv preprint arXiv:2103.12308*, 2021.

[20] J. Bastings and K. Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, 2020.

[21] S. Basu, P. Pope, and S. Feizi. Influence functions in deep learning are fragile. In *International Conference on Learning Representations*, 2021.

[22] K. Beckh, S. Müller, M. Jakobs, V. Toborek, H. Tan, R. Fischer, P. Welke, S. Houben, and L. von Rueden. Explainable machine learning with prior knowledge: An overview. *arXiv preprint arXiv:2105.10172*, 2021.

[23] V. Belle and I. Papantonis. Principles and practice of explainable machine learning. *Frontiers in big Data*, page 39, 2021.

[24] J. Bien and R. Tibshirani. Prototype selection for interpretable classification. *The Annals of Applied Statistics*, 5 (4):2403–2424, 2011.

[25] A. Biswas and D. Parikh. Simultaneous active learning of classifiers and attributes via relative feedback. In *Conference on Computer Vision and Pattern Recognition*, pages 644–651, 2013.

[26] A. Bontempelli, S. Teso, F. Giunchiglia, and A. Passerini. Learning in the Wild with Incremental Skeptical Gaussian Processes. In *International Joint Conference on Artificial Intelligence*, 2020.

[27] A. Bontempelli, F. Giunchiglia, A. Passerini, and S. Teso. Toward a unified framework for debugging gray-box models. In *The AAAI-22 Workshop on Interactive Machine Learning*, 2021.

[28] A. Bontempelli, S. Teso, F. Giunchiglia, and A. Passerini. Concept-level debugging of part-prototype networks. *arXiv preprint arXiv:2205.15769*, 2022.

[29] R. M. J. Byrne. Counterfactuals in explainable artificial intelligence (XAI): evidence from human reasoning. In *International Joint Conference on Artificial Intelligence*, pages 6276–6282, 2019.

[30] M. Cakmak and A. L. Thomaz. Eliciting good teaching from humans for machine learners. *Artificial Intelligence*, 217:198–215, 2014.

[31] O.-M. Camburu, T. Rocktäschel, T. Lukasiewicz, and P. Blunsom. e-snli: natural language inference with natural language explanations. In *Conference on Neural Information Processing Systems*, pages 9560–9572, 2018.

[32] O.-M. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz, and P. Blunsom. The struggles of feature-based explanations: Shapley values vs. minimal sufficient subsets. *arXiv preprint arXiv:2009.11023*, 2020.

[33] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[34] T. Chakraborti, S. Sreedharan, S. Grover, and S. Kambhampati. Plan explanations as model reconciliation–an empirical study. In *International Conference on Human-Robot Interaction*, pages 258–266, 2019.

[35] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian. Neural network attributions: A causal perspective. In *International Conference on Machine Learning*, pages 981–990, 2019.

[36] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, and J. K. Su. This looks like that: Deep learning for interpretable image recognition. *Conference on Neural Information Processing Systems*, 32:8930–8941, 2019.

[37] L. Chen and P. Pu. Critiquing-based recommenders: survey and emerging trends. *User Modeling and User-Adapted Interaction*, 22(1):125–150, 2012.

[38] Z. Chen, Y. Bei, and C. Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[39] G. Ciravegna, F. Giannini, M. Gori, M. Maggini, and S. Melacci. Human-driven fol explanations of deep learning. In *International Joint Conference on Artificial Intelligence*, pages 2234–2240, 2020.

[40] M. Craven and J. Shavlik. Extracting tree-structured representations of trained networks. *Conference on Neural Information Processing Systems*, 8:24–30, 1995.

[41] E. M. Daly, M. Mattetti, Ö. Alkan, and R. Nair. User driven model adjustment via boolean rule explanations. In *Conference on Artificial Intelligence*, volume 35, pages 5896–5904, 2021.

[42] S. Dash, O. Gunluk, and D. Wei. Boolean decision rules via column generation. *Conference on Neural Information Processing Systems*, 31:4655–4665, 2018.

[43] A. S. d'Avila Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, 6, 2019.

[44] G. De Toni, P. Viappiani, B. Lepri, and A. Passerini. Generating personalized counterfactual interventions for algorithmic recourse by eliciting user preferences. *arXiv preprint arXiv:2205.13743*, 2022.

[45] D. DeCoste and B. Schölkopf. Training invariant support vector machines. *Machine learning*, 46(1):161–190, 2002.

[46] A. J. DeGrave, J. D. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, pages 1–10, 2021.

[47] G. DeJong and R. Mooney. Explanation-based learning: An alternative view. *Machine learning*, 1(2):145–176, 1986.

[48] G. V. den Broeck, A. Lykov, M. Schleich, and D. Suciu. On the tractability of SHAP explanations. In *Conference on Artificial Intelligence*, pages 6505–6513, 2021.

[49] H. Deng. Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, 7(4): 277–287, 2019.

[50] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32:13589–13600, 2019.

[51] G. Druck, G. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602, 2008.

[52] G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. In *Conference on Empirical Methods in Natural Language Processing*, pages 81–90, 2009.

[53] J. A. Fails and D. R. Olsen Jr. Interactive machine learning. In *International Conference on Intelligent User Interfaces*, pages 39–45, 2003.

[54] B. Finzel, D. E. Tafler, S. Scheele, and U. Schmid. Explanation as a process: user-centric construction of multi-level and multi-modal explanations. In *German Conference on Artificial Intelligence (Künstliche Intelligenz)*, pages 80–94, 2021.

[55] B. Frénay and M. Verleysen. Classification in the presence of label noise: A survey. *Trans. Neural Networks Learn. Syst.*, 25(5):845–869, 2014.

[56] F. Friedrich, W. Stammer, P. Schramowski, and K. Kersting. A typology to explore and guide explanatory interactive machine learning. *arXiv preprint arXiv:2203.03668*, 2022.

[57] A. S. d. Garcez, K. B. Broda, and D. M. Gabbay. *Neural-symbolic learning systems: foundations and applications*. Springer Science & Business Media, 2012.

[58] D. Garreau and U. Luxburg. Explaining the explainer: A first theoretical analysis of LIME. In *International Conference on Artificial Intelligence and Statistics*, pages 1287–1296, 2020.

[59] A. Geiger, H. Lu, T. Icard, and C. Potts. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34, 2021.

[60] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[61] R. Ghaeini, X. Fern, H. Shahbazi, and P. Tadepalli. Saliency learning: Teaching the model where to pay attention. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4016–4025, 2019.

[62] B. Ghai, Q. V. Liao, Y. Zhang, R. Bellamy, and K. Mueller. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *ACM on Human-Computer Interaction*, 4(CSCW3):1–28, 2021.

[63] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. Explaining explanations: An overview of interpretability of machine learning. In *International Conference on Data Science and Advanced Analytics*, pages 80–89, 2018.

[64] L. Guan, M. Verma, S. Guo, R. Zhang, and S. Kambhampati. Widening the pipeline in human-guided reinforcement learning with explanation and context-aware data augmentation. In *International Conference on Neural Information Processing Systems*, 2021.

[65] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.

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

[67] H. Guo, N. Rajani, P. Hase, M. Bansal, and C. Xiong. FastIF: Scalable Influence Functions for Efficient Model Interpretation and Debugging. In *Conference on Empirical Methods in Natural Language Processing*, pages 10333–10350, 2021.

[68] L. Guo, E. M. Daly, Ö. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg. Building trust in interactive machine learning via user contributed interpretable rules. In *International Conference on Intelligent User Interfaces*, 2022.

[69] P. Hase and M. Bansal. When can models learn from explanations? a formal framework for understanding the roles of explanation data. *arXiv preprint arXiv:2102.02201*, 2020.

[70] P. Hase, C. Chen, O. Li, and C. Rudin. Interpretable image recognition with hierarchical prototypes. In *Conference on Human Computation and Crowdsourcing*, volume 7, pages 32–40, 2019.

[71] C. He, D. Parra, and K. Verbert. Interactive recommender systems: A survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 56:9–27, 2016.

[72] J. Heo, S. Joo, and T. Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, 32:2925–2936, 2019.

[73] J. Heo, J. Park, H. Jeong, K. J. Kim, J. Lee, E. Yang, and S. J. Hwang. Cost-effective interactive attention learning with neural attention processes. In *International Conference on Machine Learning*, pages 4228–4238, 2020.

[74] M. Herde, D. Huseljic, B. Sick, and A. Calma. A survey on cost types, interaction schemes, and annotator performance models in selection algorithms for active learning in classification. *IEEE Access*, 9:166970–166989, 2021.

[75] R. R. Hoffman et al. Trust in automation. *IEEE Intelligent Systems*, 28(1):84–88, 2013.

[76] A. Hoffmann, C. Fanconi, R. Rade, and J. Kohler. This looks like that... does it? shortcomings of latent space prototype interpretability in deep networks. *arXiv preprint arXiv:2105.02968*, 2021.

[77] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller. Causability and explainabilty of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9:e1312, 07 2019.

[78] D. Honeycutt, M. Nourani, and E. Ragan. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Conference on Human Computation and Crowdsourcing*, volume 8, pages 63–72, 2020.

[79] S. Hooker, D. Erhan, P. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. In *Conference on Neural Information Processing Systems*, pages 9734–9745, 2019.

[80] Y. Izza, A. Ignatiev, and J. Marques-Silva. On explaining decision trees. *arXiv preprint arXiv:2010.11034*, 2020.

[81] S. Kambhampati, S. Sreedharan, M. Verma, Y. Zha, and L. Guan. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. *arXiv preprint arXiv:2109.09904*, 2021.

[82] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[83] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[84] D. Kaushik, E. Hovy, and Z. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2019.

[85] R. Khanna, B. Kim, J. Ghosh, and S. Koyejo. Interpreting black box predictions using fisher kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 3382–3390, 2019.

[86] B. Kim, C. Rudin, and J. A. Shah. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Conference on Neural Information Processing Systems*, pages 1952–1960, 2014.

[87] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[88] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894, 2017.

[89] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[90] S. Kraft, K. Broelemann, A. Theissler, G. Kasneci, G. Esslingen am Neckar, S. H. AG, G. Wiesbaden, and G. Aalen. Sparrow: Semantically coherent prototypes for image classification. 2021.

[91] T. Kulesza, S. Stumpf, M. Burnett, W.-K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel, and K. McIntosh. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *Symposium on Visual Languages and Human-Centric Computing*, pages 41–48, 2010.

[92] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *International Conference on Intelligent User Interfaces*, pages 126–137, 2015.

[93] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pages 5491–5500, 2020.

[94] I. Lage and F. Doshi-Velez. Learning interpretable concept-based models with human feedback. *arXiv preprint arXiv:2012.02898*, 2020.

[95] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez. Human evaluation of models built for interpretability. In *AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67, 2019.

[96] H. Lakkaraju, S. H. Bach, and J. Leskovec. Interpretable decision sets: A joint framework for description and prediction. In *International Conference on Knowledge Discovery and Data Mining*, pages 1675–1684, 2016.

[97] 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*, 10(1):1–8, 2019.

[98] P. Lertvittayakumjorn and F. Toni. Explanation-based human debugging of nlp models: A survey. *arXiv preprint arXiv:2104.15135*, 2021.

[99] P. Lertvittayakumjorn, L. Specia, and F. Toni. Find: human-in-the-loop debugging deep text classifiers. In *Conference on Empirical Methods in Natural Language Processing*, pages 332–348, 2020.

[100] W. Liang, J. Zou, and Z. Yu. Alice: Active learning with contrastive natural language explanations. In *Conference on Empirical Methods in Natural Language Processing*, pages 4380–4391, 2020.

[101] Q. V. Liao and K. R. Varshney. Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*, 2021.

[102] B. Y. Lim. *Improving understanding and trust with intelligibility in context-aware applications*. PhD thesis, Carnegie Mellon University, 2012.

[103] T. Lombrozo. The structure and function of explanations. *Trends in cognitive sciences*, 10(10):464–470, 2006.

[104] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[105] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2:56–67, 2020.

[106] O. Mac Aodha, S. Su, Y. Chen, P. Perona, and Y. Yue. Teaching categories to human learners with visual explanations. In *Conference on Computer Vision and Pattern Recognition*, pages 3820–3828, 2018.

[107] A. Mahinpei, J. Clark, I. Lage, F. Doshi-Velez, and W. Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

[108] A. Margeloiu, M. Ashman, U. Bhatt, Y. Chen, M. Jamnik, and A. Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021.

[109] C. J. Michael, D. Acklin, and J. Scheuerman. On interactive machine learning and the potential of cognitive feedback, 2020.

[110] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38, 2019.

[111] S. Milli, L. Schmidt, A. D. Dragan, and M. Hardt. Model reconstruction from model explanations. In *Conference on Fairness, Accountability, and Transparency*, pages 1–9, 2019.

[112] T. M. Mitchell, R. M. Keller, and S. T. Kedar-Cabelli. Explanation-based generalization: A unifying view. *Machine learning*, 1(1):47–80, 1986.

[113] M. Mitsuhara, H. Fukui, Y. Sakashita, T. Ogata, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. Embedding human knowledge into deep neural network via attention map. *arXiv preprint arXiv:1905.03540*, 2019.

[114] G. Montavon, W. Samek, and K.-R. Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[115] G. Nanfack, P. Temple, and B. Frénay. Global explanations with decision rules: a co-learning approach. In *Conference on Uncertainty in Artificial Intelligence*, pages 589–599, 2021.

[116] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682*, 2018.

[117] M. Nauta, A. Jutte, J. Provoost, and C. Seifert. This looks like that, because... explaining prototypes for interpretable image recognition. *arXiv preprint arXiv:2011.02863*, 2020.

[118] M. Nauta, R. van Bree, and C. Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 14933–14943, 2021.

[119] A. Parkash and D. Parikh. Attributes for classifier feedback. In *european conference on computer vision*, pages 354–368. Springer, 2012.

[120] J. Pearl. *Causality*. Cambridge university press, 2009.

[121] G. Plumb, M. Al-Shedivat, Á. A. Cabrera, A. Perer, E. Xing, and A. Talwalkar. Regularizing black-box models for improved interpretability. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

[122] T. Popordanoska, M. Kumar, and S. Teso. Machine guides, human supervises: Interactive learning with global explanations. *arXiv preprint arXiv:2009.09723*, 2020.

[123] H. Raghavan and J. Allan. An interactive algorithm for asking and incorporating feature feedback into support vector machines. In *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 79–86, 2007.

[124] H. Raghavan, O. Madani, and R. Jones. Active learning with feedback on features and instances. *The Journal of Machine Learning Research*, 7:1655–1686, 2006.

[125] G. Ras, N. Xie, M. van Gerven, and D. Doran. Explainable deep learning: A field guide for the uninitiated. *Journal of Artificial Intelligence Research*, 73:329–397, 2022.

[126] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access, 2017.

[127] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[128] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Conference on Artificial Intelligence*, volume 32, 2018.

[129] L. Rieger, C. Singh, W. Murdoch, and B. Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *International Conference on Machine Learning*, pages 8116–8126. PMLR, 2020.

[130] 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*, pages 2662–2670, 2017.

[131] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[132] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

[133] D. Rymarczyk, Ł. Struski, J. Tabor, and B. Zieliński. Protopshare: Prototype sharing for interpretable image classification and similarity discovery. *arXiv preprint arXiv:2011.14340*, 2020.

[134] S. Santurkar, D. Tsipras, M. Elango, D. Bau, A. Torralba, and A. Madry. Editing a classifier by rewriting its prediction rules. *Conference on Neural Information Processing Systems*, 34, 2021.

[135] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler. Neuro-symbolic artificial intelligence. *AI Commun.*, 34(3): 197–209, 2021.

[136] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[137] P. Schramowski, W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

[138] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, pages 618–626, 2017.

[139] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *International Conference on Computer Vision*, pages 2591–2600, 2019.

[140] B. Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, 2011.

[141] B. Settles. Active learning: Synthesis lectures on artificial intelligence and machine learning. *Long Island, NY: Morgan & Clay Pool*, 2012.

[142] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. Glocalx-from local to global explanations of black box ai models. *Artificial Intelligence*, 294:103457, 2021.

[143] X. Shao, A. Skryagin, P. Schramowski, W. Stammer, and K. Kersting. Right for better reasons: Training differentiable models by constraining their influence function. In *Conference on Artificial Intelligence*, 2021.

[144] X. Shao, K. Stelzner, and K. Kersting. Right for the right latent factors: Debiasing generative models via disentanglement. *arXiv preprint arXiv:2202.00391*, 2022.

[145] A. Shih, A. Choi, and A. Darwiche. A symbolic approach to explaining bayesian network classifiers. In *International Joint Conference on Artificial Intelligence*, pages 5103–5111, 2018.

[146] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.

[147] C. Singh, W. J. Murdoch, and B. Yu. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations*, 2018.

[148] R. R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *Conference on Human Factors in Computing Systems*, pages 830–831, 2002.

[149] L. Sixt, M. Granz, and T. Landgraf. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pages 9046–9057. PMLR, 2020.

[150] E. Slany, Y. Ott, S. Scheele, J. Paulus, and U. Schmid. Caipi in practice: Towards explainable interactive medical image classification. *arXiv preprint arXiv:2204.02661*, 2022.

[151] K. Small, B. C. Wallace, C. E. Brodley, and T. A. Trikalinos. The constrained weight space svm: learning with ranked features. In *International Conference on International Conference on Machine Learning*, pages 865–872, 2011.

[152] K. Sokol and P. Flach. One explanation does not fit all. *KI-Künstliche Intelligenz*, pages 1–16, 2020.

[153] K. Sokol and P. A. Flach. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In *Conference on Fairness, Accountability, and Transparency*, pages 56–67, 2020.

[154] W. Stammer, P. Schramowski, and K. Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Conference on Computer Vision and Pattern Recognition*, pages 3619–3629, 2021.

[155] W. Stammer, M. Memmel, P. Schramowski, and K. Kersting. Interactive disentanglement: Learning concepts by interacting with their prototype representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10317–10328, 2022.

[156] E. Štrumbelj and I. Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

[157] S. Stumpf, V. Rajaram, L. Li, M. Burnett, T. Dietterich, E. Sullivan, R. Drummond, and J. Herlocker. Toward harnessing user feedback for machine learning. In *International Conference on Intelligent User Interfaces*, pages 82–91, 2007.

[158] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328, 2017.

[159] S. Teso. Toward faithful explanatory active learning with self-explainable neural nets. In *Workshop on Interactive Adaptive Learning*, pages 4–16, 2019.

[160] S. Teso and K. Kersting. Explanatory interactive machine learning. In *Conference on AI, Ethics, and Society*, pages 239–245, 2019.

[161] S. Teso, A. Bontempelli, F. Giunchiglia, and A. Passerini. Interactive label cleaning with example-based explanations. In *International Conference on Neural Information Processing Systems*, 2021.

[162] N. Tintarev and J. Masthoff. Effective explanations of recommendations: User-centered design. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, RecSys '07, page 153–156, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937308. doi:10.1145/1297231.1297259. URL https://doi.org/10.1145/1297231.1297259.

[163] N. Tintarev and J. Masthoff. *Explaining Recommendations: Design and Evaluation*, pages 353–382. Springer, 2015.

[164] B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

[165] K. R. Varshney. Trustworthy machine learning and artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students*, 25(3):26–29, 2019.

[166] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

[167] G. Vilone and L. Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf. Fusion*, 76:89–106, 2021.

[168] J. D. Viviano, B. Simpson, F. Dutil, Y. Bengio, and J. P. Cohen. Saliency is a possible red herring when diagnosing poor generalization. In *International Conference on Learning Representations*, 2021.

[169] L. von Rueden, S. Mayer, K. Beckh, B. Georgiev, S. Giesselbach, R. Heese, B. Kirsch, M. Walczak, J. Pfrommer, A. Pick, R. Ramamurthy, J. Garcke, C. Bauckhage, and J. Schuecker. Informed machine learning - a taxonomy and survey of integrating prior knowledge into learning systems. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[170] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[171] B. Wagner and A. d'Avila Garcez. Neural-symbolic integration for fairness in ai. In *CEUR Workshop*, volume 2846, 2021.

[172] E. Wang, P. Khosravi, and G. Van den Broeck. Towards probabilistic sufficient explanations. In *Extending Explainable AI Beyond Deep Models and Classifiers Workshop at ICML (XXAI)*, 2020.

[173] G. Wang. Humans in the Loop: The Design of Interactive AI Systems, 2019. URL https://hai.stanford.edu/news/humans-loop-design-interactive-ai-systems.

[174] N. Wang, D. V. Pynadath, E. Rovira, M. J. Barnes, and S. G. Hill. Is it my looks? or something i said? the impact of explanations, embodiment, and expectations on trust and performance in human-robot teams. In *International Conference on Persuasive Technology*, pages 56–69, 2018.

[175] N. Wang et al. Trust calibration within a human-robot team: Comparing automatically generated explanations. In *Proc. of HRI*, pages 109–116, 2016.

[176] M. Ware, E. Frank, G. Holmes, M. Hall, and I. H. Witten. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3):281–292, 2001.

[177] A. Waytz, J. Heafner, and N. Epley. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117, 2014.

[178] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE transactions on evolutionary computation*, 1997.

[179] G. Wu, K. Luo, S. Sanner, and H. Soh. Deep language-based critiquing for recommender systems. In *Conference on Recommender Systems*, pages 137–145, 2019.

[180] M. Wu, M. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. In *Conference on Artificial Intelligence*, volume 32, 2018.

[181] M. Wu, S. Parbhoo, M. Hughes, R. Kindle, L. Celi, M. Zazzi, V. Roth, and F. Doshi-Velez. Regional tree regularization for interpretability in deep neural networks. In *Conference on Artificial Intelligence*, volume 34, pages 6413–6421, 2020.

[182] T. Wu, D. S. Weld, and J. Heer. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *Transactions on Computer-Human Interaction*, 26(4):1–27, 2019.

[183] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Annual Meeting of the Association for Computational Linguistics*, 2021.

[184] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang. Causality learning: A new perspective for interpretable machine learning, 2021.

[185] H. Yao, Y. Chen, Q. Ye, X. Jin, and X. Ren. Refining neural networks with compositional explanations. *arXiv preprint arXiv:2103.10415*, 2021.

[186] C. Yeh, B. Kim, and P. Ravikumar. Human-centered concept explanations for neural networks. In H. Pascal and S. M. Kamruzzaman, editors, *Neuro-Symbolic Artificial Intelligence: The State of the Art*, volume 342 of *Frontiers in Artificial Intelligence and Applications*, pages 337–352. IOS Press, 2021.

[187] C.-K. Yeh, J. S. Kim, I. E. Yen, and P. Ravikumar. Representer point selection for explaining deep neural networks. In *Conference on Neural Information Processing Systems*, pages 9311–9321, 2018.

[188] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, and J. Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In *Conference on Neural Information Processing Systems*, pages 1039–1050, 2018.

[189] O. Zaidan, J. Eisner, and C. Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *Conference of the North American Chapter of the Association for Computational Linguistics*, pages 260–267, 2007.

[190] M. Zeni, W. Zhang, E. Bignotti, A. Passerini, and F. Giunchiglia. Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge. *Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1):32, 2019.

[191] Y. Zha, L. Guan, and S. Kambhampati. Learning from ambiguous demonstrations with self-explanation guided reinforcement learning. *arXiv preprint arXiv:2110.05286*, 2021.

[192] J. Zhang and S. P. Curley. Exploring explanation effects on consumers' trust in online recommender agents. *International Journal of Human–Computer Interaction*, 34(5):421–432, 2018.

[193] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. In *AI for Social Good Workshop at ICML'19*, 2019.

[194] Y. Zhang, X. Chen, et al. Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, 14(1):1–101, 2020.

[195] H. Zylberajch, P. Lertvittayakumjorn, and F. Toni. Hildif: Interactive debugging of nli models using influence functions. *Workshop on Interactive Learning for Natural Language Processing*, page 1, 2021.