Explaining Explanations: An Overview of Interpretability of Machine Learning

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139
{lgilpin, davidbau, bzy, abajwa, specter, lkagal}@ mit.edu

***Abstract*—There has recently been a surge of work in explanatory artificial intelligence (XAI). This research area tackles the important problem that complex machines and algorithms often cannot provide insights into their behavior and thought processes. XAI allows users and parts of the internal system to be more transparent, providing explanations of their decisions in some level of detail. These explanations are important to ensure algorithmic fairness, identify potential bias/problems in the training data, and to ensure that the algorithms perform as expected. However, explanations produced by these systems is neither standardized nor systematically assessed. In an effort to create best practices and identify open challenges, we describe foundational concepts of explainability and show how they can be used to classify existing literature. We discuss why current approaches to explanatory methods especially for deep neural networks are insufficient. Finally, based on our survey, we conclude with suggested future research directions for explanatory artificial intelligence.**

I. Introduction

As autonomous machines and black-box algorithms begin making decisions previously entrusted to humans, it becomes necessary for these mechanisms to explain themselves. Despite their success in a broad range of tasks including advertising, movie and book recommendations, and mortgage qualification, there is general mistrust about their results. In 2016, Angwin et al. [1] analyzed Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a widely used criminal risk assessment tool, and found that its predictions were unreliable and racially biased. Along with this, deep neural networks (DNNs) have been shown to be easily fooled into misclassifying inputs with no resemblance to the true category [2]. Extending this observation, a number of techniques have been shown for changing a network's classification of any image to any target class by making imperceptible alterations to the pixels [3], [4], [5]. Adversarial examples are not confined to images; natural language networks can also be fooled [6]. Trojaning attacks have been demonstrated [7] in which inputs remain unchanged, but imperceptible changes are hidden in deep networks to cause them to make targeted errors. While some defense methods have been developed, more attack methods have also emerged [8], [9], [10], [11], and susceptibility to unintuitive errors remains a pervasive problem in DNNs. The potential for such unexpected behavior and unintentional discrimination highlights the need for explanations.

As a first step towards creating explanation mechanisms, there is a new line of research in interpretability, loosely defined as the science of comprehending what a model did (or might have done). Interpretable models and learning methods show great promise; examples include visual cues to find the "focus" of deep neural networks in image recognition and proxy methods to simplify the output of complex systems. However, there is ample room for improvement, since identifying dominant classifiers and simplifying the problem space does not solve all possible problems associated with understanding opaque models.

We take the stance that interpretability alone is insufficient. In order for humans to trust black-box methods, we need *explainability* – models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions. While interpetability is a substantial first step, these mechanisms need to *also* be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited. Although interpretability and explainability have been used interchangeably, we argue there are important reasons to distinguish between them. Explainable models are interpretable by default, but the reverse is not always true.

Some existing deployed systems and regulations make the need for explanatory systems urgent and timely. With impending regulations like the European Union's "Right to Explanation" [12], calls for diversity and inclusion in AI systems [13], findings that some automated systems may reinforce inequality and bias [14], and requirements for safe and secure AI in safety-critical tasks [15], there has been a recent explosion of interest in interpreting the representations and decisions of black-box models. These models are everywhere, and the development of interpretable and explainable models is scattered throughout various disciplines. Examples of general "explainable systems" include interpretable AI, explainable ML, causality, safe AI, computational social science, and automatic scientific discovery. Further, research in explanations and their evaluation are found in machine learning, human computer interaction (HCI), crowd sourcing, machine teaching, AI ethics, technology policy, and many other disciplines. This paper aims to broadly engage the greater machine learning community on the intersection of these topics, to set best practices, to define key concepts, and

to propose evaluation criteria for standardizing explanatory systems often considered in isolation.

In this survey, we present a set of definitions, construct a taxonomy, and present best practices to start to standardize interpretability and explanatory work in AI. We review a number of approaches towards explainable AI systems and provide a taxonomy of how one can think about diverse approaches towards explainability. In Section 2, we define key terms including "explanation", "interpretability", and "explainability". We compare and contrast our definitions with those accepted in the literature. In Section 3, we review some classical AI approaches (e.g., causal modeling, constraint reasoning, intelligent user interfaces, planning) but focus mainly on explainable models for deep learning. We provide a summary of related work papers in Section 4, highlighting differences between definitions of key terms including "explanation", "interpretability", and "explainability". In Section 5, we present a novel taxonomy that examines *what is being explained* by these explanations. We conclude with a discussion addressing open questions and recommend a path to the development and adoption of explainable methods for safety-critical or mission-critical applications.

II. Background and Foundational Concepts

In this section, we provide background information about the key concepts of interpretability and explanability, and describe the meaningful differences between them.

A. What is an Explanation?

Philosophical texts show much debate over what constitutes an explanation. Of particular interest is what makes an explanation "good enough" or what really defines an explanation. Some say a good explanation depends on the question [16]. This set of essays discusses the nature of explanation, theory, and the foundations of linguistics. Although for our work, the most important and interesting work is on "Why questions." In particular, when you can phrase what you want to know from an algorithm as a why question, there is a natural qualitative representation of when you have answered said question–when you can no longer keep asking why. There are two why-questions of interest; why and why-should. Similarly to the explainable planning literature, philosophers wonder about the why-shouldn't and why-should questions, which can give the kinds of explainability requirements we want.

There is also discussion in philosophy about what makes the best explanation. While many say it is inference [17], similar views point to the use of abductive reasoning to explain all the possible outcomes.

B. Interpretability vs. Completeness

An explanation can be evaluated in two ways: according to its *interpretability*, and according to its *completeness*.

The goal of *interpretability* is to describe the internals of a system in a way that is understandable to humans. The success of this goal is tied to the cognition, knowledge, and biases of the user: for a system to be interpretable, it must produce descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user.

The goal of *completeness* is to describe the operation of a system in an accurate way. An explanation is more complete when it allows the behavior of the system to be anticipated in more situations. When explaining a self-contained computer program such as a deep neural network, a perfectly complete explanation can always be given by revealing all the mathematical operations and parameters in the system.

The challenge facing explainable AI is in creating explanations that are both complete and interpretable: it is difficult to achieve interpretability and completeness simultaneously. The most accurate explanations are not easily interpretable to people; and conversely the most interpretable descriptions often do not provide predictive power.

Herman [18] notes that we should be wary of evaluating interpretable systems using merely human evaluations of interpretability, because human evaluations imply a strong and specific bias towards simpler descriptions. He cautions that reliance on human evaluations can lead researchers to create *persuasive* systems rather than transparent systems. He presents the following ethical dilemmas that are a central concern when building interpretable systems:

1) When is it unethical to manipulate an explanation to better persuade users?
2) How do we balance our concerns for transparency and ethics with our desire for interpretability?

We believe that it is fundamentally unethical to present a simplified description of a complex system in order to increase trust if the limitations of the simplified description cannot be understood by users, and worse if the explanation is optimized to hide undesirable attributes of the system. Such explanations are inherently misleading, and may result in the user justifiably making dangerous or unfounded conclusions.

To avoid this trap, explanations should allow a *tradeoff* between interpretability and completeness. Rather than providing only simple descriptions, systems should allow for descriptions with higher detail and completeness at the possible cost of interpretability. Explanation methods should not be evaluated on a single point on this tradeoff, but according to how they behave on the curve from maximum interpretability to maximum completeness.

C. Explainability of Deep Networks

Explanations of the operation of deep networks have focused on either explaining the *processing* of the data by a network, or explaining the *representation* of data inside a network. An explanation of processing answers "Why does this particular input lead to that particular output?" and is analogous to explaining the execution trace of a program. An explanation about representation answers "What information does the network contain?" and can be compared to explaining the internal data structures of a program.

A third approach to interpretability is to create *explanation-producing* systems with architectures that are designed to

simplify interpretation of their own behavior. Such architectures can be designed to make either their processing, representations, or other aspects of their operation easier for people to understand.

III. Review

Due to the growing number of subfields, as well as the policy and legal ramifications [12] of opaque systems, the volume of research in interpretability is quickly expanding. Since it is intractable to review all the papers in the space, we focus on explainable methods in deep neural architectures, and briefly highlight review papers from other subfields.

A. Explanations of Deep Network Processing

Commonly used deep networks derive their decisions using a large number of elementary operations: for example, ResNet [19], a popular architecture for image classification, incorporates about 5×10^7 learned parameters and executes about 10^{10} floating point operations to classify a single image. Thus the fundamental problem facing explanations of such processing is to find ways to reduce the complexity of all these operations. This can be done by either creating a *proxy model* which behaves similarly to the original model, but in a way that is easier to explain, or by creating a *salience map* to highlight a small portion of the computation which is most relevant.

1) Linear Proxy Models: The proxy model approach is exemplified well by the LIME method by Ribeiro [20]. With LIME, a black-box system is explained by probing behavior on perturbations of an input, and then that data is used to construct a local linear model that serves as a simplified proxy for the full model in the neighborhood of the input. Ribeiro shows that the method can be used to identify regions of the input that are most influential for a decision across a variety of types of models and problem domains. Proxy models such as LIME are predictive: the proxy can be run and evaluated according to its faithfulness to the original system. Proxy models can also be measured according to their model complexity, for example, number of nonzero dimensions in a LIME model. Because the proxy model provides a quantifiable relationship between complexity and faithfulness, methods can be benchmarked against each other, making this approach attractive.

2) Decision Trees: Another appealing type of proxy model is the decision tree. Efforts to decompose neural networks into decision trees have recently extended work from the 1990s, which focused on shallow networks, to generalizing the process for deep neural networks. One such method is DeepRED [21], which demonstrates a way of extending the CRED [22] algorithm (designed for shallow networks) to arbitrarily many hidden layers. DeepRED utilizes several strategies to simplify its decision trees: it uses RxREN [23] to prune unnecessary input, and it applies algorithm C4.5 [24], a statistical method for creating a parsimonious decision tree. Although DeepRED is able to construct complete trees that are closely faithful to the original network, the generated trees can be quite large, and the implementation of the method takes substantial time and memory and is therefore limited in scalability.

Another decision tree method is ANN-DT [25] which uses sampling to create a decision tree: the key idea is to use sampling to expand training using a nearest neighbor method.

3) Automatic-Rule Extraction: Automatic rule extraction is another well-studied approach for summarizing decisions. Andrews et al [26] outlines existing rule extraction techniques, and provides a useful taxonomy of five dimensions of rule-extraction methods including their expressive power, translucency and the quality of rules. Another useful survey can be found in the master's thesis by Zilke [27].

Decompositional approaches work on the neuron-level to extract rules to mimic the behavior of individual units. The KT method [28] goes through each neuron, layer-by-layer and applies an if-then rule by finding a threshold. Similar to DeepRED, there is a merging step which creates rules in terms of the inputs rather than the outputs of the preceding layer. This is an exponential approach which is not tangible for deep neural networks. However, a similar approach proposed by Tsukimoto [29] achieves polynomial-time complexity, and may be more tangible. There has also been work on transforming neural network to fuzzy rules [30], by transforming each neuron into an approximate rule.

Pedagogical approaches aim to extract rules by directly mapping inputs to outputs rather than considering the inner workings of a neural network. These treat the network as a black box, and find trends and functions from the inputs to the outputs. Validity Interval Analysis is a type of sensitivity analysis to mimic neural network behavior [31]. This method finds stable intervals, where there is some correlation between the input and the predicted class. Another way to extract rules using sampling methods [32], [33]. Some of these sampling approaches only work on binary input [34] or use genetic algorithms to produce new training examples [35]. Other approaches aim to reverse engineer the neural network, notably, the RxREN algorithm, which is used in DeepRED[21].

Other notable rule-extraction techniques include the MofN algorithm [36], which tries to find rules that explain single neurons by clustering and ignoring insignificant neurons. Similarly, The FERNN [37] algorithm uses the C4.5 algorithm [24] and tries to identify the meaningful hidden neurons and inputs to a particular network.

Although rule-extraction techniques increase the transparency of neural networks, they may not be truly faithful to the model. With that, there are other methods that are focused on creating trust between the user and the model, even if the model is not "sophisicated."

4) Salience Mapping: The salience map approach is exemplified by occlusion procedure by Zeiler [38], where a network is repeatedly tested with portions of the input occluded to create a map showing which parts of the data actually have influence on the network output. When deep network parameters can be inspected directly, a salience map can be created more efficiently by directly computing the input gradient (Simonyan [39]). Since such derivatives can miss important aspects of the information that flows through a network, a number of other approaches have been designed to

propagate quantities other than gradients through the network. Examples are LRP [40], DeepLIFT [41], CAM [42], Grad-CAM [43], Integrated gradients [44], and SmoothGrad [45]. Each technique strikes a balance between showing areas of high network activation, where neurons fire strongest, and areas of high network sensitivity, where changes would most affect the output. A comparison of some of these methods can be found in Ancona [46].

B. Explanations of Deep Network Representations

While the number of individual operations in a network is vast, deep networks are internally organized into a smaller number of subcomponents: for example, the billions of operations of ResNet are organized into about 100 layers, each computing between 64 and 2048 channels of information per pixel. The explanation of deep network representations aims to understand the role and structure of the data flowing through these bottlenecks. This work can be divided by the granularity examined: representations can be understood *by layer*, where all the information flowing through a layer is considered together, and *by unit*, where single neurons or single filter channels are considered individually, and *by vector*, where other vector directions in representation space are considered individually.

1) Role of Layers: Layers can be understood by testing their ability to help solve different problems from the problems the network was originally trained on. For example Razavian [47] found that the output of an internal layer of a network trained to classify images of objects in the ImageNet dataset produced a feature vector that could be directly reused to solve a number of other difficult image processing problems including fine-grained classification of different species of birds, classification of scene images, attribute detection, and object localization. In each case, a simple model such as an SVM was able to directly apply the deep representation to the target problem, beating state-of-the-art performance without training a whole new deep network. This method of using a layer from one network to solve a new problem is called *transfer learning*, and it is of immense practical importance, allowing many new problems to be solved without developing new datasets and networks for each new problem. Yosinksi [48] described a framework for quantifying transfer learning capabilities in other contexts.

2) Role of Individual Units: The information within a layer can be further subdivided into individual neurons or individual convolutional filters. The role of such individual units can be understood qualitatively, by creating visualizations of the input patterns that maximize the response of a single unit, or quantitatively, by testing the ability of a unit to solve a transfer problem. Visualizations can be created by optimizing an input image using gradient descent [39], by sampling images that maximize activation [49], or by training a generative network to create such images [50]. Units can also be characterized quantitatively by testing their ability to solve a task. One example of a such a method is network dissection [51], which measures the ability of individual units solve a segmentation problem over a broad set of labeled visual concepts. By quantifying the ability of individual units to locate emergent concepts such as objects, parts, textures, and colors that are not explicit in the original training set, network dissection can be used characterize the kind of information represented by visual networks at each unit of a network.

A review of explanatory methods focused on understanding unit representations used by visual CNNs can be found in [52], which examines methods for visualization of CNN representations in intermediate network layers, diagnosis of these representations, disentanglement representation units, the creation of explainable models, and semantic middle-to-end learning via human-computer interaction.

Pruning of networks [53] has also been shown to be a step towards understanding the role of individual neurons in networks. In particular, large networks that train successfully contain small subnetworks with initializations conducive to optimization. This demonstrates that there exist training strategies that make it possible to solve the same problems with much smaller networks that may be more interpretable.

3) Role of Representation Vectors: Closely related to the approach of characterizing individual units is characterizing other directions in the representation vector space formed by linear combinations of individual units. Concept Activation Vectors (CAVs) [54] are a framework for interpretation of a neural nets representations by identifying and probing directions that align with human-interpretable concepts.

C. Explanation-Producing Systems

Several different approaches can be taken to create networks that are designed to be easier to explain: networks can be trained to use *explicit attention* as part of their architecture; they can be trained to learn *disentangled representations*; or they can be directly trained to create *generative explanations*.

1) Attention Networks: Attention-based networks learn functions that provide a weighting over inputs or internal features to steer the information visible to other parts of a network. Attention-based approaches have shown remarkable success in solving problems such as allowing natural language translation models to process words in an appropriate non-sequential order [55], and they have also been applied in domains such as fine-grained image classification [56] and visual question answering [57]. Although units that control attention are not trained for the purpose of creating human-readable explanations, they do directly reveal a map of which information passes through the network, which can serve as a form of explanation. Datasets of human attention have been created [58], [59]; these allow systems to be evaluated according to how closely and their internal attention resembles human attention.

While attention can be observed as a way of extracting explanations, another interesting approach is to train attention explicitly in order to create a network that has behavior that conforms to desired explanations. This is the technique proposed by Ross [60], where input sensitivity of a network is adjusted and measured in order to create networks that are

"right for the right reasons;" the method can be used to steer the internal reasoning learned by a network. They also propose that the method can be used to learn a sequence of models that discover new ways to solve a problem that may not have been discovered by previous instances.

2) Disentangled Representations: Disentangled representations have individual dimensions that describe meaningful and independent factors of variation. The problem of separating latent factors is an old problem that has previously been attacked using a variety of techniques such as Principal Component Analysis [61], Independent Component Analysis [62], and Nonnegative Matrix Factorization [63]. Deep networks can be trained to explicitly learn disentangled representations. One approach that shows promise is Variational Autoencoding [64], which trains a network to optimize a model to match the input probability distribution according to information-theoretic measures. Beta-VAE [65] is a tuning of the method that has been observed to disentangle factors remarkably well. Another approach is InfoGAN [66], which trains generative adversarial networks with an objective that reduces entanglement between latent factors. Special loss functions have been suggested for encouraging feed-forward networks to also disentangle their units; this can be used to create interpretable CNNs that have individual units that detect coherent meaningful patches instead of difficult-to-interpret mixtures of patterns [67]. Disentangled units can enable the construction of graphs [68] and decision trees [69] to elucidate the reasoning of a network. Architectural alternatives such as capsule networks [70] can also organize the information in a network into pieces that disentangle and represent higher-level concepts.

3) Generated Explanations: Finally, deep networks can also be designed to generate their own human-understandable explanations as part of the explicit training of the system. Explanation generation has been demonstrated as part of systems for visual question answering [71] as well as in fine-grained image classification [72]. In addition to solving their primary task, these systems synthesize a "because" sentence that explains the decision in natural language. The generators for these explanations are trained on large data sets of human-written explanations, and they explain decisions using language that a person would use.

Multimodal explanations that incorporate both visual pointing and textual explanations can be generated; this is the approach taken in [59]. This system builds upon the winner of the 2016 VQA challenge [73], with several simplification and additions. In addition to the question answering task and the internal attention map, the system trains an additional long-form explanation generator together with a second attention map optimized as a visual pointing explanation. Both visual and textual explanations score well individually and together on evaluations of user trust and explanation quality. Interestingly, the generation of these highly readable explanations is conditioned on the output of the network: the explanations are generated based on the decision, after the decision of the network has already been made.

IV. Related Work

We provide a summary of related review papers, and an overview of interpretability and explainability in other domains.

A. Interpretability

A previous survey has attempted to define taxonomies and best practices for a "strong science" of interpretability [74]. The motivation of this paper is similar to ours, noting that "the volume of research on interpretability is rapidly growing" and that there is no clear existing definition or evaluation criteria for interpretability. The authors define interpretability as "the ability to explain or to present in understandable terms to a human" and suggest a variety of definitions for explainability, converging on the notion that interpretation is the act of discovering the evaluations of an explanation. The authors attempt to reach consensus on the definition of interpretable machine learning and how it should be measured. While we are inspired by the taxonomy of this paper, we focus on the explainability aspect rather than interpretability.

The main contribution of this paper is a taxonomy of modes for interpretability evaluations: application-grounded, human-grounded, and functionally grounded. The authors state interpretability is required when a problem formulation is incomplete, when the optimization problem – the key definition to solve the majority of machine learning problems – is disconnected from evaluation. Since their problem statement is the incompleteness criteria of models, resulting in a disconnect between the user and the optimization problem, evaluation approaches are key.

The first evaluation approach is application-grounded, involving real humans on real tasks. This evaluation measures how well human-generated explanations can aid other humans in particular tasks, with explanation quality assessed in the true context of the explanation's end tasks. For instance, a doctor should evaluate diagnosis systems in medicine.

The second evaluation approach is human-grounded, using human evaluation metrics on simplified tasks. The key motivation is the difficulty of finding target communities for application testing. Human-grounded approaches may also be used when specific end-goals, such as identifying errors in safety-critical tasks, are not possible to realize fully.

The final evaluation metric is functionally grounded evaluation, without human subjects. In this experimental setup, proxy or simplified tasks are used to prove some formal definition of interpretability. The authors acknowledge that choosing which proxy to use is a challenge inherent to this approach. There lies a delicate tradeoff between choosing an interpretable model and a less interpretable proxy method which is more representative of model behavior; the authors acknowledge this point and briefly mention decision trees as a highly interpretable model.

The authors then discuss open problems, best practices and future work in interpretability research, while heavily encouraging data-driven approaches for discovery in interpretability. Although the contribution of the interpretability definition,

we distinguish our taxonomy by defining different focuses of explanations a model can provide, and how those explanations should be evaluated.

B. Explainable AI for HCI

One previous review paper of explainable AI performed a sizable data-driven literature analysis of explainable systems [75]. In this work, the authors move beyond the classical AI interpretability argument, focusing instead on how to create practical systems with efficacy for real users. The authors motivate AI systems that are "explainable by design" and present their findings with three contributions: a data-driven network analysis of 289 core papers and 12,412 citing papers for an overview of explainable AI research, a perspective on trends using network analysis, and a proposal for best practices and future work in HCI research pertaining to explainablity.

Since most of the paper focuses on the literature analysis, the authors highlight only three large areas in their related work section: explainable artificial intelligence (XAI), intelligibility and interpretability in HCI, and analysis methods for trends in research topics.

The major contribution of this paper is a sizable literature analysis of explainable research, enabled by the citation network the authors constructed. Papers were aggregated based on a keyword search on variations of the terms "intelligible," "interpretable," "transparency," "glass box," "black box," "scrutable," "counterfacutals," and "explainable," and then pruned down to 289 core papers and 12,412 citing papers. Using network analysis, the authors identified 28 significant clusters and 9 distinct research communities, including early artificial intelligence, intelligent systems/agents/user interfaces, ambient intelligence, interaction design and learnability, interpretable ML and classifier explainers, algorithmic fairness/accountability/transparency/policy/journalism, causality, psychological theories of explanations, and cognitive tutors. In contrast, our work is focused on the research in interpretable ML and classifier explainers for deep learning.

With the same sets of core and citing papers, the authors performed LDA-based topic modeling on the abstract text to determine which communities are related. The authors found the largest, most central and well-studied network to be intelligence and ambient systems. In our research, the most important subnetworks are the Explainable AI: Fair, Accountable, and Transparent (FAT) algorithms and Interpretable Machine Learning (iML) subnetwork and the theories of explanations subnetworks. In particular, the authors provide a distinction between FATML and interpretability; while FATML is focused on societal issues, interpretability is focused on methods. Theory of explanations joins causality and cognitive psychology with the common threads of counterfactual reasoning and causal explanations. Both these threads are important factors in our taxonomy analysis.

In the final section of their paper, the authors name two trends of particular interest to us: ML production rules and a road map to rigorous and usable intelligibility. The authors note a lack of classical AI methods being applied to interpretability, encouraging broader application of those methods to current research. Though this paper focused mainly on setting an HCI research agenda in explainability, it raises many points relevant to our work. Notably, the literature analysis discovered subtopics and subdisciplines in psychology and social science, not yet identified as related in our analysis.

C. Explanations for Black-Box Models

A recent survey on methods for explaining black-box models [76] outlined a taxonomy to provide classifications of the main problems with opaque algorithms. Most of the methods surveyed are applied to neural-network based algorithms, and therefore related to our work.

The authors provide an overview of methods that explaining decision systems based on opaque and obscure machine learning models. Their taxonomy is detailed, distinguishing small differing components in explanation approaches (e.g. Decision tree vs. single tree, neuron activation, SVM, etc.) Their classification examines four features for each explanation method:

1) The type of the problem faced.
2) The explanatory capability used to open the black box.
3) The type of black box model that can be explained.
4) The type of input data provided to the black box model.

They primarily divide the explanation methods according to the types of problem faced, and identify four groups of explanation methods: methods to explain black box models; methods to explain black box outcomes; methods to inspect black boxes; and methods to design transparent boxes. Using their classification features and these problem definitions, they discuss and further categorize methods according to the type of explanatory capability adopted, the black box model "opened", and the input data. Their goal is to review and classify the main black box explanation architectures, so their classifications can serve as a guide to identifying similar problems and approaches. We find this work a meaningful contribution that is useful for exploring the design space of explanation methods. Our classification is less finely-divided; rather than subdividing implementation techniques, we examine the focus of the explanatory capability and what each approach *can* explain, with an emphasis on understanding how different types of explainability methods can be evaluated.

D. Explainability in Other Domains

Explainable planning [77] is an emerging discipline that exploits the model-based representations that exist in the planning community. Some of the key ideas were proposed years ago in plan recognition [78]. Explainable planning urges the familiar and common basis for communication with users, while acknowledging the gap between planning algorithms and human problem-solving. In this paper, the authors outline and provide examples of a number of different types of questions that explanations could answer, like "Why did you do A" or "Why DIDN'T you do B", "Why CAN'T you do C", etc. In addition, the authors emphasize that articulating a plan in natural language is NOT usually the same thing as explaining

the plan. A request for explanation is "an attempt to uncover a piece of knowledge that the questioner believes must be available to the system and that the questioner does not have". We discuss the questions an explanation can and should answer in our conclusion.

Automatic explanation generation is also closely related to computers and machines that can tell stories. In John Reeves' thesis [79], he created the THUNDER program to read stories, construct character summaries, infer beliefs, and understand conflict and resolution. Other work examines how to represent the necessary structures to do story understanding [80]. The Genesis Story-Understanding System [81] is a working system that understands, uses, and composes stories using higher-level concept patterns and commonsense rules. Explanation rules are used to supply missing causal or logical connections.

At the intersection of human robot interaction and story-telling is verbalization; generating explanations for human-robot interaction [82]. Similar approaches are found in abductive reasoning; using a case-based model [83] or explanatory coherence [84]. This is also a well-studied field in brain and cognitive science by filling in the gaps of knowledge by imagining new ideas [85] or using statistical approaches [86].

V. Taxonomy

The approaches from the literature that we have examined fall into three different categories. Some papers propose explanations that, while admittedly non-representative of the underlying decision processes, provide some degree of *justification* for emitted choices that may be used as response to demands for explanation in order to build human trust in the system's accuracy and reasonableness. These systems *emulate* the *processing* of the data to draw connections between the inputs and outputs of the system.

The second purpose of an explanation is to explain the *representation* of data inside the network. These provide insight about the internal operation of the network and can be used to facilitate explanations or interpretations of activation data within a network. This is comparative to explaining the internal data structures of the program, to start to gain insights about why certain intermediate representations provide information that enables specific choices.

The final type of explanation is *explanation-producing* networks. These networks are specifically built to explain themselves, and they are designed to simplify the interpretation of an opaque subsystem. They are steps towards improving the transparency of these subsystems; where processing, representations, or other parts are justified and easier to understand.

The taxonomy we present is useful given the broad set of existing approaches for achieving varying degrees of interpretability and completeness in machine learning systems. Two distinct methods claiming to address the same overall problem may, in fact, be answering very different questions. Our taxonomy attempts to subdivide the problem space, based on existing approaches, to more precisely categorize what has already been accomplished.

We show the classifications of our reviewed methods per category in Table I. Notice that the processing and explanation-producing roles are much more populated than the representation role. We believe that this disparity is largely due to the fact that it is difficult to evaluate representation-based models. User-study evaluations are not always appropriate. Other numerical methods, like demonstrating better performance by adding or removing representations, are difficult to facilitate.

The position of our taxonomy is to promote research and evaluation across categories. Instead of other explanatory and interpretability taxonomies that assess the purpose of explanations [74] and their connection to the user [75], we instead assess the focus on the method, whether the method tries to explain the *processing* of the data by a network, explain the *representation* of data inside a network or to be a self-explaining architecture to gain additional meta predictions and insights about the method.

We promote this taxonomy, particularly the explanation-producing sub-category, as a way to consider designing neural network architectures and systems. We also highlight the lack of standardized evaluation metrics, and propose research crossing areas of the taxonomy as future research directions.

Processing	**Representation**	**Explanation Producing**
Proxy Methods	Role of layers	Scripted conversations
Decision Trees	Role of neurons	Attention-based
Salience mapping	Role of vectors	Disentangled rep.
Automatic-rule extraction		Human evaluation

TABLE I
THE CLASSIFICATIONS OF TOP LEVEL METHODS INTO OUR TAXONOMY.

VI. Evaluation

Although we outline three different focuses of explanations for deep networks, they do not share the same evaluation criteria. Most of the work surveyed conducts one of the following types of evaluation of their explanations.

1) Completeness compared to the original model. A proxy model can be evaluated directly according to how closely it approximates the original model being explained.
2) Completeness as measured on a substitute task. Some explanations do not directly explain a model's decisions, but rather some other attribute that can be evaluated. For example, a salience explanation that is intended to reveal model sensitivity can be evaluated against a brute-force measurement of the model sensitivity.
3) Ability to detect models with biases. An explanation that reveals sensitivity to a specific phenomenon (such as a presence of a specific pattern in the input) can be tested for its ability to reveal models with the presence or absence of a relevant bias (such as reliance or ignorance of the specific pattern).
4) Human evaluation. Humans can evaluate explanations for reasonableness, that is how well an explanation matches human expectations. Human evaluation can also

evaluate completeness or substitute-task completeness from the point of view of enabling a person to predict behavior of the original model; or according to helpfulness in revealing model biases to a person.

As we can see in Table II, the tradeoff between *interpretability* and its *completeness* can be seen not only as a balance between simplicity and accuracy in a proxy model. The tradeoff can also be made by anchoring explanations to substitute tasks or evaluating explanations in terms of their ability to surface important model biases. Each of the three types of explanation methods can provide explanations that can be evaluated for completeness (on those critical model characteristics), while still being easier to interpret than a full accounting for every detailed decision of the model.

Processing	**Representation**	**Explanation Producing**
Completeness to Model Completeness on substitute task	Completeness on substitute task Detect biases	Human evaluation Detect biases

TABLE II
SUGGESTED EVALUATIONS FOR THE CLASSIFICATIONS IN OUR TAXONOMY

A. Processing

Processing models can also be regarded as emulation-based methods. Proxy methods should be evaluated on their faithfulness to the original model. A handful of these metrics are described in [20]. The key idea is that evaluating completeness to a model should be local. Even if a model, in our case, a deep neural network, is too complex globally, you can still explain in a way that makes sense locally by approximating local behavior. Therefore, processing model explanations want to minimize the "complexity" of explanations (essentially, minimize length) as well as "local completeness" (error of interpretable representation relative to actual classifier, near instance being explained).

Salience methods that highlight sensitive regions for processing are often evaluated qualitatively. Although they do not directly predict the output of the original method, these methods can also be evaluated for faithfulness, since their intent is to explain model sensitivity. For example, [46] conducts an occlusion experiment as ground truth, in the model is tested on many version of an input image where each portion of the image is occluded. This test determines in a brute-force but computationally inefficient way which parts of an input cause a model to change its outputs the most. Then each salience method can be evaluated according to how closely the method produces salience maps that correlate with this occlusion-based sensitivity.

B. Representation

Representation-based methods typically characterize the role of portions of the representation by testing the representations on a transfer task. For example, representation layers are characterized according to their ability to serve as feature input for a transfer problem, and both Network Dissection representation units and Concept Activation Vectors are measured according to their ability to detect or correlate with specific human-understandable concepts.

Once individual portions of a representation are characterized, they can be tested for explanatory power by evaluating whether their activations can faithfully reveal a specific bias in a network. For example, Concept Activation Vectors [54] are evaluated by training several versions of the same network on datasets that are synthesized to contain two different types of signals that can be used to determine the class (the image itself, and an overlaid piece of text which gives the class name with varying reliability). The faithfulness of CAVs to the network behavior can be verified by evaluating whether classifiers that are known to depend on the text (as evidenced by performance on synthesized tests) exhibit high activations of CAV vectors corresponding to the text, and that classifiers that do not depend on the text exhibits low CAV vectors.

C. Explanation-Producing

Explanation-producing systems can be evaluated according to how well they match user expectations. For example, network attention can be compared to human attention [58], and disentangled representations can be tested on synthetic datasets that have known latent variables, to determine whether those variables are recovered. Finally, systems that are trained explicitly to generate human-readable explanations can be tested by similarity to test sets, or by human evaluation.

One of the difficulties of evaluating explanatory power of explanation-producing systems is that, since the system itself produces the explanation, evaluations necessarily couple evaluation of the system along with evaluation of the explanation. An explanation that seems unreasonable could indicate either a failure of the system to process information in a reasonable way, or it could indicate the failure of the explanation generator to create a reasonable description. Conversely, an explanation system that is not faithful to the decisionmaking process could produce a reasonable description even if the underlying system is using unreasonable rules to make the decision. An evaluation of explanations based on their reasonableness alone can miss these distinctions. In [74], a number of user-study designs are outlined that can help bridge the gap between the model and the user.

VII. CONCLUSIONS

One common viewpoint in the deep neural network community is that the level of interpretability and theoretical understanding needed to for transparent explanations of large DNNs remains out of reach; for example, as a response to Ali Rahimi's Test of Time NIPS address, Yann LeCunn responded that "The engineering artifacts have almost always preceded the theoretical understanding" [87]. However, we assert that, for machine learning systems to achieve wider acceptance among a skeptical populace, it is crucial that such systems be able to provide or permit satisfactory explanations of their decisions. The progress made so far has been promising, with efforts in explanation of deep network processing, explanation

of deep network representation, and system-level explanation production yielding encouraging results.

We find, though, that the various approaches taken to address different facets of explainability are siloed. Work in the explainability space tends to advance a particular category of technique, with comparatively little attention given to approaches that merge different categories of techniques to achieve more effective explanation. Given the purpose and type of explanation, it is not obvious what the best type of explanation metric is and should be. We encourage the use of diverse metrics that align with the purpose and completeness of the targeted explanation. Our view is that, as the community learns to advance its work collaboratively by combining ideas from different fields, the overall state of system explanation will improve dramatically, resulting in methods that provide behavioral extrapolation, build trust in deep learning systems, and provide usable insight into deep network operation enabling system behavior understanding and improvement.

ACKNOWLEDGEMENTS

The work was partially funded by DARPA XAI program FA8750-18-C0004, the National Science Foundation under Grants No. 1524817, the MIT-IBM Watson AI Lab, and the Toyota Research Institute (TRI). The authors also wish to express their appreciation for Jonathan Frankle for sharing his insightful feedback on earlier versions of the manuscript.

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Accessed May 24, 2019], 2016.

[2] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 427–436.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[4] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 86–94.

[5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387.

[6] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," *CoRR*, vol. abs/1707.07328, 2017. [Online]. Available: http://arxiv.org/abs/1707.07328

[7] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," 2017.

[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2018.

[10] A. D. Kraft, "Vision by alignment," Ph.D. dissertation, MIT, 77 Massachusetts Ave., 2 2018, an optional note.

[11] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 3–14.

[12] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a" right to explanation"," *arXiv preprint arXiv:1606.08813*, 2016.

[13] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Information, communication & society*, vol. 15, no. 5, pp. 662–679, 2012.

[14] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[15] S. Russell, D. Dewey, and M. Tegmark, "Research priorities for robust and beneficial artificial intelligence," *Ai Magazine*, vol. 36, no. 4, pp. 105–114, 2015.

[16] S. Bromberger, *On what we know we don't know: Explanation, theory, linguistics, and how questions shape them*. University of Chicago Press, 1992.

[17] P. R. Thagard, "The best explanation: Criteria for theory choice," *The journal of philosophy*, vol. 75, no. 2, pp. 76–92, 1978.

[18] B. Herman, "The promise and peril of human evaluation for model interpretability," *arXiv preprint arXiv:1711.07414*, 2017.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[21] J. R. Zilke, E. L. Mencía, and F. Janssen, "Deepred–rule extraction from deep neural networks," in *International Conference on Discovery Science*. Springer, 2016, pp. 457–473.

[22] M. Sato and H. Tsukimoto, "Rule extraction from neural networks via decision tree induction," in *Neural Networks, 2001. Proceedings. IJCNN'01. International Joint Conference on*, vol. 3. IEEE, 2001, pp. 1870–1875.

[23] M. G. Augasta and T. Kathirvalavakumar, "Reverse engineering the neural networks for rule extraction in classification problems," *Neural processing letters*, vol. 35, no. 2, pp. 131–150, 2012.

[24] S. L. Salzberg, "C4. 5: Programs for machine learning by j. ross quinlan. morgan kaufmann publishers, inc., 1993," *Machine Learning*, vol. 16, no. 3, pp. 235–240, 1994.

[25] G. P. Schmitz, C. Aldrich, and F. S. Gouws, "Ann-dt: an algorithm for extraction of decision trees from artificial neural networks," *IEEE Transactions on Neural Networks*, vol. 10, no. 6, pp. 1392–1401, 1999.

[26] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[27] J. R. Zilke, "Extracting Rules from Deep Neural Networks," Master's thesis, Technische Universitat Darmstadt, 2016.

[28] L. Fu, "Rule generation from neural networks," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 24, no. 8, pp. 1114–1124, 1994.

[29] H. Tsukimoto, "Extracting rules from trained neural networks," *IEEE Transactions on Neural Networks*, vol. 11, no. 2, pp. 377–389, 2000.

[30] J. M. Benítez, J. L. Castro, and I. Requena, "Are artificial neural networks black boxes?" *IEEE Transactions on neural networks*, vol. 8, no. 5, pp. 1156–1164, 1997.

[31] S. Thrun, "Extracting rules from artificial neural networks with distributed representations," in *Advances in neural information processing systems*, 1995, pp. 505–512.

[32] U. Johansson, R. Konig, and L. Niklasson, "Automatically balancing accuracy and comprehensibility in predictive modeling," in *Information Fusion, 2005 8th International Conference on*, vol. 2. IEEE, 2005, pp. 7–pp.

[33] M. W. Craven, "Extracting comprehensible models from trained neural networks," Ph.D. dissertation, University of Wisconsin, Madison, 1996.

[34] I. A. Taha and J. Ghosh, "Symbolic interpretation of artificial neural networks," *IEEE Transactions on knowledge and data engineering*, vol. 11, no. 3, pp. 448–463, 1999.

[35] T. Hailesilassie, "Rule extraction algorithm for deep neural networks: A review," *arXiv preprint arXiv:1610.05267*, 2016.

[36] G. G. Towell and J. W. Shavlik, "Extracting refined rules from knowledge-based neural networks," *Machine learning*, vol. 13, no. 1, pp. 71–101, 1993.

[37] R. Setiono and W. K. Leow, "Fernn: An algorithm for fast extraction of rules from neural networks," *Applied Intelligence*, vol. 12, no. 1-2, pp. 15–25, 2000.

[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[39] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[40] 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*, vol. 10, no. 7, p. e0130140, 2015.

[41] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

[42] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 2016, pp. 2921–2929.

[43] 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," *See https://arxiv. org/abs/1610.02391 v3*, vol. 7, no. 8, 2016.

[44] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," *arXiv preprint arXiv:1703.01365*, 2017.

[45] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017. [Online]. Available: http://arxiv.org/abs/1706.03825

[46] M. Ancona, E. Ceolini, A. C. Öztireli, and M. H. Gross, "A unified view of gradient-based attribution methods for deep neural networks," *CoRR*, vol. abs/1711.06104, 2017. [Online]. Available: http://arxiv.org/abs/1711.06104

[47] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014, pp. 512–519.

[48] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Object detectors emerge in deep scene cnns," *arXiv preprint arXiv:1412.6856*, 2014.

[50] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune, "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks," in *Advances in Neural Information Processing Systems*, 2016, pp. 3387–3395.

[51] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, "Network dissection: Quantifying interpretability of deep visual representations," in *Computer Vision and Pattern Recognition*, 2017.

[52] Q.-s. Zhang and S.-C. Zhu, "Visual interpretability for deep learning: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 19, no. 1, pp. 27–39, 2018.

[53] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Training pruned neural networks," *CoRR*, vol. abs/1803.03635, 2018. [Online]. Available: http://arxiv.org/abs/1803.03635

[54] B. Kim, J. Gilmer, F. Viegas, U. Erlingsson, and M. Wattenberg, "Tcav: Relative concept importance testing with linear concept activation vectors," *arXiv preprint arXiv:1711.11279*, 2017.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 6000–6010.

[56] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 842–850.

[57] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *Advances In Neural Information Processing Systems*, 2016, pp. 289–297.

[58] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" *Computer Vision and Image Understanding*, vol. 163, pp. 90–100, 2017.

[59] D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, "Multimodal explanations: Justifying decisions and pointing to the evidence," *CoRR*, vol. abs/1802.08129, 2018. [Online]. Available: http://arxiv.org/abs/1802.08129

[60] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," *arXiv preprint arXiv:1703.03717*, 2017.

[61] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*. Springer, 1986, pp. 115–128.

[62] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[63] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational statistics & data analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[64] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[65] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," 2016.

[66] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2016, pp. 2172–2180.

[67] Q. Zhang, Y. N. Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," 2018.

[68] Q. Zhang, R. Cao, Y. N. Wu, and S.-C. Zhu, "Growing interpretable part graphs on convnets via multi-shot learning." 2017.

[69] Q. Zhang, Y. Yang, Y. Liu, Y. N. Wu, and S.-C. Zhu, "Unsupervised learning of neural networks to explain neural networks," *arXiv preprint arXiv:1805.07468*, 2018.

[70] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, 2017, pp. 3859–3869.

[71] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[72] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.

[73] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *arXiv preprint arXiv:1606.01847*, 2016.

[74] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv*, 2017. [Online]. Available: https://arxiv.org/abs/1702.08608

[75] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 582.

[76] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," *arXiv preprint arXiv:1802.01933*, 2018.

[77] M. Fox, D. Long, and D. Magazzeni, "Explainable planning," *CoRR*, vol. abs/1709.10256, 2017. [Online]. Available: http://arxiv.org/abs/1709.10256

[78] H. A. Kautz, "Generalized plan recognition."

[79] J. F. Reeves, "Computational morality: A process model of belief conflict and resolution for story understanding," 1991.

[80] E. T. Mueller, "Story understanding," *Encyclopedia of Cognitive Science*.

[81] P. Winston and D. Holmes, "The genesis manifesto: Story understanding and human intelligence," 2017.

[82] S. Rosenthal, S. P. Selvaraj, and M. M. Veloso, "Verbalization: Narration of autonomous robot experience."

[83] D. B. Leake, "Focusing construction and selection of abductive hypotheses," 1993.

[84] H. T. Ng and R. J. Mooney, "The role of coherence in constructing and evaluating abductive explanations," in *Working Notes, AAAI Spring Symposium on Automated Abduction, Stanford, California*, 1990.

[85] R. W. Magid, M. Sheskin, and L. E. Schulz, "Imagination and the generation of new ideas," *Cognitive Development*, vol. 34, pp. 99–110, 2015.

[86] J. Koster-Hale and R. Saxe, "Theory of mind: a neural prediction problem," *Neuron*, vol. 79, no. 5, pp. 836–848, 2013.

[87] Y. LeCun, "My take on ali rahimi's test of time award talk at nips," 2017.