Interpretability Needs a New Paradigm

Anonymous authors
Paper under double-blind review

Abstract

Interpretability is the study of explaining models in understandable terms to humans. At present, interpretability is divided into two paradigms: the intrinsic paradigm, which believes that only models designed to be explained can be explained, and the post-hoc paradigm, which believes that black-box models can be explained. At the core of this debate is how each paradigm ensures its explanations are *faithful*, i.e., true to the model's behavior. This is important, as false but convincing explanations lead to unsupported confidence in artificial intelligence (AI), which can be dangerous. This paper's position is that we should think about new paradigms while staying vigilant regarding faithfulness. First, by examining the history of paradigms in science, we see that paradigms are constantly evolving. Then, by examining the current paradigms, we can understand their underlying beliefs, the value they bring, and their limitations. Finally, this paper presents 3 emerging paradigms for interpretability. The first paradigm designs models such that faithfulness can be easily measured. Another optimizes models such that explanations become faithful. The last paradigm proposes to develop models that produce both a prediction and an explanation.

1 Introduction

In 1874, Georg Cantor proposed set theory and showed there exists at least two kinds of infinity. This divided the mathematical field. The Intuitionists, who named Cantor's theory nonsense, thought that math was a pure creation of the mind and that these infinities weren't real. Henri Poincaré said: "Later generations will regard Mengenlehre (set theory) as a disease from which one has recovered" (Gray, 1991). Leopold Kronecker called Cantor a "scientific charlatan" and "corruptor of the youth" (Dauben, 1977).

The other group, the Formalists, thought that by using Cantor's set theory, all math could be proven from this fundamental foundation. David Hilbert said: "No one shall expel us from the paradise that Candor has created" (Hilbert, 1926) and "In opposition to the foolish Ignoramus (we will not know; i.e., intuitionists), our slogan shall be: We must know – we will know" (Hilbert, 1930; Reidemeister, 1971; Smith, 2014).

Today, we know infinities are important concepts; thus, the Intuitionists were wrong. However, Kurt Gödel showed that the Formalists were also wrong. Unfortunately, there exist true statements which can never be proven (Gödel, 1931, Gödel's incompleteness theorem).

There are many examples in science and mathematics where there have been strong debates and beliefs due to conflicting paradigms. Science historian Thomas Kuhn defines a scientific paradigm as: "universally recognized scientific achievements that, for a time, provide model problems and solutions to a community of practitioners" (Kuhn, 1996).

Time and time again, when there are conflicting paradigms, it is only "for a time". Eventually, we find neither paradigm is true, or both paradigms are true (under a more nuanced understanding). In retrospect, it is more constructive to develop an understanding as to which paradigms may be right under what conditions, as opposed to an all-or-nothing approach of arguing about a singular right paradigm. Alternatively, we could come up with a new paradigm, a new school of thought, a new direction; which replaces or bridges the old way of thinking.

In this paper, we re-examine the current direction and paradigms of interpretability and invite the reader to consider whether it is time for a new paradigm.

1.1 Interpretability and faithfulness

Interpretability is the ability to explain a model in understandable terms to humans (Doshi-Velez & Kim, 2017). Model explanations have particularly become important for AI safety, as machine learning is increasingly being used by the industry and affects the lives of most humans. This and additional motivations are elaborated on in Section 2.

Within interpretability, there currently exist two paradigms, called *post-hoc* and *intrinsic* (Lipton, 2018). Section 3 properly describe their stance. Put briefly, the *intrinsic* paradigm believes that only models designed to be explained can be explained (Rudin, 2019). In contrast, the *post-hoc* paradigms believe this constraint is unnecessary and too restrictive to achieve competitive performance (Madsen et al., 2022b).

The position in this paper is that, while both paradigms have yielded some insights on specific domains, their broader impact has been limited because their underlying beliefs are problematic and we should therefore look for new directions. Section 4 contains the primary support for this position. To prove that other paradigms can exist, Section 5 then presents three emerging paradigms for interpretability and discusses how they might overcome past challenges, their beliefs, drawbacks, and future directions. Importantly, while this demonstrates that developing new paradigms is possible, Section 5 should not be considered a final list of new directions.

At the core of this discussion is how each paradigm approaches *faithfulness*. A faithful explanation means the explanation accurately reflects the model's logic, and ensuring and validating this often presents a major challenge because the model's logic is inaccessible to humans (Jacovi & Goldberg, 2020). Faithfulness is particularly important, as false but convincing explanations can lead to unsupported confidence in models, increasing the risk of AI.

In addition to faithfulness is *comprehensibility*, another equally important desirable (Doshi-Velez & Kim, 2017), measuring how understandable an explanation is to humans (also known as human-groundedness) (Robnik-Šikonja & Bohanec, 2018; Lipton, 2018). However, this position paper focuses primarily on faithfulness, as the paradigms are rooted in faithfulness as discussed in Section 3.2.1, and the issue of comprehensibility often first materializes when considering a specific explanation, this is discussed more in Section 6. However, comments on comprehensibility are made when appropiate throughout the paper.

For these reasons, emerging paradigms (Section 4) attempt to bring new perspectives regarding how to achieve faithfulness. This creates a new opportunity to do interpretability research centered around ensuring faithfulness. However, it also creates a new risk as we may take faithfulness for granted once again, as has been the case with both the intrinsic (Jacovi & Goldberg, 2020) and post-hoc paradigms (Madsen et al., 2022b). To prevent this, this paper also takes the position that we should be vigilant about faithfulness when it comes to new paradigms to prevent repeating past mistakes.

2 Why interpretability is needed

Before discussing the current paradigms and their shortcomings, it's necessary to first consider if interpretability is needed at all. Many ethical motivations for interpretability are also served by bias and fairness metrics, so if the current paradigms of interpretability do not work (as we argue in Section 4), perhaps we should drop the idea of interpretability completely. If the models can be made accurate, unbiased, and fair enough, do we need to explain the models? In this section, we will argue that interpretability is required by examining the limitations of bias and fairness metrics and the scientific motivations for interpretability.

2.1 Limitations of bias and fairness metrics

There is no doubt that bias and fairness metrics present a vital role in validating models' behavior. However, a shared limitation is that they always measure known attributes (Barocas et al., 2019). For example, gender-bias metrics use gender attributes. This presents two challenges. Can we procure such attributes (known as protected attributes)? How do we prevent unanticipated biases?

2.1.1 Protected attribute procurement

Attributes like gender, race, age, disability, etc., are under U.S. law known as "protected attributes" (Xiang & Raji, 2019), and collecting and using these attributes is heavily regulated in most of the world. Andrus et al. (2021) write, "In many situations, however, information about demographics can be extremely difficult for practitioners to even procure.". Therefore, systematically measuring bias and fairness is not always practical (Andrus et al., 2021).

On the other hand, explanations often don't depend on knowing these protected attributes in advance and can provide a more qualitative analysis. For example, consider a résumé screening model, and an adversarial explanation (Ye et al., 2021) which tells us that removing "Woman" from "Member of Woman's Chess Club" changes the prediction from reject to recommend; then this would indicate a potentially harmful bias (Kodiyan, 2019). Therefore, explanations can serve a similar practical purpose to a fairness or bias metric without performing systematical correlations.

2.1.2 Unknown attribute bias

Although protected attributes are important to consider and are often legally protected, many more relevant attributes are involved in ensuring a fair and unbiased system. Unfortunately, it is impossible to consider every possible bias in advance. As an alternative, interpretability offers a more qualitative and explorative validation.

Continuing the example with résumés and automated hiring recommendations, during investigations by Fuller et al. (2021), the authors found that a hospital only accepted candidates with computer programming experience when they needed workers to enter patient data into a computer. Another example was a clerk position where applicants were rejected if they did not mention floor-buffing (i.e., a cleaning method for floors) (Fuller, 2021).

These examples present cases of systematic unintended bias. However, they do not relate to any protected attributes, and they are so specific they can only be discovered through qualitative explanations and investigations. That said, systematic fairness/bias metrics can quantify the damage once potential biases are identified using interpretability. Afterward, those metrics can be integrated into a quality assessment system to prevent future harm.

2.2 Interpretability for scientific discovery and understanding

Interpretability is not only used for ethics and adjacent purposes, where bias and fairness metrics have an important role. Interpretability is also used for scientific discovery and learning about what makes models work.

2.2.1 Scientific Discovery

An example of scientific discovery is interpretability in drug discovery (Preuer et al., 2019; Jiménez-Luna et al., 2020; Dara et al., 2022). A common approach is to use feature attribution to identify regions in genomic sequences responsible for a particular behavior, such as producing a protein. While these explanations do not guarantee that such connections exist in reality, they can provide important initial hypotheses for scientists enabling them to make more informed choices about the direction of their research.

2.2.2 Model understanding

An emerging field of interpretability is mechanistic interpretability, which identifies parts of a neural network that have a particular responsibility (Cammarata et al., 2020). For example, identifying a collection of neurons responsible for copying content in a generative language model, etc. (Elhage et al., 2021). Such insights may not be directly relevant to downstream tasks, but they help us understand current model limitations and can lead to better model design.

	Intrinsic paradigm	Post-hoc paradigm
defintion	The model is designed to provide explanations by making the explanation part of the model architecture.	The model is produced without regard for explanation, and the explanations are then created after model training.
underlying beliefs	Only models that were designed to be explained can be explained.	Although it may be very challenging, black-box models can be explained.
	Intrinsic models can have the same performance as a black-box model.	Black-box models will be more generally applicable than intrinsic models.

Table 1: Comparison of the definitions and underlying beliefs of the intrinsic and post-hoc paradigms. The beliefs relate to a) requirements for a faithful explanation and b) model capabilities. It should be apparent that these two views are seemingly incompatible.

3 The current paradigms of interpretability

This paper uses a common definition of interpretability, "the ability to explain or to present in understandable terms to a human" by Doshi-Velez & Kim (2017). However, even this definition of interpretability is not agreed upon.

Lipton says, "the term interpretability holds no agreed upon meaning, and yet machine learning conferences frequently publish papers which wield the term in a quasi-mathematical way" (Lipton, 2018). In 2017, a UK Government House of Lords review of AI noted after substantial expert evidence that "the terminology used by our witnesses varied widely. Many used the term transparency, while others used interpretability or explainability, sometimes interchangeably" (House of Lords, 2017, 91).

To the credit of the field, there have been many attempts at rectifying this with unified taxonomy Mohseni et al. (2021); Ali et al. (2023); Graziani et al. (2023). Unfortunately, there is still no universally agreed-upon definition of interpretability, nor the current paradigms of interpretability (Carvalho et al., 2019; Flora et al., 2022). As such, this section defines the *intrinsic* and *post-hoc* paradigms, as well as describe their underlying beliefs, which are summarized in Table 1.

3.1 Definitions

Jacovi & Goldberg (2020) write: "A distinction is often made between two methods of interpretability: (1) interpreting existing models via post-hoc techniques; and (2) designing inherently interpretable models. (Rudin, 2019)". Based on this and other sources (Schwalbe & Finzel, 2024; Dang et al., 2024; Molnar, 2020; Bonifácio, 2024; Madsen et al., 2022b; Arya et al., 2019; Carvalho et al., 2019; Murdoch et al., 2019), this paper refers to these two ideas respectively as 1) the *post-hoc* paradigm and 2) the *intrinsic* paradigm.

3.1.1 The intrinsic paradigm

The intrinsic paradigm works on creating so-called *inherently interpretable models*. These models are architecturally constrained, such that the explanation emerges from the architecture itself.

Classical examples are decision trees, linear regression, and prototypes (e.g. kNNs, Fix & Hodges 1951; Bien & Tibshirani 2009; Blei et al. 2003). In the field of neural networks, some examples are: 1) Generalized Additive Models (Agarwal et al., 2021; Lou et al., 2013; 2012) 2) Attention-based feature attribution (Bahdanau et al. 2015, Section 5.2.1; Luong et al. 2015, Appendix A; Vaswani et al. 2017, Appendix; Jain & Wallace 2019), where attention points to which input tokens are important. 3) Concept bottlenecks (Koh et al.,

2020; Zarlenga et al., 2022) 4) Neural Modular Networks (Andreas et al., 2016; Gupta et al., 2020; Fashandi, 2023), which produce a prediction via a sequence of sub-models, each with known behavior. 5) Prototypical Networks (Kim et al., 2014; Alvarez-Melis & Jaakkola, 2018; Chen et al., 2019), which predicts by finding similar training observations.

Occasionally, the term "*ante-hoc*" is used instead of *intrsic*, where "*ante-hoc*" means anything that isn't post-hoc (Retzlaff et al., 2024). This is much more encompassing than just architecturally constrained models, including also models with changes to their optimization procedure. However, such categorization is unsuitable when discussing paradigm shifts, as it's so encompassing that there are by definition no other paradigms. Additionally, *intrsic* captures more precisely the current literature; for example, almost all interpretability surveys only discuss the post-hoc paradigm or architecturally constrained models (i.e. the intrinsic paradigm) (Retzlaff et al., 2024; Schwalbe & Finzel, 2024; Dang et al., 2024; Molnar, 2020; Bonifácio, 2024; Madsen et al., 2022b; Arya et al., 2019; Carvalho et al., 2019; Murdoch et al., 2019).

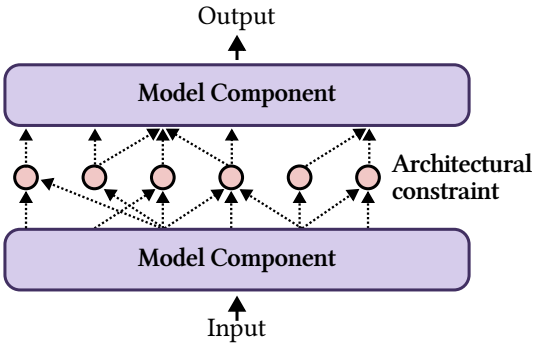


Figure 1: Abstract diagram of the intrinsic paradigm, where the model is architecturally constrained, such that the constraint itself is the explanation. In cases of Decision Trees the entire model is constrained, but often (e.g. Prototype Networks or Attention) only part of the model is constrained.

3.1.2 The post-hoc paradigm

Post-hoc explanations are computed after the model has been trained. They are developed independently of the model's architecture and how it was trained. However, there are often some simple criteria, like "the model should be differentiable", "the training dataset is known", or "inputs are represented as tokens" (Madsen et al., 2022b). Although general applicability is technically not a requirement, if a method is so specific that it only works on one specific model, it's likely an *intrinsic explanation*.

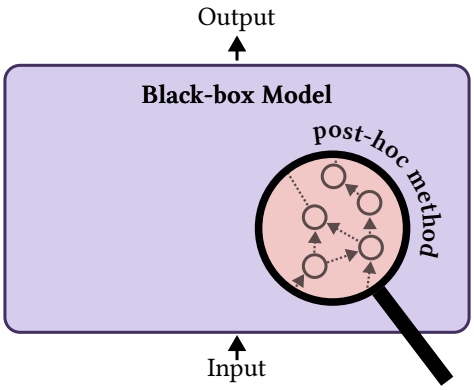


Figure 2: Abstract diagram of the post-hoc paradigm, where a post-hoc method is used to explain a black-box model. The post-hoc method is usually an algorithm, like the gradient w.r.t. the input, but it can also be an auxiliary model.

Common examples in the field of neural networks are: 1) gradient-based or occlusion-based feature attribution, such as Integrated Gradient (Sundararajan et al., 2017), Shapely approximations (Lundberg & Lee, 2017), LIME (Ribeiro et al., 2016), and Grad-CAM (Selvaraju et al., 2020), which indicates what input features are important. 2) Influence functions Koh & Liang (2017), which indicates what training observations are important. 3) Probing methods, such as BERTology (Belinkov & Glass, 2019; Belinkov et al., 2020), which show where information or concepts are stored. 4) Surrogate models (Alizadeh et al., 2020; Kazhdan et al., 2020), such as LORE (Guidotti et al., 2018), DeepRED (Zilke et al., 2016), and BETA (Lakkaraju et al., 2017), which distill the black-box model to an inherently interpretable model like decision tree (Craven & Shavlik, 1995). 5) Concept discovery, such as T-CAV (Kim et al., 2018) and ACE (Ghorbani et al., 2019), which identifies abstract properties which are relevant to the classification, like stribs' relevance in zebra-classification. All of these explanations apply to general models after training. For example, gradient-based feature attribution methods work by differentiating the prediction with respect to the input. The idea is that if a small change in input causes a big change in the output, then that input is important (Baehrens et al., 2010; Karpathy et al., 2015; Seo et al., 2018).

3.2 Beliefs

As with all paradigms, there are fundamental underlying beliefs, which are why the paradigm's followers partake in their paradigm of choice. At the core of these beliefs are two central questions. When are explanations faithful and what are the requirements for faithfulness? And, how do these requirements affect the model's general performance capabilities? This section defines these beliefs and the motivations that lead to them, section Section 4 then discusses the limitations of the intrinsic and post-hoc paradigms in the context of these beliefs.

3.2.1 When are explanations faithful?

The intrinsic paradigm believes that: *only models designed to be explained, can be explained*, and the only approach to achieve this is via architectural constraints, i.e. *inherently interpretable models*. Specifically, they argue that using black-box models and post-hoc explanations is dangerous, as these models and methods do not guarantee faithfulness (Rudin, 2019).

The *post-hoc* paradigm takes a less strict stance and believes that even models that were not designed to be explained (i.e., black-box models) can still be explained. Although faithful post-hoc explanations may be much more challenging to produce, *post-hoc* paradigm believes it's possible.

In conclusion, the intrinsic paradigm considers explanations to be part of the model design, and post-hoc explanations are always applied after the model design. Madsen et al. (2022b) frame intrinsic as proactive and post-hoc as retroactive. Hence, the two schools of thought are incompatible frameworks, and they can philosophically be considered as paradigms (Kuhn, 1996).

3.2.2 What is the effect on the model's general performance capabilities?

It would seem that *intrinsic explanation* is the obvious choice. If we can control the model such that the faithfulness of explanations can be guaranteed, why consider *post-hoc explanation*?

The commonly mentioned idea is that the *post-hoc* paradigm believes that by constraining the models in the manners that the *intrinsic paradigm* requires, there is a trade-off in performance (DARPA, 2016). However, this trade-off does not have to be the case in practice (Rudin, 2019, section 2).

A more accurate take, which is rarely explicitly discussed, is that the common industry prefers off-the-shelf general-purpose models and only later thinks about interpretability (Bhatt et al., 2019). Additionally, most research only considers predictive performance, not interpretability. Therefore, *intrinsic* researchers are always catching up to black-box models. From the *post-hoc* perspective, it would make more sense to work on generally applicable interpretability methods for both off-the-shelf and future black-box models.

From the intrinsic perspective, while the industry might prefer off-the-shelf models now, they shouldn't. Not validating models through intrinsic explanations can have serious consequences (Rudin, 2019) and eventually

damage their business. Additionally, with increasing legal requirements to provide explanations, the industry may have to use inherently explainable models (Goodman & Flaxman, 2017).

For these reasons, the *intrinsic* paradigm believes we should not let the industry's needs dictate our research direction, as their goals may be too short-sighted. In the long run, intrinsic models may be the only reasonable option.

In conclusion, the *post-hoc* paradigm has good intentions of providing general explanations for general-purpose models. However, from the *intrinsic* paradigm perspective, those good intentions are meaningless if it is fundamentally impossible to provide guaranteed faithful explanations without an *inherently interpretable model*.

4 Why interpretability needs a new paradigm

It tends to be the case that when there are multiple paradigms, it is because neither of the paradigms fits the needs. However, for the case of the *post-hoc* and *intrinsic* paradigms, it could be argued that they serve different needs. For example, *intrinsic* explanations should be preferred for critical applications (Rudin, 2019), and *post-hoc* explanations could be used for verifiable situations, such as drug discovery, where the hypothesis generated by the explanations is verified using physical experiments.

4.1 The case against the intrinsic paradigm

The industry primarily uses post-hoc explanations, including for high-stakes applications such as insurance risk assessment and financial loan assessment (Bhatt et al., 2019; Krishna et al., 2022). This is because such industries usually do not have the in-house expertise to develop custom high-performing inherently interpretable models for their specific task. They must rely on existing inherently interpretable models, which are not generally competitive, or use more advanced off-the-shelf neural black-box models, like pre-trained language models, which will be competitive. In practice, the industry is thus often not in a position to choose inherently interpretable models.

Another challenge with the intrinsic paradigm is that its models are often not completely interpretable because only a part of the model is architecturally constrained to be interpretable. The rest of the neural network, still use black-box components (e.g. Dense layer, Recurrent layer, etc.) which are not interpretable. As such, the intrinsic promise should not be taken at face value (Jacovi & Goldberg, 2020).

To summarize, intrinsic methods are either not competitive in terms of predictive accuracy, general-purpose enough for the industry (Bhatt et al., 2019), or their intrinsic claims are unsupported (Jacovi & Goldberg, 2020). We will here give a few examples where this can be observed.

An example of a lack of general-purpose performance is General Additive Models (GAMs). GAMs map each input feature via non-linear models to separate latent representations and then combine these via a linear model to the final prediction Lou et al. (2012). GAMs have been used successfully in practice (Caruana et al., 2015; Lou et al., 2013), sometimes by extending it to all feature-pairs (Schug et al., 2023). The limitation is that they only work well on tasks that do not require high-order combinatorial feature modeling, which is unfortunately often the case.

An example of unsupported faithfulness is classic attention-based models. Attention itself is interpretable, as it's a weighted sum, and explains the importance of each intermediate representation. However, attention is often used as token-importance (Bahdanau et al. 2015, Section 5.2.1; Luong et al. 2015, Appendix A; Vaswani et al. 2017, Appendix; Jain & Wallace 2019). This is not faithful, as the intermediate representations are produced by a black-box recurrent neural network (e.g. LSTM Hochreiter & Schmidhuber 1997) which can mix or move the relationship between tokens and the intermediate representations. Therefore, the attention scores do not necessarily represent token-importance (Bastings & Filippova, 2020).

Another example is Neural Modular Networks, which produce an executable problem composed of sub-networks, such as `find-max-num(filter(find()))`, which is interpretable (Fashandi, 2023; Andreas et al., 2016; Gupta et al., 2020). However, each sub-networks (`find-max-num`, `filter`, `find`) is itself a black-box

model with little guarantee that it operates as intended (Amer & Maul, 2019; Subramanian et al., 2020; Lyu et al., 2024).

The dynamic between faithfulness and general-purpose predictive accuracy is often observed with concept bottlenecks (Koh et al., 2020; Zarlenga et al., 2022), where a layer in a neural network restricts the intermediate representation to activations of pre-determined concepts; for example, wing-color, beak-length, etc. in a bird-classification task. This requires all relevant concepts to be known and labeled. However, this is rarely satisfied and works have shown that the concepts leak extraneous information unrelated to the concepts (Margeloiu et al., 2021; Mahinpei et al., 2021). Recent works have attempted to control this leakage by allocating vector space for unknown concepts, but their faithfulness is still lacking (Ismail et al., 2024).

Overall, there are few success stories within the intrinsic paradigm, where intrinsically faithful explanations have been provided without impacting predictive accuracy and model generality.

4.2 The case against the post-hoc paradigm

Although post-hoc explanations directly address the interpretability challenge of black-box components and models, and could therefore provide more complete explanations, there are also few success stories with post-hoc, where post-hoc explanations are consistently faithful.

Most notable is perhaps post-hoc feature attribution explanations (also known as importance measures, IMs), where the explanation indicates which input features are the most important for making a prediction. The pursuit of such explanations has produced countless papers (Erhan et al., 2009; Štrumbelj & Kononenko, 2014; Zeiler & Fergus, 2014; Karpathy et al., 2015; Li et al., 2016; Shrikumar et al., 2017; Smilkov et al., 2017; Ahern et al., 2019; Thorne et al., 2019; ElShawi et al., 2019; Sangroya et al., 2020), among the most popular are methods like LIME (Ribeiro et al., 2016), Shapely approximations (Lundberg & Lee, 2017), Grad-CAM (Selvaraju et al., 2020), and Integrated Gradient (Sundararajan et al., 2017).

However, repeatedly, the faithfulness of these IM explanations is criticized (Adebayo et al., 2018; 2021; Kindermans et al., 2019; Hooker et al., 2019; Yeh et al., 2019). For example, there is great disagreement between alleged faithful IMs, which is hard to reconcile (Jain & Wallace, 2019; Krishna et al., 2022), and they are not robust to adversarial model attacks (Bordt et al., 2022; Slack et al., 2020). Other works show their faithfulness is both task and model-dependent and thus don't provide the generality that the *post-hoc* paradigm desires (Bastings et al., 2022; Madsen et al., 2022a). Finally, theoretical works suggest that IMs are subject to a *no free lunch theorem* (Han et al., 2022), or it may be impossible to provide faithful post-hoc IMs (Bilodeau et al., 2024).

Similar to the work of IM, is the visualization of neurons in computer vision, which shows that neurons represent high-level concepts, such as nose or dog. This is done by visualizing convolutional weights or the input image that maximizes a neuron's activation (Olah et al., 2017; Nguyen et al., 2016; Yosinski et al., 2015), which provides very convincing evidence. However, it has been shown empirically, theoretically, and through human-computer-interaction (HCI) studies that these visualizations do not provide comprehensible explanations regarding the neurons' responsibility (Geirhos et al., 2023; Borowski et al., 2021; Zimmermann et al., 2021)[1].

Another notable example is probing explanations, where models are verified by relating the model's behavior or intermediate representation to, for example, linguistic properties (part-of-speech, etc.) (Belinkov & Glass, 2019; Belinkov et al., 2020). This idea has produced an entire subfield called BERTology (Rogers et al., 2020). BERTology in particular has attained substantial attention (Coenen et al., 2019; Clark et al., 2019; Rogers et al., 2020; Clouatre et al., 2022; McCoy et al., 2019; Conneau et al., 2018; Tenney et al., 2019), with most of the works finding that neural networks can learn linguistic properties indirectly.

Unfortunately, like post-hoc feature attribution, there are many reasons to be highly skeptical (Belinkov, 2021). For example, using an untrained model or a randomized dataset shows an equally high correlation with linguistic properties, compared with training a regular model (Zhang & Bowman, 2018; Hewitt & Liang,

[1]Neural networks likely do encode high-level concepts, but these visualizations are not effective (i.e. low comprehensibility) for identifying the responsibility of specific neurons.

	Learn-to-faithfully-explain paradigm	Faithfulness measurable model paradigm	Self-explaining model paradigm
defintion	The model is optimized such that an explanation method becomes faithful.	The model is designed to enable measuring faithfulness of a category of explanations.	The model directly outputs both its prediction and an explanation for that prediction.
underlying beliefs	The relaxed faithfulness metric used for optimization is a sufficient approximation. Models can be optimized such that explanations become faithful without losing performance.	It is computationally feasible to optimize explanations for optimal faithfulness. Models can be optimized to be faithfulness measurable without loss of predictive performance.	Models can be trained to model and articulate their own reasoning accurately. Self-explanation capabilities do not negatively impact regular predictions.

Table 2: Comparison of the definitions and underlying beliefs of the new paradigms. The beliefs relate to a) explanation requirements and b) model capabilities. These new paradigms can be compared with the old paradigms in Table 1.

2019). These discoveries have put the entire methodology into question, although there is work trying to adapt to these new critiques (Voita & Titov, 2020).

4.3 Summary

Post-hoc feature attribution and probing explanations are just two cases where post-hoc shows initial promise through countless papers, only to be debunked repeatedly. The trend is oscillating between proposing new explanation methods and debunking them. Of course, it's impossible to prove that there will never be a great post-hoc method. However, the lack of guarantees also makes it impossible to know when a faithful post-hoc method is proposed. Similarly, intrinsic explanations also receive criticism after a while, as has been the case with attention and Neural Modular Networks.

5 Are new paradigms possible?

Although both the intrinsic and post-hoc paradigms have significant issues, parts of their underlying beliefs have merit. The intrinsic paradigm believes that *we can't expect models that were not designed to be explained, to be explained*, while post-hoc believes *black-box models tend to be more general purpose while providing high predictive performance*. These beliefs have merit, and it's worth considering how to incorporate their spirit into new paradigms.

It can seem unlikely that such a paradigm can even exist. However, there is already some work that satisfies these desirables. In particular, we have identified 3 alternative paradigms, summarized in Table 2. These emerging directions have unfortunately not received much focus, likely due to favoritism towards existing paradigms (Kuhn, 1996).

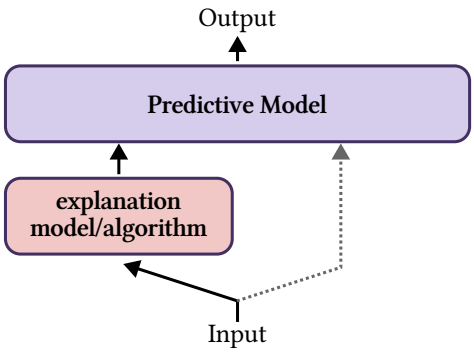
All three paradigms work with what would be black-box models. However, their idea is to optimize these models so that they are designed to be explained. How they differ is in their exact formulation of this approach.

It's important to note, that it's only with hindsight we can truly know if a new idea will become the next major paradigm, and it may be a fourth unknown idea that will become the next major paradigm. As such, the main purpose of this section is not to promote the next paradigm but rather to establish that it is possible to develop new interpretability paradigms.

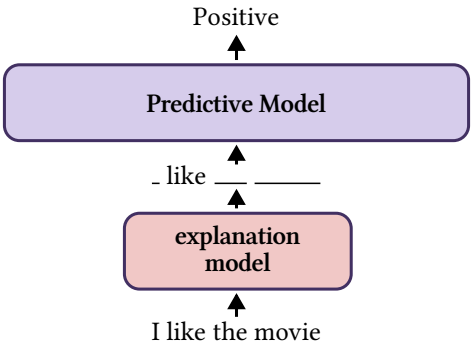
5.1 The learn-to-faithfully-explain paradigm

This paradigm is the most direct application of the optimization idea. An existing explanation algorithm (Bhalla et al., 2023) or model is used (Yoon et al., 2019; Chen et al., 2018), and the predictive model is then optimized to maximize both the predictive performance and the faithfulness. In addition to faithfulness, it's possible to also optimize for comprehensibility-proxies such as sparsity (Bhalla et al., 2023; Jethani et al., 2021).

Importantly, this approach does not require the architectural constraints that the intrinsic paradigm applies, as the explanation comes from an external explanation method, not the architectural design. The explanation method can be similar or even identical to those from the post-hoc paradigm. However, because the model is optimized to enable these explanations to be faithful, it's not post-hoc, and there are more reasons to think that the explanations is faithful.



(a) Abstract diagram, showing the joint or disjoined amortized explanation setup. Care needs to be taken that the explanation model doesn't perform the prediction.



(b) Concrete example, where an explanation model produces a feature attribution explanation, showing only "like" is important.

Figure 3: The learn-to-faithfully explain paradigm. In most cases, this paradigm works by generating an explanation from the input, using either a model or an algorithm, this explanation is then fed into the predictive model, which has been optimized to respect the explanation.

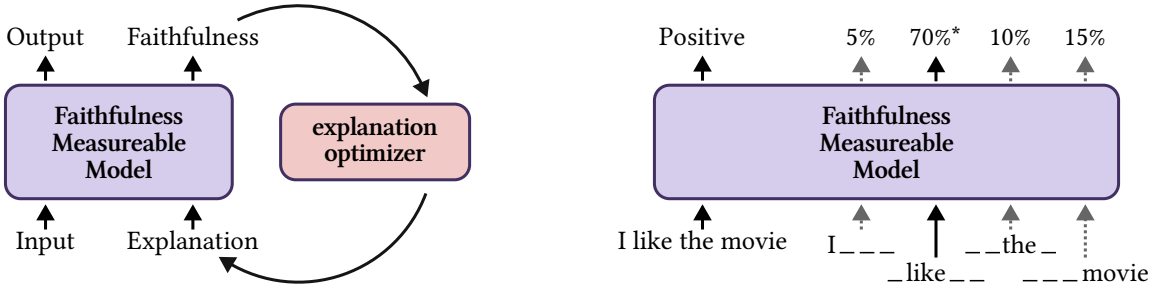
Early work on this jointly trains an explanation model and a prediction model (Yoon et al., 2019; Chen et al., 2018). This direction has been called joint amortized explanation methods (JAMs). However, Jethani et al. (2021) point out that the explanation model often learns to encode the prediction, which means the explanation model becomes part of the black-box problem rather than the solution. A solution can be to use a disjoint setup (Jethani et al., 2021), where the explanation model can't encode the prediction, a setup that the following works have adapted (Jethani et al., 2022; Covert et al., 2022). However, the explanation model may still output unfaithful explanations for out-of-distribution inputs. An alternative is to produce the explanation algorithmically (Bhalla et al., 2023), for example by having an explanation algorithm remove unnecessary features, and the prediction model learns to support sparse features.

Regardless of the specific approach used to produce the explanation, the challenges are formalizing the faithfulness objective correctly such that the optimization works as intended, ensuring that the explanations are truly faithful and that the model properties that make explanations faithful also hold for out-of-distribution data (Covert et al., 2022; Bhalla et al., 2023). To validate faithfulness present methods use ground-truths (Jethani et al., 2022; Bhalla et al., 2023) but this is only feasible for simple or synthetic tasks. Another reasonable concern is that adding faithfulness to the optimization objective decreases the predictive accuracy as capacity in the predictive model needs to be used for this objective. As of yet, there is no strong analysis of this concern. However, existing work suggests there are no performance penalties (Covert et al., 2022; Jethani et al., 2022), and the predictive model becomes more robust to adversarial attacks (Bhalla et al., 2023).

5.2 The faithfulness measurable model paradigm

Normally, measuring faithfulness is extremely challenging (Jacovi & Goldberg, 2020). However, the faithfulness measurable model (FMM) paradigm designs the model such that faithfulness can be easily and precisely measured without requiring architectural constraints. Importantly, because faithfulness is easy to measure by design in FMMs, it's possible to identify the explanation that maximizes faithfulness using optimization algorithms (Zhou & Shah, 2023), which makes the model indirectly intrinsically explainable (Madsen et al., 2024b; Hase et al., 2021; Vafa et al., 2021). In essence, this paradigm reformulates the intrinsic paradigm from 'inherently explainable" to "inherently measurable".

Additionally, because faithfulness can be easily and precisely measured, it's possible to present the faithfulness metric along with the explanation to the user. This can add confidence to the explanation, thus increasing the comprehensibility. While this is technically possible in any paradigm, this paradigm is designed for it.



(a) Abstract diagram, showing the optimization loop between the faithfulness measurable model and the optimizer.

(b) Concrete example, showing how the faithfulness of all potential feature attribution explanations is measured, the optimal explanation ("like") is then selected.

Figure 4: The faithfulness measurable model paradigm. In this paradigm, the predictive model can also measure how faithful a given explanation is. The explanation can thus be produced by optimizing an initial (maybe random) explanation towards maximal faithfulness.

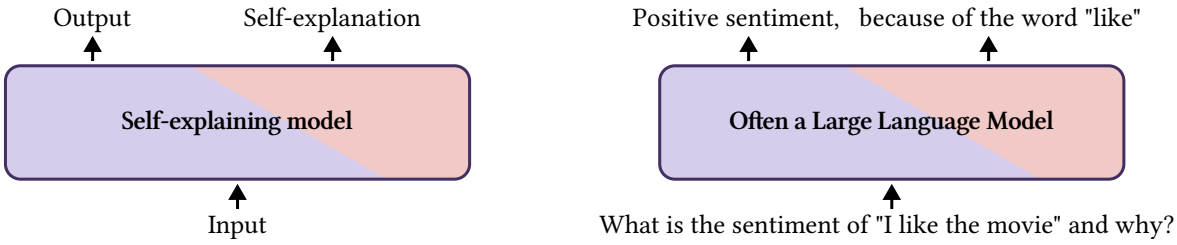
Madsen et al. (2024b) and Hase et al. (2021) show that this idea can be achieved using simple data argumentation, and there is no need for architectural constraints. The central idea is to use the erasure metric (Samek et al., 2017) to measure the faithfulness of feature attribution explanations. The erasure metric says: if information (pixels, tokens, etc.) is truly important, then when removing it the prediction should change significantly. The common challenge is that removing information causes out-of-distribution issues (Hooker et al., 2019; Madsen et al., 2022a). However, by using data argumentation during training, it's possible to extend the model to support the partial inputs created by the erasure metric. Importantly this can be achieved without architectural constraints, thus it remains possible to use general-purpose models such as RoBERTa (Madsen et al., 2024b) and GPT-2 (Vafa et al., 2021).

The challenge in this paradigm is about coming up with a way to integrate the faithfulness metric in the model while ensuring there is no performance impact and that the model operates in-distribution (Madsen et al., 2024b). Additionally, developing efficient optimization procedures for optimizing explanations is difficult, due to the discrete nature of many explanations (Hase et al., 2021; Zhou & Shah, 2023).

5.3 The self-explaining model paradigm

Rather than using external algorithms or models to produce explanations, Elton (2020) proposes in this paradigm that models should directly output explanations themselves, meaning they become *self-explaining*. This differs from the intrinsic paradigm, as the explanation is produced only by basic inference, not architectural constraints. It is also not post-hoc, as models do not explain themselves without some training towards this. Elton (2020) idea is to consider different sub-models, which output prediction, explanation, and confidence; which all produce these from a latent representation. However, today the most common implementation of this idea is instruction-tuned large language models (e.g., ChatGPT, Gemini, etc.) (OpenAI, 2023; Jiang

et al., 2023; Meta, 2023), which are able to explain themselves in great detail and very convincingly (Chen et al., 2023), because they have been optimized towards this objective (Agarwal et al., 2024).



(a) Abstract diagram, where output and explanation is produced by the same model. This idea is not limited to LLMs, Elton (2020) writes about alternatives.

(b) Concrete example, showing a feature attribution explanation saying "like" is important. Feature attributions is not the normal use for LLM self-explanations.

Figure 5: The self-explanation paradigm, where the same model is trained to produce both the regular predictive output and an explanation, called a self-explanation. This paradigm is often seen with Large Language Models, where both the predictive output and the self-explanations appear as generated text.

Unfortunately, because the explanations are produced by a black-box model this paradigm can be quite dangerous. Therefore, there must be solid evidence that the explanations are faithful for this approach to be valid. However, despite this immediate danger, the model that generates the explanation can in principle have access to all of the logic that produces the prediction. At a minimum, the same weights produce both the prediction and the explanation.

Importantly, self-explanations must relate to the model's reasoning logic, not just the world or abstract concepts. However, presently there is little evidence that this is satisfied (Turpin et al., 2023; Lanham et al., 2023; Madsen et al., 2024a). This is not surprising, as the self-explanations are explicitly trained based on humans' annotating how these explanations should look. However, humans don't have any insight into how the model operates (Jacovi & Goldberg, 2020). As such, the model converges towards very convincing self-explanations with no regard for faithfulness (Agarwal et al., 2024; Chen et al., 2023).

Works addressing the faithfulness challenges of self-explanation are now emerging, with approaches using a combination of in-context learning, self-correction, and post-training to align the model to produce faithful explanations Pan et al. (2024); Chuang et al. (2024); Chen et al. (2024); Paul et al. (2024); Tanneru et al. (2024). However, the improvements have been minor and it's still considered a hard problem (Tanneru et al., 2024). Additionally, just measuring faithfulness of general self-explanation remains a challenge (Huang et al., 2023; Parcalabescu & Frank, 2023; Pan et al., 2024). Despite these challenges, this direction may be worthwhile as self-explanations are a natural solution to generating natural language explanations which are often considered highly comprehensible (Luo et al., 2024; Agarwal et al., 2024).

6 Limitations

This position paper primarily focuses on faithfulness, without talking much about comprehensibility. The reason is that the paradigms' underlying beliefs are rooted in concerns regarding faithfulness and performance (Section 3.2) and concerns regarding comprehensibility often first materialize when talking about a specific explanation type, which can generally be produced by any paradigm. For example, feature attributions have been produced within all 5 paradigms ((Bahdanau et al., 2015; Baehrens et al., 2010; Chen et al., 2018; Hase et al., 2021; Huang et al., 2023)) and there is significant HCI literature discussing the comprehensibility of feature attributions (Sen et al., 2020; Schuff et al., 2022b; Rong et al., 2024; Kaur et al., 2020; Gilpin et al., 2018; Prasad et al., 2021; Schuff et al., 2022a; Lertvittayakumjorn & Toni, 2019; Lage et al., 2019).

That being said some paradigms may be better suited for some explanation types. For example, the self-explanation paradigm produces highly convincing natural language explanations (Chen et al., 2023; Agarwal et al., 2024), but is likely ill-suited for feature attribution (Huang et al., 2023). Readers are encouraged to explore such connections.

Additionally, new orthogonal paradigms on comprehensibility have been proposed. For example, readers are encouraged to study recent works like Schut et al. (2023); Kim (2022), which propose the new idea that it is not enough to frame explanations in terms that humans already understand. We should also develop new language and mental abstractions for humans to understand machines. Such ideas are orthogonal to this position paper, as they can be applied to any of the paradigms discussed.

Finally, this position paper does not discuss which specific faithfulness metric to use. Similarly to comprehensibility, this is because which faithfulness metric to use depends on the type of explanation and the paradigms discussed in this position paper cover many or potentially all explanation types. Additionally, developing faithfulness metrics themselves is ongoing research; this is particularly true for the self-explanation paradigm (Turpin et al., 2023; Lanham et al., 2023; Madsen et al., 2024a). Readers are encouraged to study the literature on the meta-evaluation of faithfulness metrics (Hedström et al., 2023; Wang & Wang, 2022), surveys (Zhou et al., 2021), and principles of developing faithfulness metrics (Jacovi & Goldberg, 2020; 2021).

7 Conclusion

Although some evidence exists for the new paradigms presented in Section 5, these are, first and foremost, just ideas. It's only in retrospect that we can truly know if one paradigm results in meaningful progress in the field. It is also entirely possible that neither of these ideas is what moves the interpretability field forward.

For these reasons, the core position of this paper is that we should develop new directions and paradigms in interpretability, instead of focusing on the existing post-hoc and intinsic paradigms, which are currently dominating.

That being said, we must also be vigilant regarding faithfulness to avoid repeating past mistakes (Jacovi & Goldberg, 2020). These new paradigms will present new arguments for why their method is faithful. As we are unfamiliar with these arguments, identifying their flaws is difficult, and it will be easy to get swayed by them.

Historically, a common tactic in post-hoc works was convincing visualization of explanations that aligned with our intuitions (Olah et al., 2017; Yosinski et al., 2015; Nguyen et al., 2016). However, such visualizations are empty arguments, as humans can't know what a true explanation looks like (Geirhos et al., 2023). Likewise, intrinsic works have made seemingly strong theoretical arguments for why their methods are faithful, but these arguments failed to capture the whole model. Even the new *learn-to-faithfully-explain paradigm* have already shown sharp corners, where the explainer model unintentionally encodes the prediction and is therefore unfaithful (Jethani et al., 2021).

To prevent false arguments, a sound start is to always have a specific and measurable definition of faithfulness, which works for all methods within a given explanation category (e.g., counterfactual or feature attribution).

Finally, while these new paradigms are promising, it's unlikely they will completely erase the current paradigms. We still teach both the particle and wave paradigms in physics. Most scientists don't worry about whether there are true statements in math that cannot be proven.

Likewise, there will likely always be situations where intrinsic or post-hoc interpretability makes sense. For example, basic statistics and linear regressions can be framed as intrinsic interpretability. Hence, if a company or researchers decide to use a model because of its intrinsically explainable properties, then we should only praise them – as long as they also measure the faithfulness of the explanations.

Broader Impact Statement

This paper presents new paradigms for interoperability, focused on ensuring that explanations are faithful to the model. Should this happen, this would have a significant positive impact on society.

However, the paper also discusses existing paradigms, namely post-hoc and intrinsic, citing reasons for why they may not be fruitful directions. This could be problematic if one of those paradigms is truly fruitful, but the field just hasn't found the right method yet, and researchers feel discouraged from working in these directions because of our paper.

Discouraging work is not the intended outcome of this paper. To prevent this, we specifically write in the conclusion that the field should encourage work on interpretability in general, as long as the faithfulness is well supported. As such, there should be no reason for concern regarding this paper's potential negative impact.

References

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 2018-Decem, pp. 9505–9515. Curran Associates, Inc., 10 2018. URL http://arxiv.org/abs/1810.03292.

Julius Adebayo, Michael Muelly, Harold Abelson, and Been Kim. Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation. In *International Conference on Learning Representations*, pp. 1–13, 2021. URL https://openreview.net/forum?id=xNOVfCCvDpM.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. *arXiv*, 2024. URL http://arxiv.org/abs/2402.04614.

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural Additive Models: Interpretable Machine Learning with Neural Nets. *Advances in Neural Information Processing Systems*, 6(NeurIPS):4699–4711, 2021. ISSN 10495258.

Isaac Ahern, Adam Noack, Luis Guzmán-Nateras, Dejing Dou, Boyang Li, and Jun Huan. Normlime: A new feature importance metric for explaining deep neural networks. *arXiv*, 9 2019. ISSN 23318422. URL http://arxiv.org/abs/1909.04200.

Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99(April), 2023. ISSN 15662535. doi: 10.1016/j.inffus.2023.101805.

Reza Alizadeh, Janet K. Allen, and Farrokh Mistree. Managing computational complexity using surrogate models: a critical review. *Research in Engineering Design*, 31(3):275–298, 2020. ISSN 14356066. doi: 10.1007/s00163-020-00336-7. URL https://doi.org/10.1007/s00163-020-00336-7.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):7775–7784, 2018. ISSN 10495258.

Mohammed Amer and Tomás Maul. A review of modularization techniques in artificial neural networks. *Artificial Intelligence Review*, 52(1):527–561, 6 2019. ISSN 0269-2821. doi: 10.1007/s10462-019-09706-7. URL http://link.springer.com/10.1007/s10462-019-09706-7.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural Module Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 39–48. IEEE, 6 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.12. URL http://ieeexplore.ieee.org/document/7780381/.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, We can't understand: Challenges to demographic data procurement in the pursuit of fairness. *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 249–260, 2021. doi: 10.1145/3442188.3445888.

Vijay Arya, Rachel K. E. Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Q. Vera Liao, Ronny Luss, Aleksandra Mojsilović, Sami Mourad, Pablo Pedemonte, Ramya Raghavendra, John Richards, Prasanna Sattigeri, Karthikeyan Shanmugam, Moninder Singh, Kush R. Varshney, Dennis Wei, and Yunfeng Zhang. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv*, 9 2019. ISSN 23318422. URL http://arxiv.org/abs/1909.03012.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11: 1803–1831, 12 2010. ISSN 15324435. URL http://arxiv.org/abs/0912.1128.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, pp. 1–15. International Conference on Learning Representations, ICLR, 9 2015. URL https://arxiv.org/abs/1409.0473.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. URL https://fairmlbook.org/.

Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–155, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.blackboxnlp-1.14. URL https://www.aclweb.org/anthology/2020.blackboxnlp-1.14.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. "Will You Find These Shortcuts?" A Protocol for Evaluating the Faithfulness of Input Salience Methods for Text Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 976–991, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.64. URL https://aclanthology.org/2022.emnlp-main.64.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *arXiv*, pp. 1–12, 2 2021. URL http://arxiv.org/abs/2102.12452.

Yonatan Belinkov and James Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 4 2019. ISSN 2307-387X. doi: 10.1162/tacl{_}a{_}00254. URL https://doi.org/10.1162/tacl_a_00254.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. Interpretability and Analysis in Neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 1–5, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-tutorials.1. URL https://www.aclweb.org/anthology/2020.acl-tutorials.1.

Usha Bhalla, Suraj Srinivas, and Himabindu Lakkaraju. Verifiable Feature Attributions: A Bridge between Post Hoc Explainability and Inherent Interpretability. *arXiv*, pp. 1–14, 2023. URL http://arxiv.org/abs/2307.15007.

Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable Machine Learning in Deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 648–657, 9 2019. doi: 10.1145/3351095.3375624. URL https://dl.acm.org/doi/10.1145/3351095.3375624.

Jacob Bien and Robert Tibshirani. Classification by Set Cover: The Prototype Vector Machine. *arXiv*, pp. 1–24, 2009. URL http://arxiv.org/abs/0908.2284.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):1–38, 1 2024. ISSN 0027-8424. doi: 10.1073/pnas.2304406120. URL https://pnas.org/doi/10.1073/pnas.2304406120http://arxiv.org/abs/2212.11870.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. ISSN 15324435. URL https://jmlr.org/papers/v3/blei03a.html.

Stella Mendes Meireles Bonifácio. *Explainable AI: A case study on a Citizen's Complaint Text Classification Model*. PhD thesis, Universidade de Brasilia, 2024. URL http://repositorio.unb.br/handle/10482/50962.

Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. *ACM International Conference Proceeding Series*, 1:891–905, 1 2022. doi: 10.1145/3531146.3533153. URL http://arxiv.org/abs/2201.10295http://dx.doi.org/10.1145/3531146.3533153.

Judy Borowski, Roland S. Zimmermann, Judith Schepers, Robert Geirhos, Thomas S.A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary Natural Images Explain Cnn Activations Better Than State-of-the-Art Feature Visualization. *ICLR 2021 - 9th International Conference on Learning Representations*, pp. 1–41, 2021.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, and Ludwig Schubert. Thread: Circuits. *Distill*, 5(3), 3 2020. ISSN 2476-0757. doi: 10.23915/distill.00024. URL https://distill.pub/2020/circuits.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noémie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015-Augus:1721–1730, 2015. doi: 10.1145/2783258.2788613.

Diogo V. Carvalho, Eduardo M. Pereira, and Jaime S. Cardoso. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8):832, 7 2019. ISSN 2079-9292. doi: 10.3390/electronics8080832. URL https://www.mdpi.com/2079-9292/8/8/832.

Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, 32, 6 2019. ISSN 10495258. URL http://arxiv.org/abs/1806.10574.

Jianbo Chen, Le Song, Martin J. Wainwright, and Michael I. Jordan. Learning to explain: An information-theoretic perspective on model interpretation. *35th International Conference on Machine Learning, ICML 2018*, 2:1386–1418, 2 2018. URL http://arxiv.org/abs/1802.07814.

Yanda Chen, Ruiqi Zhong, Narutatsu Ri, Chen Zhao, He He, Jacob Steinhardt, Zhou Yu, and Kathleen McKeown. Do Models Explain Themselves? Counterfactual Simulatability of Natural Language Explanations. *arXiv*, 2023. URL http://arxiv.org/abs/2307.08678.

Yanda Chen, Chandan Singh, Xiaodong Liu, Simiao Zuo, Bin Yu, He He, and Jianfeng Gao. Towards Consistent Natural-Language Explanations via Explanation-Consistency Finetuning. *arXIv*, 2024. URL http://arxiv.org/abs/2401.13986.

Yu-Neng Chuang, Guanchu Wang, Chia-Yuan Chang, Ruixiang Tang, Fan Yang, Mengnan Du, Xuanting Cai, and Xia Hu. Large Language Models As Faithful Explainers. *ArXiv*, 2 2024. URL http://arxiv.org/abs/2402.04678.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://www.aclweb.org/anthology/W19-4828.

Louis Clouatre, Prasanna Parthasarathi, Amal Zouaq, and Sarath Chandar. Local Structure Matters Most: Perturbation Study in NLU. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3712–3731, Stroudsburg, PA, USA, 7 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.293. URL https://aclanthology.org/2022.findings-acl.293.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. Visualizing and Measuring the Geometry of BERT. *Advances in Neural Information Processing Systems*, 32:8594–8603, 6 2019. ISSN 23318422. URL https://proceedings.neurips.cc/paper/2019/file/159c1ffe5b61b41b3c4d8f4c2150f6c4-Paper.pdfhttp://arxiv.org/abs/1906.02715.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics. ISBN 9781948087322. doi: 10.18653/v1/P18-1198. URL http://aclweb.org/anthology/P18-1198.

Ian Connick Covert, Chanwoo Kim, and Su-In Lee. Learning to Estimate Shapley Values with Vision Transformers. *The Eleventh International Conference on Learning Representations*, pp. 1–48, 2022. URL http://arxiv.org/abs/2206.05282https://openreview.net/forum?id=5ktFNz_pJLK.

Mark W. Craven and Jude W. Shavlik. Extracting Tree-Structured Representations of Trained Networks. *NIPS 1995: Proceedings of the 8th International Conference on Neural Information Processing Systems*, pp. 24–30, 1995.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, Yong Liu, Jing Shao, Hui Xiong, and Xuming Hu. Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey. *arXiv*, 12 2024. URL http://arxiv.org/abs/2412.02104.

Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, C H Madhu Babu, and Mohamed Jawed Ahsan. Machine learning in drug discovery: a review. *Artificial Intelligence Review*, 55(3):1947–1999, 2022.

DARPA. Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53. *Defense Advanced Research Projects Agency (DARPA)*, pp. 1–52, 2016. URL https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf.

Joseph W. Dauben. Georg Cantor and Pope Leo XIII: Mathematics, Theology, and the Infinite. *Journal of the History of Ideas*, 38(1):85, 1 1977. ISSN 00225037. doi: 10.2307/2708842. URL https://www.jstor.org/stable/2708842?origin=crossref.

Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*, 2 2017. URL http://arxiv.org/abs/1702.08608.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A Mathematical Framework for Transformer Circuits. *Anthropic*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. ILIME: Local and Global Interpretable Model-Agnostic Explainer of Black-Box Decision. In Tatjana Welzer, Johann Eder, Vili Podgorelec, and Aida Kamišalić Latifić (eds.), *Advances in Databases and Information Systems*, pp. 53–68. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28730-6. doi: 10.1007/978-3-030-28730-6{_}4. URL http://link.springer.com/10.1007/978-3-030-28730-6_4.

Daniel C. Elton. Self-explaining AI as an Alternative to Interpretable AI. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12177 LNAI:95–106, 2 2020. ISSN 16113349. doi: 10.1007/978-3-030-52152-3{_}10. URL http://link.springer.com/10.1007/978-3-030-52152-3_10.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, 2009. URL http://igva2012.wikispaces.asu.edu/file/view/Erhan+2009+Visualizing+higher+layer+features+of+a+deep+network.pdf.

Homa Fashandi. Neural module networks: A review. *Neurocomputing*, 552:126518, 2023. ISSN 18728286. doi: 10.1016/j.neucom.2023.126518. URL https://doi.org/10.1016/j.neucom.2023.126518.

E Fix and JL Hodges. Nonparametric discrimination: consistency properties. Technical Report Feb, University of California, 1951. URL https://web.archive.org/web/20200706215717/https://apps.dtic.mil/dtic/tr/fulltext/u2/a800276.pdf.

Montgomery Flora, Corey Potvin, Amy McGovern, and Shawn Handler. Comparing Explanation Methods for Traditional Machine Learning Models Part 1: An Overview of Current Methods and Quantifying Their Disagreement. *arXiv*, pp. 1–22, 2022. URL http://arxiv.org/abs/2211.08943.

Joseph Fuller. Companies Need More Workers. Why Do They Reject Millions of Résumés? *The project on workforce*, 2021. URL https://www.pw.hks.harvard.edu/post/companies-need-more-workers-wsj.

Joseph B Fuller, Manjari Ramen, Eva Sage-gavin, and Kristen Hines. Hidden Workers: Untapped Talent. *Harvard Business School Project on Managing the Future of Work and Accenture*, 2021. URL https://www.pw.hks.harvard.edu/post/hidden-workers-untapped-talent.

Robert Geirhos, Roland S. Zimmermann, Blair Bilodeau, Wieland Brendel, and Been Kim. Don't trust your eyes: on the (un)reliability of feature visualizations. *arXiv*, 2023. URL http://arxiv.org/abs/2306.04719.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pp. 80–89, 2018. doi: 10.1109/DSAA.2018.00018.

Kurt Gödel. On Formally Undecidable Propositions of Principia Mathematica and Related Systems I. *Monatshefte für Mathematik*, 1931.

Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017. ISSN 07384602. doi: 10.1609/aimag.v38i3.2741.

Jeremy Gray. Did poincaré say "set theory is a disease"? *The Mathematical Intelligencer*, 13(1):19–22, 12 1991. ISSN 0343-6993. doi: 10.1007/BF03024067. URL https://maa.org/press/periodicals/convergence/quotations/poincare-jules-henri-1854-1912-0http://link.springer.com/10.1007/BF03024067.

Mara Graziani, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano, John O. Prior, Lode Lauwaert, Wessel Reijers, Adrien Depeursinge, Vincent Andrearczyk, and Henning Müller. *A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences*, volume 56. Springer Netherlands, 2023. ISBN 0123456789. doi: 10.1007/s10462-022-10256-8. URL https://doi.org/10.1007/s10462-022-10256-8.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local Rule-Based Explanations of Black Box Decision Systems. *arXiv*, 2018. URL http://arxiv.org/abs/1805.10820.

Nitish Gupta, Kevin Lin, Dan Roth, Sameer Singh, and Matt Gardner. Neural Module Networks for Reasoning over Text. In *International Conference on Learning Representations (ICLR)*, 12 2020. URL https://openreview.net/forum?id=SygWvAVFPr.

Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post Hoc Explanations. *Advances in Neural Information Processing Systems*, 35(NeurIPS), 2022. ISSN 10495258. URL http://arxiv.org/abs/2206.01254.

Peter Hase, Harry Xie, and Mohit Bansal. The Out-of-Distribution Problem in Explainability and Search Methods for Feature Importance Explanations. *Advances in Neural Information Processing Systems*, 5 (NeurIPS):3650–3666, 2021. ISSN 10495258.

Anna Hedström, Philine Bommer, Kristoffer K. Wickstrøm, Wojciech Samek, Sebastian Lapuschkin, and Marina M. C. Höhne. The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus. *Transactions on Machine Learning Research*, 2023. URL http://arxiv.org/abs/2302.07265.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. ISBN 9781950737901. doi: 10.18653/v1/D19-1275. URL https://www.aclweb.org/anthology/D19-1275.

David Hilbert. Über das Unendliche. *Mathematische Annalen*, 95(1):161–190, 12 1926. ISSN 0025-5831. doi: 10.1007/BF01206605. URL http://link.springer.com/10.1007/BF01206605.

David Hilbert. Hilbert's Radio Address to Society of German Scientists and Physicians, 1930. URL http://smith-at-sfsu.net/Documents/HilbertRadio/HilbertRadio.mp3.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://www.mitpressjournals.org/doi/abs/10.1162/neco.1997.9.8.1735.

Sara Hooker, Dumitru Erhan, Pieter-Jan Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 6 2019. URL http://arxiv.org/abs/1806.10758.

UK Government House of Lords. AI in the UK: Ready, Willing and Able?, 2017. URL https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10007.htm.

Shiyuan Huang, Siddarth Mamidanna, Shreedhar Jangam, Yilun Zhou, and Leilani H. Gilpin. Can Large Language Models Explain Themselves? A Study of LLM-Generated Self-Explanations. *arXiv*, 2023. URL http://arxiv.org/abs/2310.11207.

Aya Abdelsalam Ismail, Julius Adebayo, Hector Corrada Bravo, Stephen Ra, Kyunghyun Cho, Prescient Design, and Guide Labs. Concept Bottleneck Generative Models. In *The Twelfth International Conference on Learning Representations*, pp. 1–19, 2024. URL https://openreview.net/forum?id=L9U5MJJleF.

Alon Jacovi and Yoav Goldberg. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Stroudsburg, PA, USA, 4 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://www.aclweb.org/anthology/2020.acl-main.386.

Alon Jacovi and Yoav Goldberg. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310, 2021. ISSN 2307387X. doi: 10.1162/tacl{_}a{_}00367.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North*, volume 1, pp. 3543–3556, Stroudsburg, PA, USA, 2 2019. Association for Computational Linguistics. ISBN 9781950737130. doi: 10.18653/v1/N19-1357. URL http://aclweb.org/anthology/N19-1357.

Neil Jethani, Mukund Sudarshan, Yindalon Aphinyanaphongs, and Rajesh Ranganath. Have We Learned to Explain?: How Interpretability Methods Can Learn to Encode Predictions in their Interpretations. *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 130:1459–1467, 2021. ISSN 26403498.

Neil Jethani, Mukund Sudarshan, Ian Covert, Su In Lee, and Rajesh Ranganath. Fastshap: Real-Time Shapley Value Estimation. *ICLR 2022 - 10th International Conference on Learning Representations*, pp. 1–23, 2022.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv*, pp. 1–9, 2023. URL http://arxiv.org/abs/2310.06825.

José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and Understanding Recurrent Networks. *arXiv*, pp. 1–12, 6 2015. URL http://arxiv.org/abs/1506.02078.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–14, 2020. doi: 10.1145/3313831.3376219.

Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, and Pietro Liò. MEME: Generating RNN Model Explanations via Model Extraction. In *NeurIPS 2020 Workshop on Human And Model in the Loop Evaluation and Training Strategies*, 12 2020. ISBN 2012.06954v1. URL https://github.com/dmitrykazhdan/MEME-RNN-XAIhttp://arxiv.org/abs/2012.06954.

Been Kim. Beyond interpretability: developing a language to shape our relationships with AI. In *The International Conference on Learning Representations*, 2022. URL https://iclr.cc/Conferences/2022/Schedule?showEvent=7237.

Been Kim, Cynthia Rudin, and Julie Shah. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in Neural Information Processing Systems*, 3(January): 1952–1960, 2014. ISSN 10495258.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *35th International Conference on Machine Learning, ICML 2018*, 6:4186–4195, 11 2018. URL http://arxiv.org/abs/1711.11279.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of Saliency Methods. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11700 LNCS, pp. 267–280. Springer, 11 2019. doi: 10.1007/978-3-030-28954-6{_}14. URL http://link.springer.com/10.1007/978-3-030-28954-6_14.

Akhil Alfons Kodiyan. An overview of ethical issues in using AI systems in hiring with a case study of Amazon's AI based hiring tool. *Researchgate Preprint*, pp. 1–19, 2019.

Pang Wei Koh and Percy Liang. Understanding Black-box Predictions via Influence Functions. *34th International Conference on Machine Learning, ICML 2017*, 4:2976–2987, 3 2017. URL http://arxiv.org/abs/1703.04730.

Pang Wei Koh, Thao Nguye, Yew Siang Tang, Stephen Mussmann, Emma Pierso, Been Kim, and Percy Liang. Concept Bottleneck Models. *37th International Conference on Machine Learning, ICML 2020*, PartF16814:5294–5304, 2020.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv*, 2022. URL http://arxiv.org/abs/2202.01602.

Thomas S Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, 3rd editio edition, 1996. ISBN 978-0-226-45807-6.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An Evaluation of the Human-Interpretability of Explanation. *arXiv*, 1 2019. URL http://arxiv.org/abs/1902.00006.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & Explorable Approximations of Black Box Models. In *KDD'17, Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017. URL http://arxiv.org/abs/1707.01154.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring Faithfulness in Chain-of-Thought Reasoning. *arXiv*, 2023. URL http://arxiv.org/abs/2307.13702.

Piyawat Lertvittayakumjorn and Francesca Toni. Human-grounded Evaluations of Explanation Methods for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5194–5204, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. ISBN 9781950737901. doi: 10.18653/v1/D19-1523. URL https://www.aclweb.org/anthology/D19-1523.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding Neural Networks through Representation Erasure. *arXiv*, 2016. URL http://arxiv.org/abs/1612.08220.

Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 9 2018. ISSN 0001-0782. doi: 10.1145/3233231. URL https://dl.acm.org/doi/10.1145/3233231.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–158, 2012. doi: 10.1145/2339530.2339556.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Part F1288:623–631, 2013. doi: 10.1145/2487575.2487579.

Scott Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, pp. 4766–4775, 5 2017. URL http://arxiv.org/abs/1705.07874.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. Local Interpretations for Explainable Natural Language Processing: A Survey. *ACM Computing Surveys*, 56(9), 2024. ISSN 15577341. doi: 10.1145/3649450.

Minh Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, 2015. doi: 10.18653/v1/d15-1166.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards Faithful Model Explanation in NLP: A Survey. *Computational Linguistics*, 50(2):657–723, 6 2024. ISSN 0891-2017. doi: 10.1162/coli{_}a{_}00511. URL http://arxiv.org/abs/2209.11326https://direct.mit.edu/coli/article/50/2/657/119158/Towards-Faithful-Model-Explanation-in-NLP-A-Survey.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1731–1751, Abu Dhabi, United Arab Emirates, 12 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-emnlp.125.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Computing Surveys*, 55(8):1–42, 8 2022b. ISSN 0360-0300. doi: 10.1145/3546577. URL https://dl.acm.org/doi/10.1145/3546577.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from Large Language Models faithful? *The 62nd Annual Meeting of the Association for Computational Linguistics*, 1 2024a. URL https://openreview.net/forum?id=0fB5OROAIq.

Andreas Madsen, Siva Reddy, Sarath Chandar, and Anonymous. Faithfulness Measurable Masked Language Models. In *Forty-first International Conference on Machine Learning*, 2024b. URL https://openreview.net/forum?id=tw1PwpuAuNhttp://arxiv.org/abs/2310.07819.

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and Pitfalls of Black-Box Concept Learning Models. In *Proceeding at the International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*, volume 139, 2021. URL http://arxiv.org/abs/2106.13314.

Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do Concept Bottleneck Models Learn as Intended? In *XAI in Action: Past, Present, and Future Applications*, 2021. URL http://arxiv.org/abs/2105.04289.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. ISBN 9781950737482. doi: 10.18653/v1/P19-1334. URL https://www.aclweb.org/anthology/P19-1334.

Meta. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv*, 2023. URL http://arxiv.org/abs/2307.09288.

Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4), 2021. ISSN 21606463. doi: 10.1145/3387166.

Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable.* LeanPup, 2020. ISBN 9798411463330. URL https://christophm.github.io/interpretable-ml-book.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116(44):22071–22080, 10 2019. ISSN 10916490. doi: 10.1073/pnas.1900654116. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1900654116.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks. *Visualization for Deep Learning workshop at ICML*, 2016. URL http://arxiv.org/abs/1602.03616.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature Visualization. *Distill*, 2(11), 11 2017. ISSN 2476-0757. doi: 10.23915/distill.00007. URL https://distill.pub/2017/feature-visualization.

OpenAI. GPT-4 Technical Report. *OpenAI*, 4:1–100, 3 2023. URL http://arxiv.org/abs/2303.08774.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically Correcting Large Language Models: Surveying the Landscape of Diverse Automated Correction Strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024. ISSN 2307387X. doi: 10.1162/tacl{_}a{_}00660.

Letitia Parcalabescu and Anette Frank. On Measuring Faithfulness of Natural Language Explanations. *arXiv*, 2023. URL http://arxiv.org/abs/2311.07466.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15012–15032, Stroudsburg, PA, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.882. URL http://arxiv.org/abs/2402.13950https://aclanthology.org/2024.findings-emnlp.882.

Grusha Prasad, Yixin Nie, Mohit Bansal, Robin Jia, Douwe Kiela, and Adina Williams. To what extent do human explanations of model behavior align with actual model behavior? In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 1–14, Stroudsburg,

PA, USA, 2021. Association for Computational Linguistics. ISBN 9781955917063. doi: 10.18653/v1/2021.blackboxnlp-1.1. URL https://aclanthology.org/2021.blackboxnlp-1.1.

Kristina Preuer, Günter Klambauer, Friedrich Rippmann, Sepp Hochreiter, and Thomas Unterthiner. Interpretable deep learning in drug discovery. *Explainable AI: interpreting, explaining and visualizing deep learning*, pp. 331–345, 2019.

Kurt Werner Friedrich Reidemeister. *Hilbert: Gedenkband*. Springer, 1971. ISBN 978-3540052920.

Carl O. Retzlaff, Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Röttger, Heimo Müller, and Andreas Holzinger. Post-hoc vs ante-hoc explanations: xAI design guidelines for data scientists. *Cognitive Systems Research*, 86(June 2023):101243, 8 2024. ISSN 13890417. doi: 10.1016/j.cogsys.2024.101243. URL https://doi.org/10.1016/j.cogsys.2024.101243https://linkinghub.elsevier.com/retrieve/pii/S1389041724000378.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17-Augu, pp. 1135–1144, New York, NY, USA, 8 2016. ACM. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://dl.acm.org/doi/10.1145/2939672.2939778.

Marko Robnik-Šikonja and Marko Bohanec. *Perturbation-Based Explanations of Prediction Models*. Springer International Publishing, 2018. ISBN 9783319904030. doi: 10.1007/978-3-319-90403-0{_}9. URL http://dx.doi.org/10.1007/978-3-319-90403-0_9.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 12 2020. ISSN 2307-387X. doi: 10.1162/tacl{_}a{_}00349. URL https://direct.mit.edu/tacl/article/96482.

Yao Rong, Tobias Leemann, Thai Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4): 2104–2122, 2024. ISSN 19393539. doi: 10.1109/TPAMI.2023.3331846.

Cynthia 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. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x. URL http://www.nature.com/articles/s42256-019-0048-x.

Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Muller. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 11 2017. ISSN 2162-237X. doi: 10.1109/TNNLS.2016.2599820. URL https://ieeexplore.ieee.org/document/7552539/.

Amit Sangroya, Mouli Rastogi, C Anantaram, and Lovekesh Vig. Guided-LIME: Structured sampling based hybrid approach towards explaining blackbox machine learning models. In *CEUR Workshop Proceedings*, volume 2699, 2020.

Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. Human Interpretation of Saliency-based Explanation Over Text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 611–636, New York, NY, USA, 6 2022a. ACM. ISBN 9781450393522. doi: 10.1145/3531146.3533127. URL https://dl.acm.org/doi/10.1145/3531146.3533127.

Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. Human Interpretation of Saliency-based Explanation Over Text. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 611–636, New York, NY, USA, 6 2022b. ACM. ISBN 9781450393522. doi: 10.1145/3531146.3533127. URL https://dl.acm.org/doi/10.1145/3531146.3533127.

Daniel Schug, Sai Yerramreddy, Rich Caruana, Craig Greenberg, and Justyna P. Zwolak. Extending Explainable Boosting Machines to Scientific Image Data. In *Machine Learning and the Physical Sciences Workshop, NeurIPS*, 2023. URL http://arxiv.org/abs/2305.16526.

Lisa Schut, Nenad Tomasev, Tom McGrath, Demis Hassabis, Ulrich Paquet, and Been Kim. Bridging the Human-AI Knowledge Gap: Concept Discovery and Transfer in AlphaZero. *arXiv*, pp. 1–61, 10 2023. URL http://arxiv.org/abs/2310.16410.

Gesina Schwalbe and Bettina Finzel. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5): 3043–3101, 2024. ISSN 1573756X. doi: 10.1007/s10618-022-00867-8. URL https://doi.org/10.1007/s10618-022-00867-8.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 10 2020. ISSN 15731405. doi: 10.1007/s11263-019-01228-7. URL http://arxiv.org/abs/1610.02391http://dx.doi.org/10.1007/s11263-019-01228-7.

Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. Human Attention Maps for Text Classification: Do Humans and Neural Networks Focus on the Same Words? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4596–4608, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.419. URL https://www.aclweb.org/anthology/2020.acl-main.419.

Junghoon Seo, Jeongyeol Choe, Jamyoung Koo, Seunghyeon Jeon, Beomsu Kim, and Taegyun Jeon. Noise-adding Methods of Saliency Map as Series of Higher Order Partial Derivative. In *2018 ICML Workshop on Human Interpretability in Machine Learning*, 6 2018. URL http://arxiv.org/abs/1806.03000.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pp. 4844–4866, 2017. ISBN 9781510855144. URL https://arxiv.org/.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling LIME and SHAP. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, New York, NY, USA, 2 2020. ACM. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://dl.acm.org/doi/10.1145/3375627.3375830.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: removing noise by adding noise. *ICML workshop on visualization for deep learning*, 6 2017. ISSN 23318422. URL http://arxiv.org/abs/1706.03825.

James Smith. David Hilbert's Radio Address. *Convergence*, 2014. doi: 10.4169/convergence20140202. URL https://old.maa.org/press/periodicals/convergence/david-hilberts-radio-address-hilbert-and-mathematical-inquiry.

Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014. ISSN 02193116. doi: 10.1007/s10115-013-0679-x.

Sanjay Subramanian, Ben Bogin, Nitish Gupta, Tomer Wolfson, Sameer Singh, Jonathan Berant, and Matt Gardner. Obtaining faithful interpretations from compositional neural networks. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 5594–5608, 2020. ISSN 0736587X. doi: 10.18653/v1/2020.acl-main.495. URL https://www.aclweb.org/anthology/2020.acl-main.495.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pp. 5109–5118, 3 2017. ISBN 9781510855144. URL http://arxiv.org/abs/1703.01365.

Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. On the Hardness of Faithful Chain-of-Thought Reasoning in Large Language Models. *arXiv*, 6 2024. URL http://arxiv.org/abs/2406.10625.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601, Stroudsburg, PA, USA, 5 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1452. URL https://www.aclweb.org/anthology/P19-1452.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Generating Token-Level Explanations for Natural Language Inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, volume 1, pp. 963–969, Stroudsburg, PA, USA, 2019. Association for Computational Linguistics. ISBN 9781950737130. doi: 10.18653/v1/N19-1101. URL http://aclweb.org/anthology/N19-1101.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*, pp. 1–32, 5 2023. URL http://arxiv.org/abs/2305.04388https://openreview.net/forum?id=bzs4uPLXvi.

Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. Rationales for Sequential Predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 10314–10332, Stroudsburg, PA, USA, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.807. URL https://aclanthology.org/2021.emnlp-main.807.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pp. 5999–6009. Association for Computational Linguistics (ACL), 6 2017. ISBN 9781941643327. URL http://arxiv.org/abs/1706.03762.

Elena Voita and Ivan Titov. Information-Theoretic Probing with Minimum Description Length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, Stroudsburg, PA, USA, 3 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.14. URL https://www.aclweb.org/anthology/2020.emnlp-main.14.

Yipei Wang and Xiaoqian Wang. A Unified Study of Machine Learning Explanation Evaluation Metrics. *arXiv*, 2022. URL http://arxiv.org/abs/2203.14265.

Alice Xiang and Inioluwa Deborah Raji. On the Legal Compatibility of Fairness Definitions. *Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems*, 2019. URL http://arxiv.org/abs/1912.00761.

Wenqian Ye, Fei Xu, Yaojia Huang, Cassie Huang, and Ji A. Adversarial Examples Generation for Reducing Implicit Gender Bias in Pre-trained Models. *arXiv*, 2021. URL http://arxiv.org/abs/2110.01094.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, Pradeep K Ravikumar, Arun Sai Suggala, David I Inouye, and Pradeep K Ravikumar. On the (In)fidelity and Sensitivity of Explanations. In H Wallach, H Larochelle, A Beygelzimer, F d\textquotesingle Alché-Buc, E Fox, and R Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 10967–10978. Curran Associates, Inc., Vancouver, Canada, 2019. URL https://arxiv.org/abs/1901.09392.

Jinsung Yoon, James Jordon, and Mihaela Van Der Schaar. Invase: Instance-wise variable selection using neural networks. *7th International Conference on Learning Representations, ICLR 2019*, pp. 1–24, 2019.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding Neural Networks Through Deep Visualization. In *Deep Learning Workshop at 31st International Conference on Machine Learning*, 2015. URL http://arxiv.org/abs/1506.06579.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. *Advances in Neural Information Processing Systems*, 35(3), 2022. ISSN 10495258.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8689 LNCS(PART 1):818–833, 2014. ISSN 16113349. doi: 10.1007/978-3-319-10590-1{_}53.

Kelly Zhang and Samuel Bowman. Language Modeling Teaches You More than Translation Does: Lessons Learned Through Auxiliary Syntactic Task Analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 359–361, Stroudsburg, PA, USA, 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5448. URL `http://aclweb.org/anthology/W18-5448`.

Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics (Switzerland)*, 10(5):1–19, 2021. ISSN 20799292. doi: 10.3390/electronics10050593.

Yilun Zhou and Julie Shah. The Solvability of Interpretability Evaluation Metrics. In *Findings of the Association for Computational Linguistics: EACL*, 2023. URL `http://arxiv.org/abs/2205.08696`.

Jan Ruben Zilke, Eneldo Loza Mencía, and Frederik Janssen. DeepRED – Rule Extraction from Deep Neural Networks. In Toon Calders, Michelangelo Ceci, and Donato Malerba (eds.), *Discovery Science*, pp. 457–473. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46307-0. doi: 10.1007/978-3-319-46307-0{_}29. URL `https://link.springer.com/10.1007/978-3-319-46307-0_29`.

Roland S. Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas S.A. Wallis, and Wieland Brendel. How Well do Feature Visualizations Support Causal Understanding of CNN Activations? *Advances in Neural Information Processing Systems*, 14(NeurIPS):11730–11744, 2021. ISSN 10495258.