Towards Faithfully Interpretable NLP Systems:
How Should We Define and Evaluate Faithfulness?

Alon Jacovi
Bar Ilan University
alonjacovi@gmail.com

Yoav Goldberg
Bar Ilan University and Allen Institute for AI
yoav.goldberg@gmail.com

Abstract

With the growing popularity of deep-learning based NLP models, comes a need for interpretable systems. But what is interpretability, and what constitutes a high-quality interpretation? In this opinion piece we reflect on the current state of interpretability evaluation research. We call for more clearly differentiating between different desired criteria an interpretation should satisfy, and focus on the faithfulness criteria. We survey the literature with respect to faithfulness evaluation, and arrange the current approaches around three assumptions, providing an explicit form to how faithfulness is "defined" by the community. We provide concrete guidelines on how evaluation of interpretation methods should and should not be conducted. Finally, we claim that the current binary definition for faithfulness sets a potentially unrealistic bar for being considered faithful. We call for discarding the binary notion of faithfulness in favor of a more graded one, which we believe will be of greater practical utility.

1 Introduction

Fueled by recent advances in deep-learning and language processing, NLP systems are increasingly being used for prediction and decision-making in many fields (Vig and Belinkov, 2019), including sensitive ones such as health, commerce and law (Fort and Couillault, 2016). Unfortunately, these highly flexible and highly effective neural models are also opaque. There is therefore a critical need for explaining learning-based models' decisions.

The emerging research topic of *interpretability* or *explainability*[1] has grown rapidly in recent years. Unfortunately, not without growing pains.

---

[1]Despite fine-grained distinctions between the terms, within the scope of this work we use the terms "interpretability" and "explainability" interchangeably.

One such pain is the challenge of defining—and evaluating—what constitutes a quality interpretation. Current approaches define interpretation in a rather ad-hoc manner, motivated by practical use-cases and applications. However, this view often fails to distinguish between distinct aspects of the interpretation's quality, such as readability, plausibility and faithfulness (Herman, 2017).[2] We argue (§2, §5) such conflation is harmful, and that faithfulness should be defined and evaluated *explicitly*, and independently from plausibility.

Our main focus is the *evaluation of the faithfulness* of an explanation: a faithful interpretation is one that accurately represents the reasoning process behind the model's prediction. We find this to be a pressing issue: in cases where an explanation is required to be faithful, imperfect or misleading evaluation can have disastrous effects.

While literature in this area may implicitly or explicitly evaluate faithfulness for specific explanation techniques, there is no consistent and formal definition of faithfulness. We uncover three assumptions that underlie all these attempts. By making the assumptions explicit and organizing the literature around them, we "connect the dots" between seemingly distinct evaluation methods, and

---

[2]Unfortunately, the terms in the literature are not yet standardized, and vary widely. "Readability" and "plausibility" are also referred to as "human-interpretability" and "persuasiveness", respectively (e.g., Lage et al. (2019); Herman (2017)). To our knowledge, the term "faithful interpretability" was coined in Harrington et al. (1985), reinforced by Ribeiro et al. (2016), and is, we believe, most commonly used (e.g., Gilpin et al. (2018); Wu and Mooney (2018); Lakkaraju et al. (2019)). Chakraborty et al. (2017) refers to this issue (more or less) as "accountability". Sometimes referred to as how "trustworthy" (Camburu et al., 2019) or "descriptive" (Carmona et al., 2015; Biecek, 2018) the interpretation is, or as "descriptive accuracy" (Murdoch et al., 2019). Also related to the "transparency" (Baan et al., 2019), the "fidelity" (Guidotti et al., 2018) or the "robustness" (Alvarez-Melis and Jaakkola, 2018) of the interpretation method. And frequently, simply "explainability" is inferred to require faithfulness by default.

also provide a basis for discussion regarding the desirable properties of faithfulness (§6).

Finally, we observe a trend by which faithfulness is treated as a binary property, followed by showing that an interpretation method is not faithful. We claim that this is unproductive (§7), as the assumptions are nearly impossible to satisfy fully, and it is all too easy to disprove the faithfulness of an interpretation method via a counter-example. What can be done? We argue for a more practical view of faithfulness, calling for a *graded criteria* that measures *the extent and likelihood* of an interpretation to be faithful, *in practice* (§8). While we started to work in this area, we pose the exact formalization of these criteria, and concrete evaluations methods for them, as a central challenge to the community for the coming future.

2   Faithfulness vs. Plausibility

There is considerable research effort in attempting to define and categorize the desiderata of a learned system's interpretation, most of which revolves around specific use-cases (Lipton, 2018; Guidotti et al., 2018, inter alia).

Two particularly notable criteria, each useful for a different purposes, are *plausibility* and *faithfulness*. "Plausibility" refers to how convincing the interpretation is to humans, while "faithfulness" refers to how accurately it reflects the true reasoning process of the model (Herman, 2017; Wiegreffe and Pinter, 2019). Naturally, it is possible to satisfy one of these properties without the other. For example, consider the case of interpretation via post-hoc text generation—where an additional "generator" component outputs a textual explanation of the model's decision, and the generator is learned with supervision of textual explanations (Zaidan and Eisner, 2008; Rajani et al., 2019; Strout et al., 2019). In this case, plausibility is the dominating property, while there is no faithfulness guarantee.

Despite the difference between the two criteria, many authors do not clearly make the distinction, and sometimes conflate the two.[3] Moreoever, the majority of works do not explicitly name the criteria under consideration, even when they clearly belong to one camp or the other.[4]

We argue that this conflation is dangerous. For example, consider the case of *recidivism prediction*,

where a judge is exposed to a model's prediction and its interpretation, and the judge believes the interpretation to reflect the model's reasoning process. Since the interpretation's faithfulness carries legal consequences, a *plausible* but *unfaithful* interpretation may be the worst-case scenario. The lack of explicit claims by research may cause misinformation to potential users of the technology, who are not versed in its inner workings.[5] Therefore, clear distinction between these terms is critical.

3   Inherently Interpretable?

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 (2018) argues in favor of *inherently interpretable models*, which by design claim to provide more faithful interpretations than post-hoc interpretation of black-box models.

We warn against taking this argumentation at face-value: a method being "inherently interpretable" is merely a claim that needs to be verified before it can be trusted. Indeed, while *attention mechanisms* have been considered as "inherently interpretable" (Ghaeini et al., 2018; Lee et al., 2017), recent work cast doubt regarding their faithfulness (Serrano and Smith, 2019; Jain and Wallace, 2019; Wiegreffe and Pinter, 2019).

4   Evaluation via Utility

While explanations have many different use-cases, such as model debugging, lawful guarantees or health-critical guarantees, one other possible use-case with prominent evaluation literature is Intelligent User Interfaces (IUI), via Human-Computer Interaction (HCI), of automatic models assisting human decision-makers. The goal of the explanation here is to increase the degree of trust between the user and the system, giving the user more nuance towards whether the system's decision is likely correct, or not. In the general case, the final evaluation metric is the performance of the user at their task (Abdul et al., 2018). For example, Feng and Boyd-Graber (2019) evaluate various explanations of a model in a setting of trivia question answering.

However, in the context of faithfulness, we must warn against HCI-inspired evaluation, as well: **increased performance in this setting is not in-**

---

[3]E.g., Lundberg and Lee (2017); Pörner et al. (2018); Wu and Mooney (2018).

[4] E.g., Mohseni and Ragan (2018); Arras et al. (2016); Xiong et al. (2018); Weerts et al. (2019).

[5]As Kaur et al. (2019) concretely show, even experts are prone to overly trust the faithfulness of explanations, despite no guarantee.

dicative of faithfulness; rather, it is indicative of correlation between the plausibility of the explanations and the model's performance.

To illustrate, consider the following fictional case of a non-faithful explanation system, in an HCI evaluation setting: the explanation given is a heat-map of the textual input, attributing scores to various tokens. Assume the system explanations behave in the following way: when the output is *correct*, the explanation consists of random content words; and when the output is *incorrect*, it consists of random punctuation marks. In other words, the explanation is more likely to appear plausible when the model is correct, while at the same time not reflecting the true decision process of the model. The user, convinced by the nicer-looking explanations, performs better using this system. However, the explanation consistently claimed random tokens to be highly relevant to the model's reasoning process. While the system is concretely useful, the claims given by the explanation do not reflect the model's decisions whatsoever (by design).

While the above scenario is extreme, this misunderstanding is not entirely unlikely, since *any* degree of correlation between plausibility and model performance will result in increased user performance, regardless of any notion of faithfulness.

5 Guidelines for Evaluating Faithfulness

We propose the following guidelines for evaluating the faithfulness of explanations. These guidelines address common pitfalls and sub-optimal practices we observed in the literature.

Be explicit in what you evaluate. Conflating plausibility and faithfulness is harmful. You should be explicit on which one of them you evaluate, and use suitable methodologies for each one. Of course, the same applies when designing interpretation techniques—be clear about which properties are being prioritized.

Faithfulness evaluation should not involve human-judgement on the quality of interpretation. We note that: (1) humans cannot judge if an interpretation is faithful or not: if they understood the model, interpretation would be unnecessary; (2) for similar reasons, we cannot obtain supervision for this problem, either. Therefore, human judgement should not be involved in evaluation for faithfulness, as human judgement measures plausability.

Faithfulness evaluation should not involve human-provided gold labels. We should be able to interpret incorrect model predictions, just the same as correct ones. Evaluation methods that rely on gold labels are influenced by human priors on what *should* the model do, and again push the evaluation in the direction of plausability.

Do not trust "inherent interpretability" claims. Inherent interpretability is a claim until proven otherwise. Explanations provided by "inherently interpretable" models must be held to the same standards as post-hoc interpretation methods, and be evaluated for faithfulness using the same set of evaluation techniques.

Faithfulness evaluation of IUI systems should not rely on user performance. End-task user performance in HCI settings is merely indicative of correlation between plausibility and model performance, however small this correlation is. While important to evaluate the utility of the interpretations for some use-cases, it is unrelated to faithfulness.

6 Defining Faithfulness

What does it mean for an interpretation method to be faithful? Intuitively, we would like the provided interpretation to reflect the true reasoning process of the model when making a decision. But what is a reasoning process of a model, and how can reasoning processes be compared to each other?

Lacking a standard definition, different works evaluate their methods by introducing tests to measure properties that they believe good interpretations should satisfy. Some of these tests measure aspects of faithfulness. These ad-hoc definitions are often unique to each paper and inconsistent with each other, making it hard to find commonalities.

We uncover *three assumptions* that underlie all these methods, enabling us to organize the literature along standardized axes, and relate seemingly distinct lines of work. Moreover, exposing the underlying assumptions enables an informed discussion regarding their validity and merit (we leave such a discussion for future work, by us or others).

These assumptions, to our knowledge, encapsulate the current working definitions of faithfulness used by the research community.

Assumption 1 (*The Model Assumption*). *Two models will make the same predictions if and only if they use the same reasoning process.*

4200

Corollary 1.1. *An interpretation system is unfaithful if it results in different interpretations of models that make the same decisions.*

As demonstrated by a recent example concerning NLP models, it can be used for proof by counter-example. Theoretically, if all possible models which can perfectly mimic the model's decisions also provide the same interpretations, then they could be deemed faithful. Conversely, showing that two models provide the same results but different interpretations, disprove the faithfulness of the method. Wiegreffe and Pinter (2019) show how these counter-examples can be derived with adversarial training of models which can mimic the original model, yet provide different explanations.[6]

Corollary 1.2. *An interpretation is unfaithful if it results in different decisions than the model it interprets.*

A more direct application of the *Model Assumption* is via the notion of *fidelity* (Guidotti et al., 2018; Lakkaraju et al., 2019). For cases in which the explanation is itself a model capable of making decisions (e.g., decision trees or rule lists (Sushil et al., 2018)), *fidelity* is defined as the degree to which the explanation model can mimic the original model's decisions (as an accuracy score). For cases where the explanation is *not* a computable model, Doshi-Velez and Kim (2017) propose a simple way of mapping explanations to decisions via crowd-sourcing, by asking humans to simulate the model's decision without any access to the model, and only access to the input and explanation (termed *forward simulation*). This idea is further explored and used in practice by Nguyen (2018).

Assumption 2 (*The Prediction Assumption*). *On similar inputs, the model makes similar decisions if and only if its reasoning is similar.*

Corollary 2. *An interpretation system is unfaithful if it provides different interpretations for similar inputs and outputs.*

Since the interpretation serves as a proxy for the model's "reasoning", it should satisfy the same constraints. In other words, interpretations of similar decisions should be similar, and interpretations of dissimilar decisions should be dissimilar.

This assumption is more useful to *disprove* the faithfulness of an interpretation rather than prove it, since a disproof requires finding appropriate cases

where the assumption doesn't hold, where a proof would require checking a (very large) satisfactory quantity of examples, or even the entire input space.

One recent discussion in the NLP community (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019) concerns the use of this underlying assumption for evaluating attention heat-maps as explanations. The former attempts to provide different explanations of similar decisions *per instance*. The latter critiques the former and is based more heavily on the *model assumption*, described above.

Additionally, Kindermans et al. (2019) propose to introduce a constant shift to the input space, and evaluate whether the explanation changes significantly as the final decision stays the same. Alvarez-Melis and Jaakkola (2018) formalize a generalization of this technique under the term *interpretability robustness*: interpretations should be invariant to small perturbations in the input (a direct consequence of the *prediction assumption*). Wolf et al. (2019) further expand on this notion as "consistency of the explanation with respect to the model". Unfortunately, robustness measures are difficult to apply in NLP settings due to the discrete input.

Assumption 3 (*The Linearity Assumption*).[7] *Certain parts of the input are more important to the model reasoning than others. Moreover, the contributions of different parts of the input are independent from each other.*

Corollary 3. *Under certain circumstances, heat-map interpretations can be faithful.*

This assumption is employed by methods that consider heat-maps[8] (e.g., attention maps) over the input as explanations, particularly popular in NLP. Heat-maps are *claims* about which parts of the input are more relevant than others to the model's decision. As such, we can design "stress tests" to verify whether they uphold their claims.

One method proposed to do so is *erasure*, where the "most relevant" parts of the input—according to the explanation—are erased from the input, in expectation that the model's decision will change (Arras et al., 2016; Feng et al., 2018; Serrano and Smith, 2019). Otherwise, the "least relevant" parts of the input may be erased, in expectation that the model's decision will not change (Jacovi et al.,

---

[6]We note that in context, Wiegreffe and Pinter also utilize the model assumption to show that some explanations do carry useful information on the model's behavior.

[7]This assumption has gone through justified scrutiny in recent work. As mentioned previously, we do not necessarily endorse it. Nevertheless, it is used in parts of the literature.

[8]Also referred to as feature-attribution explanations (Kim et al., 2017).

2018). Yu et al. (2019); DeYoung et al. (2019) propose two measures of *comprehensiveness* and *sufficiency* as a formal generalization of erasure: as the degree by which the model is influenced by the removal of the high-ranking features, or by inclusion of solely the high-ranking features.

7 Is Faithful Interpretation Impossible?

The aforementioned assumptions are currently utilized to evaluate faithfulness in a binary manner: whether an interpretation is strictly faithful or not. Specifically, they are most often used to show that a method is *not* faithful, by constructing cases in which the assumptions do not hold for it.[9] In other words, **there is a clear trend of proof via counter-example, for various interpretation methods, that they are not globally faithful.**

We claim that this is unproductive, as we expect these various methods to consistently result in negative (not faithful) results, continuing the current trend. This follows because an interpretation functions as an *approximation* of the model or decision's true reasoning process, so it by definition loses information. By the pigeonhole principle, there will be inputs with deviation between interpretation and reasoning.

This is observed in practice, in numerous work that show adversarial behavior, or pathological behaviours, that arise from the deeply non-linear and high-dimensional decision boundaries of current models.[10] Furthermore, because we lack supervision regarding which models or decisions are indeed mappable to human-readable concepts, we cannot ignore the approximation errors.

This poses a high bar for explanation methods to fulfill, a bar which we estimate will not be overcome soon, if at all. What should we do, then, if we desire a system that provides faithful explanations?

8 Towards Better Faithfulness Criteria

We argue that a way out of this standstill is in a more practical and nuanced methodology for defining and evaluating faithfulness. We propose the following challenge to the community: **We must develop formal definition and evaluation for faith-**

fulness that allows us the freedom to say when a method is *sufficiently faithful* to be useful in practice.

We note two possible approaches to this end:

1. **Across models and tasks:** The *degree* (as grayscale) of faithfulness at the level of specific models or tasks. Perhaps some models or tasks allow sufficiently faithful interpretation, even if that is not true for others.[11] For example, the method may not be faithful for some question-answering task, but faithful for review sentiment, perhaps based on various syntactic and semantic attributes of those tasks.

2. **Across input space:** The degree of faithfulness at the level of subspaces of the input space, such as neighborhoods of similar inputs, or singular inputs themselves. If we are able to say with some degree of confidence whether a specific decision's explanation is faithful to the model, even if the interpretation method is not considered universally faithful, it can be used with respect to those specific areas or instances only.

9 Conclusion

The opinion proposed in this paper is two-fold:

First, interpretability evaluation often conflates evaluating faithfulness and plausibility together. We should tease apart the two definitions and focus solely on evaluating faithfulness without any influence of the convincing power of the interpretation.

Second, faithfulness is often evaluated in a binary "faithful or not faithful" manner, and we believe strictly faithful interpretation is a "unicorn" which will likely never be found. We should instead evaluate faithfulness on a more nuanced "grayscale" that allows interpretations to be useful even if they are not globally and definitively faithful.

Acknowledgements

---

[9]Whether for attention (Baan et al., 2019; Pruthi et al., 2019; Jain and Wallace, 2019; Serrano and Smith, 2019; Wiegreffe and Pinter, 2019), saliency methods (Alvarez-Melis and Jaakkola, 2018; Kindermans et al., 2019), or others (Ghorbani et al., 2019; Feng et al., 2018).

[10]Kim et al. (2017); Feng et al. (2018, §6) discuss this point in the context of heat-map explanations.

[11]As noted by Wiegreffe and Pinter (2019); Vashishth et al. (2019), although in the context of attention solely.

References

Ashraf M. Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan S. Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, page 582. ACM.

David Alvarez-Melis and Tommi S. Jaakkola. 2018. On the robustness of interpretability methods. *CoRR*, abs/1806.08049.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2016. "what is relevant in a text document?": An interpretable machine learning approach. *CoRR*, abs/1612.07843.

Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Do transformer attention heads provide transparency in abstractive summarization? *CoRR*, abs/1907.00570.

Przemyslaw Biecek. 2018. DALEX: explainers for complex predictive models in R. *J. Mach. Learn. Res.*, 19:84:1–84:5.

Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. 2019. Can i trust the explainer? verifying post-hoc explanatory methods.

Vicente Iván Sánchez Carmona, Tim Rocktäschel, Sebastian Riedel, and Sameer Singh. 2015. Towards extracting faithful and descriptive representations of latent variable models. In *AAAI Spring Symposia*.

Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. 2017. Interpretability of deep learning models: a survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1–6. IEEE.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Shi Feng and Jordan Boyd-Graber. 2019. What can ai do for me? evaluating machine learning interpretations in cooperative play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI 19, page 229239, New York, NY, USA. Association for Computing Machinery.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan L. Boyd-Graber. 2018. Pathologies of neural models make interpretation difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3719–3728. Association for Computational Linguistics.

Karën Fort and Alain Couillault. 2016. Yes, we care! results of the ethics and natural language processing surveys. In *LREC*.

Reza Ghaeini, Xiaoli Z. Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. *CoRR*, abs/1808.03894.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5):93:1–93:42.

L.A. Harrington, M.D. Morley, A. Ščedrov, and S.G. Simpson. 1985. *Harvey Friedman's Research on the Foundations of Mathematics*. Studies in Logic and the Foundations of Mathematics. Elsevier Science.

Bernease Herman. 2017. The promise and peril of human evaluation for model interpretability. *CoRR*, abs/1711.07414. Withdrawn.

Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2018. Understanding convolutional neural networks for text classification. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–65.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics.

Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna M. Wallach, and Jennifer Wortman

Vaughan. 2019. Interpreting interpretability: Understanding data scientists use of interpretability tools for machine learning.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un)reliability of saliency methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, volume 11700 of *Lecture Notes in Computer Science*, pages 267–280. Springer.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. *CoRR*, abs/1902.00006.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 131–138. ACM.

Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. 2017. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability. *Commun. ACM*, 61(10):36–43.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4765–4774.

Sina Mohseni and Eric D. Ragan. 2018. A human-grounded evaluation benchmark for local explanations of machine learning. *CoRR*, abs/1801.05075.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Interpretable machine learning: definitions, methods, and applications. *ArXiv*, abs/1901.04592.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

Nina Pörner, Hinrich Schütze, and Benjamin Roth. 2018. Evaluating neural network explanation methods using hybrid documents and morphological prediction. *CoRR*, abs/1801.06422.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2019. Learning to deceive with attention-based explanations. *CoRR*, abs/1909.07913.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. *CoRR*, abs/1906.02361.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the predictions of any classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. ACM.

Cynthia Rudin. 2018. Please stop explaining black box models for high stakes decisions. *CoRR*, abs/1811.10154.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2931–2951.

Julia Strout, Ye Zhang, and Raymond J. Mooney. 2019. Do human rationales improve machine explanations? *CoRR*, abs/1905.13714.

Madhumita Sushil, Simon Šuster, and Walter Daelemans. 2018. Rule induction for global explanation of trained models. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 82–97, Brussels, Belgium. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across NLP tasks. *CoRR*, abs/1909.11218.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *CoRR*, abs/1906.04284.

Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A human-grounded evaluation of SHAP for alert processing. *CoRR*, abs/1907.03324.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 2019, Hong Kong, China, November 3-7, 2019, pages 11–20. Association for Computational Linguistics.

Lior Wolf, Tomer Galanti, and Tamir Hazan. 2019. A formal approach to explainability. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 255–261. ACM.

Jialin Wu and Raymond J. Mooney. 2018. Faithful multimodal explanation for visual question answering. *CoRR*, abs/1809.02805.

Wenting Xiong, Iftitahu Ni'mah, Juan M. G. Huesca, Werner van Ipenburg, Jan Veldsink, and Mykola Pechenizkiy. 2018. Looking deeper into deep learning model: Attribution-based explanations of textcnn. *CoRR*, abs/1811.03970.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi S. Jaakkola. 2019. Rethinking cooperative rationalization: Introspective extraction and complement control. *CoRR*, abs/1910.13294.

Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 31–40, Honolulu, Hawaii. Association for Computational Linguistics.