Explainable Artificial Intelligence: A Survey

Filip Karlo Došilović*, Mario Brčić** and Nikica Hlupić**
* Student at University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
** University of Zagreb Faculty of Electrical Engineering and Computing, Zagreb, Croatia
fillip-karlo.dosilovic@fer.hr, mario.brcic@fer.hr, nikica.hlupic@fer.hr

***Abstract* - In the last decade, with availability of large datasets and more computing power, machine learning systems have achieved (super)human performance in a wide variety of tasks. Examples of this rapid development can be seen in image recognition, speech analysis, strategic game planning and many more. The problem with many state-of-the-art models is a lack of transparency and interpretability. The lack of thereof is a major drawback in many applications, e.g. healthcare and finance, where rationale for model's decision is a requirement for trust. In the light of these issues, explainable artificial intelligence (XAI) has become an area of interest in research community. This paper summarizes recent developments in XAI in supervised learning, starts a discussion on its connection with artificial general intelligence, and gives proposals for further research directions.**

Keywords* - *explainable artificial intelligence; interpretability; explainability; comprehensibility

I. INTRODUCTION

In the last decade, especially since 2012, artificial intelligence (AI) and machine learning (ML) systems have achieved (super)human performance in many tasks that were previously thought to be computationally unattainable [1]. Advances in the field were achieved due to the rise of available information, and major hardware improvements combined with new optimization algorithms. We can also attribute these advancements to high-quality open-source libraries which allowed developers and researchers to quickly code and test models. Improvements in speech recognition, image classification, object detection, classical (board) games, Texas Hol'em, and many more have led to their proliferation and percolation to the real-world applications outside research labs, mostly in the area of supervised learning.

We also saw advances in application critical areas, e.g. medicine, finance, self-driving cars, government, bioinformatics, churn prediction and content recommendation but these applications also brought attention to the crucial trust-related problems. Future applications will, in addition to extensions in aforementioned areas, also include cognitive assistance, interpretable science and reliable ML [2], [3] The all-pervasive utilization of these systems will significantly transform the social landscape of the world. These changes include many ethical challenges to which society will have to adapt fast in order to steer the developments to the favorable directions and outcomes. Automation will significantly change the job market [2], which may lead to more unfair wealth distribution. Content recommendation [4] and generation of fake content [5] coupled with other technologies will deeply impact the social dynamics [6]. Many things will be prescribed by such algorithms and that will affect human lives in ways maybe now unimagined so people will need to trust them in order to accept those prescriptions. Systems must, like humans, satisfy many criteria (assurances) in order to boost trust [2]: unbiasedness (fairness), reliability, safety, explanatory justifiability, privacy, usability etc. Among humans such assurances are assumed due to bias towards human decision making, since people are social creatures accustomed to life in human communities. Artificial intelligence, due to its novel status in our lives, as well as being of the human making, causes much skepticism - rightfully so. Deep learning, as one of the most successful machine learning approaches in supervised learning has been criticized in [7] for working well as approximations where answers often not to be fully trusted, pointing out to model vulnerabilities in language and vision models. Spoofability and biasedness have been demonstrated for visual recognition in [8], [9] and natural language processing [10], [11]. No robust solution has been found to these problems so far. The potential ethical pitfalls should be addressed as soon as possible since inactivity could lead to unforeseeable splits and differences in the future society. European Union introduced a right to explanation in General Data Protection Regulation (GDPR) [12] as an attempt to remedy the potential problems with the rising importance of algorithms. Since the aforementioned trust criteria are hard to formalize and quantify, usually criteria of interpretability and explainability are used as intermediate goals. Afterwards in the following stage, system's explanations can be checked if they satisfy other desirable trust criteria.

More generally, abstracted explanations can be utilized for finding useful properties and generating hypotheses about data-generating processes, such as causal relationships – which is crucial application in science as well as in future Artificial General Intelligence (AGI). Generated hypotheses can be basis for further automated or manual experimentation, knowledge discovery, and optimization. This view is supported by [13] and encompasses: checking for satisfaction of trust-criteria, optimization of ethical outcomes due to technology, assisted(automated) scientific discovery, transferring skills, etc. mentioned in[3], [14]–[16]. Previous overviews and surveys of interpretability in machine learning are given in [2], [3], [14], [15], [17]–[19].

In this paper we survey the advances in the interpretability and explainability of machine learning models under the supervised learning paradigm. Much of the recent work is in the area of deep learning, due to remarkable performance gains of these models on the one hand and intrinsic opaqueness on other hand. The paper is organized as follows: in section 2 we deal with the preliminaries and definitions. In section 3 we categorize the work in methods for interpretability. Section 4 offers discussion into the current state of the research field and lists future research ideas. Section 5 concludes the paper

II. PRELIMINARIES AND DEFINITIONS

In this section we offer definitions of: trust, interpretability, comprehensibility, and explainability.

Trust is defined in [2] as a psychological state in which an agent willingly and securely becomes vulnerable, or depends on, a trustee, having taken into consideration the characteristics of the trustee.

Authors in [15] claim that unlike normal ML objective functions, it is hard to formalize the definitions of criteria that are crucial for trust and acceptance, view backed by [2], [3]. In those cases of incomplete problem formalization, interpretability is used as a fallback or proxy for other criteria.

However, there is no unique definition of interpretability [3], [15]. In [3] interpretability is found not to be a monolithic concept, but in fact it reflects several distinct ideas and that in many papers interpretability is proclaimed axiomatically. Authors in [15] define that to interpret means to explain or to present in understandable terms. Then, interpretability in the context of ML systems is the ability to explain or to present in understandable terms to humans.

Interpretability and explainability are often used interchangeably in literature, but some papers make distinction. In [17] interpretation is the mapping of abstract concept into a domain humans can make sense of, while explanation is the collection of features of interpretable domain that have contributed for a given example to produce a decision. Edwards and Veale in [20] split explanations into model-centric and subject-centric, notions which correspond to definitions of interpretability and explainability from [17]. Similar roles in [15] take up global and local interpretability, respectively. In that view, we can see that GDPR covers only explainability. Comprehensibility [14] is used in a literature as a synonym for interpretability. Transparency [3] is used as a synonym for model interpretability, that is some sense of understanding the working logic of the model.

None of the aforementioned definitions is specific or restrictive enough to enable formalization. They implicitly depend on user's expertise, preferences and other contextual variables.

III. METHODS FOR INTERPRETABILITY AND EXPLAINABILITY

There are two categories of approaches to interpretability and explainability: integrated (transparency-based) and post-hoc.

Transparency [3] is one of the properties that can enable interpretability. Transparency was a traditional first step to protection of rights in human-based institutions and by analogy it is ported to algorithmic concerns such as unfairness and discrimination [20]. But, models in AI are becoming much more complex than human-based institutions and it becomes hard to find meaningful explanation that users might be able to understand. Also, human thinking, including our own, is not transparent to us and justifications in the form of explanations and interpretations may differ from the actual decision mechanism. In addition, predictive performance and transparency are conflicting objectives and they have to be traded-off in a model [21], [22]. In [7] it is stated that it is not clear how much transparency matters in the long run. If the systems are robust and self-contained it may not be necessary. But, if they are part of other systems, then transparency can be good for debuggability.

Post-hoc interpretability extracts informations from already learned model and it does not precisely depend on how the model works. The advantage of this approach is that it does not impact performance of the model which is treated as a black-box (BB). This is similar mode to how people make justifications for their own decisions, without fully knowing the real functioning of their decision-making mechanisms. However, special care must be taken in order to avoid systems that generate plausible but misleading explanations. Such explanations could satisfy laws like GDPR, but there is a problem of checking their veracity.

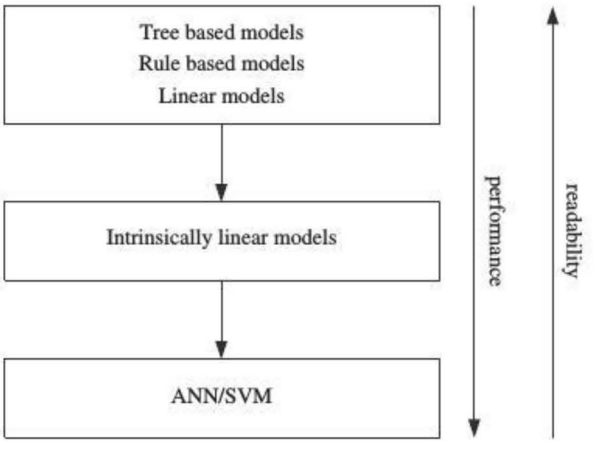

Figure 1 Performance-transparency trade-off [18]

A. *Integrated intepretability*

The best explanation of a simple model is the model itself; it perfectly represents itself and is easy to understand [23]. This approach is limited to the model families with lower complexity (flexibility), such as linear models, decision trees and rules. On the other hand, other

model families such as artificial neural networks (ANN) and support vector machines (SVM), boosted trees, and random forests are considered opaque and their complexity prevents users from tracing the logic behind predictions. The latter are most often considered as black-boxes and they are dealt with in a post-hoc manner. This trade-off between transparency and performance is conceptually depicted in Fig 1. The majority of work is done for classification tasks. Different constraints can be imposed on models in order to increase their interpretability. Model size, sparsity, and monotonicity ([14], [18]) are some of the constraints used in the literature. Even the choice of model family (representation) can be considered a constraint on models that affects the interpretability. Transparent models are both interpretable and explainable.

User-based studies on interpretability of decision-rules, tables, and tree algorithms in classification were conducted in [24] and [25]. In the latter paper, decision tables were found to be the easiest to use for inexperienced users and that the model size in general has negative impact on interpretability, answer time and confidence. Freitas in [14] gives an overview of work in comprehensible classification models where transparency plays the major role. Interpretability of decision trees, classification rules, decision tables, nearest neighbors, and Bayesian network classifiers is discussed and monotonicity constraints are advocated for improving the model transparency.

There are two sub-approaches: pure transparent and hybrid.

In pure transparent approaches we are restricted to use model families that are considered transparent. Evolutionary programming was used in [26] to search for sets of interpretable classification rules with small number of rules and conditions. Interpretable decision sets [27] are sets of independent if-then rules. Since each rule can be applied independently, interpretation is simple. The model is found by optimizing objective that takes into account both accuracy and interpretability. Oblique treed sparse additive models, region-specific predictive models, were proposed in [28]. They achieved competitive performance with kernel SVMs while providing interpretable model. Prototype selection [29] uses set covering optimization problem to achieve sparsity in samples for classification. Minimal set of prototypes is selected in order to get good nearest neighbor classifier. It was tested on recognizing handwritten digits and it showed reasonable performance.

Hybrid approaches combine transparent model families with black-box methods in order to get appropriate trade-off between the model interpretability and the predictive performance. Combination of logistic regression and SVMs was used for credit scoring in [30] in order to improve accuracy of the initial interpretable model. Multi-objective learning of hybrid classifiers was utilized in [31] to learn hybrid trees where certain leafs were substituted with black-box classifiers for boosting accuracy at the expense of interpretability.

B. Post-hoc methods

With hardware improvements and increased availability of data, predictive performance benefits of using complex, opaque models are increased. However, interpretability and explainability are issues that have to be properly addressed. In these approaches we start with trained black-box predictor and, sometimes, the used training data. Some methods deal with interpretability while others with explainability, according to the definitions in section 2. Methods are model-agnostic if they work only with the inputs and outputs of BB model, and model-specific if they use idiosyncrasies of some representation.

1) Interpretability

Transparent proxy model approach finds interpretable model that globally approximates the predictions of the black-box model. This approach offers both interpretability and explainability. In [32] a model-specific method was used to learn single decision tree from the ensemble of decision trees. The learned model was more accurate than the decision tree learned directly from the data. Rules were extracted from SVM in [33] to make more interpretable model for credit scoring. Interpretability was gained at only a small loss in performance compared to SVMs. Symbolic rules were extracted from neural network ensembles in [34]. Bayesian regression mixture with multiple elastic nets was proposed [35] and used on DNN, SVM, and random forests to explain individual decisions and look for model vulnerabilities in image recognition and text classification. Model vulnerabilities were tested with adversarial examples. Adversarial training scheme was used on DNNs in order to increase their interpretability [36]. Using adversarial examples, it was found that in normally trained DNNs neurons do not detect semantic parts but only discriminative part patches. Also, representations are not robust codes of visual concepts. After adversarial training scheme, representation is more interpretable, enabling to trace the outcomes to influential neurons. This is more transparent way of making predictions.

Indicative approaches give weaker notion of interpretability to BB model than the above methods. Some aspects of model functioning are elicited with conceptual representations. Visualization techniques for high-dimensional data like [37] can be used. In [38], visualization technique based on trained deconvolutional network was created for visualization of intermediate layers of convolutional neural networks. The insights through these visualizations enabled creation of improved architectures that outperformed existing approaches at the time. Similar was done for recurrent neural networks in [39], where character-level language models were used as an interpretable benchmark. Experiments revealed the existence of interpretable cells that keep track of long-range dependencies in text. Further work was done in [40] on interpreting neural models in natural language processing through visualization of unit's salience.

Model-agnostic visualization method based on sensitivity analysis was proposed in [41]. Input effects on model responses are inferred and visualizations of barplots for feature importance and Variable Effect Characteristic curves. The experiments were performed on neural

network ensemble and SVM models. Model-agnostic method for auditing indirect influences in BB models was presented in [42]. The procedure finds indirect influences of attributes to output through related features, even when attributes don't have the direct influence. Tamagnini et al. [43] created pedagogical visualization-based system Rivelo for interpretation of BB classifiers using instance-level explanations. User can interactively explore the set of instance-level explanations to create a mental model.

2) Explainability

These methods mostly output for each prediction also an explanation in the form of feature importances for making that decision. Layer-wise relevance propagation and sensitivity analysis were presented in [16] to explain predictions for deep learning models in the terms of input variables. In [44], deep Taylor decomposition propagates explanations from the outputs of DNN to the contributions of its inputs. Model-agnostic method for capturing the degree of influence of inputs on outputs of the system was presented in [45]. Local approximation method, SHapley Additive exPlanations (SHAP), was used in [23] to explain prediction f(x) for a single input x. SHAP is a unified framework for value estimation of additive feature attributions that generalizes several works from the literature [45]–[49]. Both model-agnostic and more efficient model-specific variant were proposed. DNN that identifies contents in the image and generates caption was described in [50]. For each word in caption, an explanation is generated in a form of highlighted relevant regions of the input image.

Other methods give explanations in other forms, such as visualization, text, examples, etc. DNNs were trained for visual question answering and explaining human activities in [51]. Justifications for decisions were given textually and evidence in images was emphasized using the attention mechanism. Textual explanations were generated together with visual classification using DNN [52]. Reinforcement learning based discriminative loss function was used for explanation model.

IV. Discussion

In the introduction we mentioned the utility of abstracted explanations for finding useful properties and generating hypotheses about data-generating processes which is important for science as well as for future Artificial General Intelligence (AGI) systems. Generated hypotheses can be further reasoned and experimented with, leading to the bootstrapped process of iterative improvement. However, as the authors in [19] point out, approaches listed in this paper only enable explanations of decisions, instead of actually generating them. It is up to the user to do the hypothesis generation, experimentation and reasoning. Research direction interesting for science and AGI is bridging neural-symbolic gap for seamless integration of learning and reasoning [53], [54]. Successful implementation would enable automatic generation of interpretations, explanations, and reasoning over them.

We have also seen that researchers use different names for similar or identical concepts. Definitions and vocabulary should be fixed in the community in order to enable easier transfer of results and information. Factors of trust-inducing criteria should be formalized, since these concepts are ambiguous and need to be split into smaller, more specific constituents. Model framework of human trust should be found, as an extension of work in [2]. This can be done in a similar data-driven approach as proposed in [15] for interpretability. There is a lack of empirical studies on user-based measures of interpretability. Also, richer loss functions need to be developed that take into account more criteria of performance in the real world. Most of the machine learning work has revolved around the scalar objectives, while the problems we are talking about are multicriteria with some criteria not even explicitly given in the form of optimization objective. Research into the extensions of interpretability research to reinforcement learning is another potential venue. Pedagogical interactive post-hoc approaches such as [43] are promising to enable people create their mental models of complex algorithms [20] which can boost trust by increasing the familiarity with models' decisions.

V. Conclusion

There are evident problems with the ethical- as well as quality-of-life implications of using AI in their current form in real-world scenarios. On several fronts, people are yet to see major impact applications such as in judicial, governmental, financial, and autonomous transport. But, for some time already, human lives have been influenced by algorithmic content recommendation which shapes opinions and tastes. With a greater spread of AI applications, trust-related problems are likely to become more pressing issues. Trust is boosted with specific criteria, but there are prominent problems with the incompleteness in problem formalization. This is a barrier to straightforward optimization approaches – some notions are so complex, multidimensional, and ambiguous that they are hard to put down in a formal way. Interpretability and explainability offer abstracted explanations for finding, checking, and reasoning over useful properties. This can be used not only for verifying trust criteria, but also for scientific discovery and in future AGI systems.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, May 2015.

[2] B. W. Israelsen, "'I can assure you ... that it's going to be all right' -- A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships," ArXiv170800495 Cs Stat, Aug. 2017.

[3] Z. C. Lipton, "The Mythos of Model Interpretability," ArXiv160603490 Cs Stat, Jun. 2016.

[4] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan, "Exploring the Filter Bubble: The Effect of Using Recommender Systems on Content Diversity," in Proceedings of the 23rd International Conference on World Wide Web, New York, NY, USA, 2014, pp. 677–686.

[5] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, and Y. Bengio, "ObamaNet: Photo-realistic lip-sync from text," ArXiv180101442 Cs, Dec. 2017.

[6] A. Kucharski, "Study epidemiology of fake news," Nature, vol. 540, p. 525, Dec. 2016.

[7] G. Marcus, "Deep Learning: A Critical Appraisal," ArXiv180100631 Cs Stat, Jan. 2018.

[8] C. Szegedy et al., "Intriguing properties of neural networks," presented at the International Conference on Learning Representations, 2014, pp. 1–10.

[9] K. Crawford, "Opinion | Artificial Intelligence's White Guy Problem," The New York Times, 25-Jun-2016.

[10] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," in Advances in Neural Information Processing Systems 29, Barcelona, Spain, 2016, pp. 4349–4357.

[11] M. J. Wolf, K. Miller, and F. S. Grodzinsky, "Why We Should Have Seen That Coming: Comments on Microsoft's Tay 'Experiment,' and Wider Implications," SIGCAS Comput Soc, vol. 47, no. 3, pp. 54–64, Sep. 2017.

[12] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation,'" ArXiv160608813 Cs Stat, Jun. 2016.

[13] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," Behav. Brain Sci., vol. 40, ed 2017.

[14] A. A. Freitas, "Comprehensible Classification Models: A Position Paper," SIGKDD Explor Newsl, vol. 15, no. 1, pp. 1–10, Mar. 2014.

[15] F. Doshi-Velez and B. Kim, "Towards A Rigorous Science of Interpretable Machine Learning," ArXiv170208608 Cs Stat, Feb. 2017.

[16] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," ITU J. ICT Discov. - Spec. Issue 1 - Impact Artif. Intell. AI Commun. Netw. Serv., vol. 1, pp. 1–10, Dec. 2017.

[17] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," Digit. Signal Process., vol. 73, pp. 1–15, Feb. 2018.

[18] D. Martens, J. Vanthienen, W. Verbeke, and B. Baesens, "Performance of classification models from a user perspective," Decis. Support Syst., vol. 51, no. 4, pp. 782–793, Nov. 2011.

[19] D. Doran, S. Schulz, and T. R. Besold, "What Does Explainable AI Really Mean? A New Conceptualization of Perspectives," ArXiv171000794 Cs, Oct. 2017.

[20] L. Edwards and M. Veale, "Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2972855, May 2017.

[21] Y. Jin and B. Sendhoff, "Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies," IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., vol. 38, no. 3, pp. 397–415, May 2008.

[22] A. A. Freitas, "A Critical Review of Multi-objective Optimization in Data Mining: A Position Paper," SIGKDD Explor Newsl, vol. 6, no. 2, pp. 77–86, Dec. 2004.

[23] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," ArXiv170507874 Cs Stat, May 2017.

[24] H. Allahyari and N. Lavesson, "User-oriented Assessment of Classification Model Understandability," in DIVA, 2011.

[25] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," Decis. Support Syst., vol. 51, no. 1, pp. 141–154, Apr. 2011.

[26] A. Cano, A. Zafra, and S. Ventura, "An interpretable classification rule mining algorithm," Inf. Sci., vol. 240, pp. 1–20, Aug. 2013.

[27] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable Decision Sets: A Joint Framework for Description and Prediction," in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2016, pp. 1675–1684.

[28] J. Wang, R. Fujimaki, and Y. Motohashi, "Trading Interpretability for Accuracy: Oblique Treed Sparse Additive Models," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2015, pp. 1245–1254.

[29] J. Bien and R. Tibshirani, "Prototype selection for interpretable classification," Ann. Appl. Stat., vol. 5, no. 4, pp. 2403–2424, Dec. 2011.

[30] T. Van Gestel, B. Baesens, P. Van Dijcke, J. Suykens, J. Garcia, and T. Alderweireld, "Linear and Non-linear Credit Scoring by Combining Logistic Regression and Support Vector Machines," J. Credit Risk, vol. 1, Jan. 2005.

[31] R. Piltaver, M. Luštrek, and M. Gams, "Multi-objective learning of accurate and comprehensible classifiers – a case study," 2014.

[32] A. V. Assche and H. Blockeel, "Seeing the Forest Through the Trees: Learning a Comprehensible Model from an Ensemble," in Machine Learning: ECML 2007, 2007, pp. 418–429.

[33] D. Martens, B. Baesens, T. Van Gestel, and J. Vanthienen, "Comprehensible credit scoring models using rule extraction from support vector machines," Eur. J. Oper. Res., vol. 183, no. 3, pp. 1466–1476, Dec. 2007.

[34] Z.-H. Zhou, Y. Jiang, and S.-F. Chen, "Extracting symbolic rules from trained neural network ensembles," AI Commun., vol. 16, no. 1, pp. 3–15, Jan. 2003.

[35] W. Guo, K. Zhang, L. Lin, S. Huang, and X. Xing, "Towards Interrogating Discriminative Machine Learning Models," ArXiv170508564 Cs Stat, May 2017.

[36] Y. Dong, H. Su, J. Zhu, and F. Bao, "Towards Interpretable Deep Neural Networks by Leveraging Adversarial Examples," ArXiv170805493 Cs, Aug. 2017.

[37] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," J. Mach. Learn. Res., vol. 9, no. Nov, pp. 2579–2605, 2008.

[38] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in Computer Vision – ECCV 2014, 2014, pp. 818–833.

[39] A. Karpathy, J. Johnson, and L. Fei-Fei, "Visualizing and Understanding Recurrent Networks," ArXiv150602078 Cs, Jun. 2015.

[40] J. Li, X. Chen, E. Hovy, and D. Jurafsky, "Visualizing and Understanding Neural Models in NLP," in Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 681–691.

[41] P. Cortez and M. J. Embrechts, "Using sensitivity analysis and visualization techniques to open black box data mining models," Inf. Sci., vol. 225, pp. 1–17, Mar. 2013.

[42] P. Adler et al., "Auditing Black-box Models for Indirect Influence," Knowl Inf Syst, vol. 54, no. 1, pp. 95–122, Jan. 2018.

[43] P. Tamagnini, J. Krause, A. Dasgupta, and E. Bertini, "Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations," in Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics, New York, NY, USA, 2017, p. 6:1–6:6.

[44] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep Taylor decomposition," Pattern Recognit., vol. 65, pp. 211–222, May 2017.

[45] A. Datta, S. Sen, and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems," in 2016 IEEE Symposium on Security and Privacy (SP), 2016, pp. 598–617.

[46] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA, 2016, pp. 1135–1144.

[47] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," ArXiv170402685 Cs, Apr. 2017.

[48] E. Štrumbelj and I. Kononenko, "Explaining Prediction Models and Individual Predictions with Feature Contributions," Knowl Inf Syst, vol. 41, no. 3, pp. 647–665, Dec. 2014.

[49] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," PLOS ONE, vol. 10, no. 7, p. e0130140, Jul. 2015.

[50] K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in International Conference on Machine Learning, 2015, pp. 2048–2057.

[51] D. H. Park, L. A. Hendricks, Z. Akata, B. Schiele, T. Darrell, and M. Rohrbach, "Attentive Explanations: Justifying Decisions and Pointing to the Evidence," ArXiv161204757 Cs, Dec. 2016.

[52] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating Visual Explanations," in Computer Vision – ECCV 2016, 2016, pp. 3–19.

[53] T. R. Besold and K.-U. Kühnberger, "Towards integrated neural–symbolic systems for human-level AI: Two research programs helping to bridge the gaps," Biol. Inspired Cogn. Archit., vol. 14, pp. 97–110, Oct. 2015.

[54] T. R. Besold et al., "Neural-Symbolic Learning and Reasoning: A Survey and Interpretation," ArXiv171103902 Cs, Nov. 2017.