On Unhelpful Explainable Machine Learning Misconceptions

Patrick Hall*†
Washington, DC
patrick.hall@h2o.ai

April 22, 2019

Introduction

As someone who has been involved in the implementation of explainable machine learning (ML) software for the past three years, I find a lot of what I read about the topic confusing and detached from my personal, hands-on experiences. This short text presents arguments, proposals, and references to address explainable ML misconceptions. Please note that this text is not an attack on any party. However, it does promote informed debate, scrutiny, and continued development of explainable ML, and not dismissal or derision of the discipline. Due to obvious community and customer demand, explainable ML methods have already been implemented in popular open source software and in commercial software.[1,2] The methods will likely be used widely and proper usage discussions are probably more helpful than "please stop" requests [24]. Moreover, this text builds the seemingly natural case for a holistic approach to ML that includes interpretable (i.e "white-box") models along with explanatory, debugging, and disparate impact analysis and remediation techniques. Much groundbreaking work has been done in several related disciplines to make machine learning more interpretable and trustworthy, and the cursory application of traditional black-box ML to human-centered problems is now an outdated, ethically problematic, and potentially dangerous practice.

To avoid ambiguity, several initial internal definitions and accompanying examples are outlined before discussing misconceptions. As illustrated in Figure 1, here *explainable ML* means post-hoc techniques used to understand trained model behavior or predictions. Examples of common explainable ML techniques include:

- Local and global feature importance methods, in particular Shapley values [25], [17], [26], [20].
- Local and global model-agnostic surrogate models, such as surrogate decision trees and Local Interpretable Model-agnostic Explanations (LIME) [6], [5], [16], [23].
- Local and global visualizations of model predictions such as accumulated local effect (ALE), 1- and 2-dimensional partial dependence, and individual conditional expectation (ICE) plots [4], [9], [11].

In this text *model debugging* refers to testing ML models to increase trust in model mechanisms and predictions. Examples of model debugging techniques include variants of sensitivity (i.e. "what-if?") and residual analysis used to test models for errors or security vulnerabilities. Model debugging should also include remediating any discovered errors or vulnerabilities. Herein *fairness* techniques refer to canonical disparate impact analysis, model selection by minimization of disparate impact, and remediation techniques such as disparate impact removal preprocessing or equalized odds post processing [8], [15]. In this text *interpretable* or *white-box* models will include linear models, decision trees, constrained or Bayesian variants of

*H2O.ai and George Washington University

[1]Like h2o-3, xgboost, and various other Python and R packages. See: https://github.com/jphall663/awesome-machine-learning-interpretability for a longer, curated list of open source software packages.
[2]For instance Datarobot, H2O Driverless AI, SAS Visual Data Mining and Machine Learning, Zest AutoML, and likely several others.

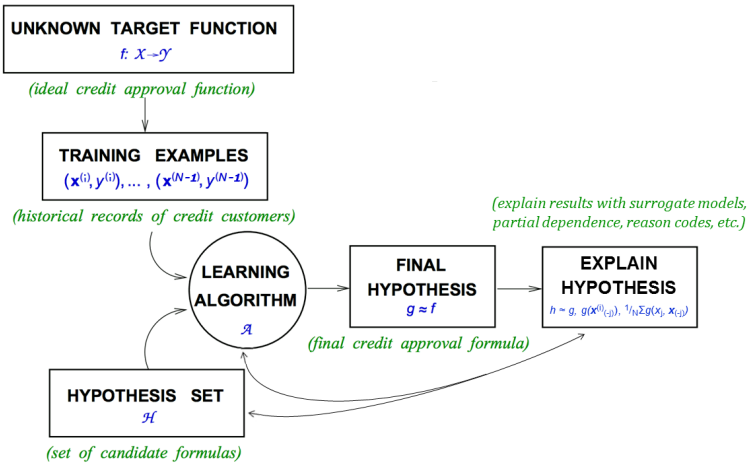


Figure 1: An augmented learning problem diagram in which several post-hoc techniques create explanations for a credit scoring model g. When used properly, explanations can increase understanding of g and help improve the accuracy, fairness, interpretability, privacy, or security of subsequent applications of $\mathcal{H}$ and $\mathcal{A}$. Adapted from Figure 1.2 of the open textbook *Learning From Data* [1].

traditional black-box ML models, or novel types of models designed to be directly interpretable. Additional examples of interpretable modeling techniques include explainable neural networks (XNNs), monotonically constrained gradient boosting machines (GBMs)[3], scalable Bayesian rule lists, or super-sparse linear integer models (SLIMs) [28], [31], [27]. Herein unconstrained, traditional black-box ML models, such as multilayer perceptron (MLP) neural networks and GBMs, are said to be uninterpretable.

1 Misconception: Explanations are Necessary and Sufficient to Establish Trust in ML

Explanations are likely necessary for trust in many cases, but certainly not sufficient for trust in all cases. Explanation, as a general concept, is related more directly to understanding and transparency than to trust.[4] Simply put, one can understand and explain something without trusting it. One can also trust something and not be able to understand or explain it. Consider the following example scenarios:

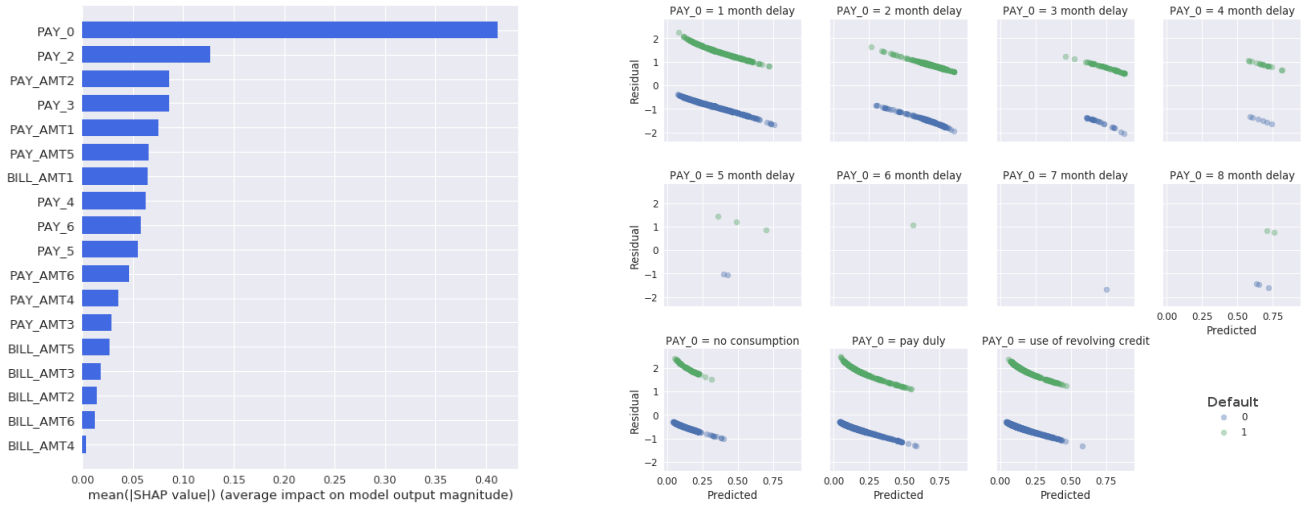
- **Explanation and understanding without trust**: In Figure 2, global Shapley explanations and residual analysis identify a pathology in an unconstrained GBM model, g_{GBM}. g_{GBM} over emphasizes the input feature `PAY_0`, or a customer's most recent repayment status. Over-weighing `PAY_0` makes g_{GBM} often unable to predict on-time payment if recent payments are delayed (`PAY_0` > 1), causing large negative residuals. g_{GBM} is also often unable to predict default if recent payments are on-time (`PAY_0` *leq* 1), causing large positive residuals. In this example scenario, g_{GBM} is explainable, but not trustworthy.

- **Trust without explanation and understanding**: Years before reliable explanation techniques were widely acknowledged and available, black-box predictive models, such as autoencoder and MLP neural networks, were used for fraud detection in the financial services industry [12]. When these models performed well, they were trusted.[5] However, they were not explainable or well-understood by contemporary standards.

[3] As implemented in XGBoost (https://xgboost.readthedocs.io/en/latest/tutorials/monotonic.html) or H2O-3 (https://github.com/h2oai/h2o-3/blob/master/h2o-py/demos/H2O_tutorial_gbm_monotonicity.ipynb).

[4] The Merriam-Webster definition of *explain*, accessed April 21st 2019, does not mention *trust*

[5] For example: https://www.sas.com/en_ph/customers/hsbc.html, https://www.kdnuggets.com/2011/03/sas-patent-fraud-detection.html.



(a) Global Shapley feature importance values for g_{GBM}. (b) g_{GBM} deviance residuals and predictions by `PAY_0`.

Figure 2: An unconstrained GBM probability of default model, g_{GBM}, over-emphasizes the importance of the input feature `PAY_0`, a customer's most recent repayment status. g_{GBM} produces large positive residuals when `PAY_0` indicates on-time payments and large negative residuals when `PAY_0` indicates late payments.

If trust in models is your goal, then explanations alone are not sufficient. However, as discussed in Section 6 and illustrated in Figure 3, in an ideal scenario, explanation techniques would be used with a wide variety of other methods to increase accuracy, fairness, interpretability, privacy, security, and trust in ML models.

2 Misconception: Explainable ML is Unnecessary

Local feature importance values for negative credit decisions, i.e. adverse action codes, are mandated under the Fair Credit Reporting Act for many credit lending decisions in the United States. If machine learning is used for such decisions, it has to be explained with local feature importance. In a number of other application domains, broader interpretability is also legal necessity. Explanation, along with white-box models, model debugging, disparate impact analysis, and the documentation they enable, can also be required under the Civil Rights Acts of 1964 and 1991, the Americans with Disabilities Act, the Genetic Information Nondiscrimination Act, the Health Insurance Portability and Accountability Act, the Equal Credit Opportunity Act, the Fair Housing Act, Federal Reserve SR 11-7, the European Union (EU) Greater Data Privacy Regulation (GDPR) Article 22, and other regulatory statutes [30].

3 Misconception: All the Key Terms in Explainable ML are Undefined

Helpful definitions that apply to explainable ML have been put forward, including:

- **Interpretable**: "The ability to explain or to present in understandable terms to a human" – in "Towards a Rigorous Science of Interpretable Machine Learning" by Doshi-Velez and Kim (2017) [7].
- **A Good Explanation**: "When you can no longer keep asking why" – in "Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning" by Gilpin et al. (2018) [10]. (Gilpin et al. also provide several clear constructs for describing more specific types of explanations.)

While the explainable ML field is far from embracing a clear and accepted taxonomy of concepts or an exhaustive and precise vocabulary, these two well-founded definitions appear to link explanations to some

ML process being interpretable. Moreover, many authors have made significant attempts to grapple with a variety of general concepts related to interpretability and explanations, including "A Survey Of Methods For Explaining Black Box Models" by Guidotti et al. (2018), Zachary Lipton's "The Mythos of Model Interpretability" (2016), Christoph Molnar's *Interpretable Machine Learning* (2018), and Adrian Weller's "Challenges for Transparency" (2017) [13], [18], [21], [29].

4 Misconception: Explainable ML is Just Models of Models

Models of models, or surrogate models, can be helpful explanatory tools, but they are usually approximate, low-fidelity explainers. Aside from 1.) a global summary of a complex model provided by a surrogate model can be helpful sometimes and 2.) much work in explainable ML has been directed toward improving the fidelity and usefulness of surrogate models [6], [5], [16], [28], **many explainable ML techniques have nothing to do with surrogate models!**

One of the most exciting breakthroughs for supervised learning problems in explainable ML is the application of a coalitional game theory concept, Shapley values, to compute feature contributions which are consistent globally and accurate locally using the trained model itself [26], [20]. An extension of this idea, called tree SHAP, has already been implemented for popular tree ensemble methods [19]. There are many other explainable ML methods that operate on trained models directly such as partial dependence and ICE plots [9], [11]. Notably, surrogate models and explanatory techniques that operate directly on trained models can be combined, for instance by using partial dependence, ICE, and surrogate decision trees to investigate and confirm modeled interactions [14].

For a curated list of many different types of white-box modeling techniques, and model-specific, model-agnostic, and surrogate model explainable ML techniques, please see:

https://github.com/jphall663/awesome-machine-learning-interpretability

Misconception Corollary: Explainable ML is just LIME. LIME, in it's most popular implementation, uses local linear surrogate models [23]. LIME is important, and imperfect, but just one of many explainable ML tools. And again, LIME can sometimes be combined with model-specific methods to yield deeper insights. Consider that tree SHAP can provide locally accurate and consistent point estimates for local feature importance whereas LIME can provide approximate information about modeled local linear trends around the same point.

5 Misconception: Explainable ML Methods Simply Provide Cover to Use Black-Box ML for Nefarious Purposes

If used disingenuously, explainable ML methods probably do provide such cover [2]. But explainable ML methods were designed specifically to crack open those same nefarious and complex black-boxes. See Angwin et al. (2016) for evidence that hacking or stealing of commercial black-box models for oversight purposes is possible [3]. Such investigations would likely only be improved by advances in explanatory and fairness tools. Additionally, many important computer-based technological advances present similar double-edged sword dilemmas, e.g. social media or strong encryption. Rarely does the ability of a tool to be misused for malicious purposes disqualify it from being used as designed.

6 Misconception: Explainable ML Methods and White-box Models are Somehow Mutually Exclusive

A few well-known publications have focused either on white-box modeling techniques (e.g. [27], [31]) or on post-hoc explanations (e.g. [23], [20]), but the two can be used together in the context of a broader and more human-centered machine learning workflow as illustrated in Figure 3.

Consider the seemingly useful example case of augmenting globally interpretable models with local post-hoc explanations: A practitioner could train a single pruned decision tree as a globally interpretable model

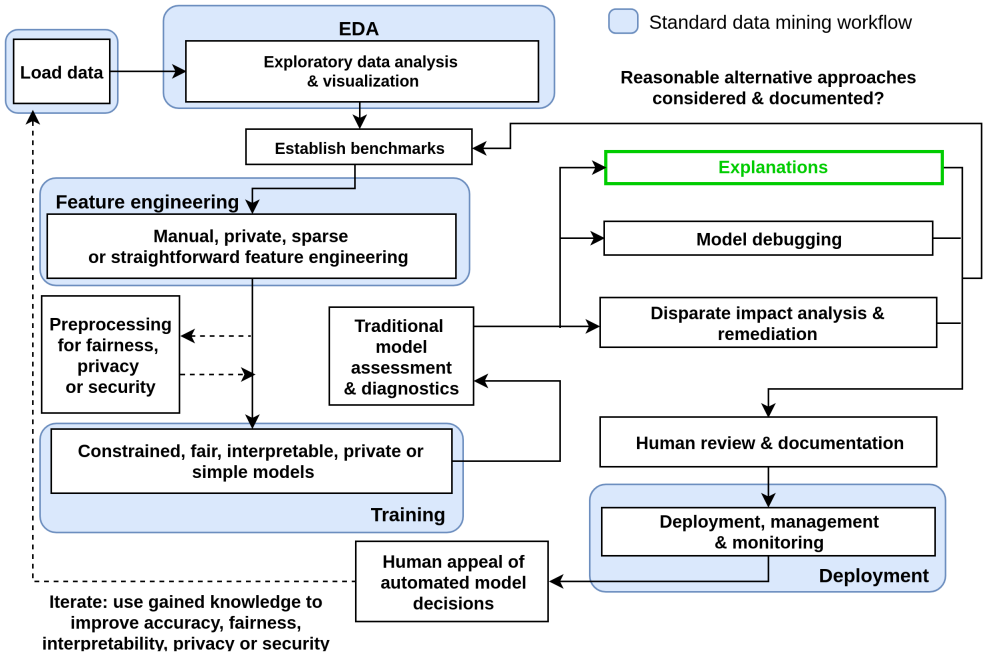


Figure 3: A diagram of a proposed human-centered machine learning workflow in which explanations (highlighted in green) are used along with interpretable models, disparate impact analysis and remediation techniques, and other review and appeal mechanisms to create a fair, accountable, and transparent ML system.

then use local explanations in the form of Shapley feature importance. This would enable the practitioner to see accurate numeric feature contributions for each model prediction in addition to the entire directed graph of the decision tree. Isn't a complete global and local understanding of the model more desirable than the white-box model or the post-hoc explanations alone?

Corollary Misconception: Explainable ML Methods and Fairness Methods are Somehow Mutually Exclusive: Like white-box models, fairness methods are often presented in different articles than post-hoc explanatory methods. However, in banks, using post-hoc explanatory tools such as partial dependence plots and local feature importance to comply with credit reporting regulations often goes hand-in-hand with using disparate impact analysis to comply with fair lending regulations.

Conclusion

Acknowledgemnts

The author thanks Pramit Choudhary and Navdeep Gill at H2O.ai for their input and insights.

References

[1] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. ***Learning from Data***. AMLBook, New York, 2012. URL: https://work.caltech.edu/textbook.html.

[2] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the Risk of Rationalization. *arXiv preprint arXiv:1901.09749*, 2019. URL: https://arxiv.org/pdf/1901.09749.pdf.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica*, 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Daniel W. Apley. Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv preprint arXiv:1612.08468*, 2016. URL: https://arxiv.org/pdf/1612.08468.pdf.

[5] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting Blackbox Models via Model Extraction. *arXiv preprint arXiv:1705.08504*, 2017. URL: https://arxiv.org/pdf/1705.08504.pdf.

[6] Mark W. Craven and Jude W. Shavlik. Extracting Tree-structured Representations of Trained Networks. *Advances in Neural Information Processing Systems*, 1996. URL: http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks.pdf.

[7] Finale Doshi-Velez and Been Kim. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608*, 2017. URL: https://arxiv.org/pdf/1702.08608.pdf.

[8] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015. URL: https://arxiv.org/pdf/1412.3756.pdf.

[9] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. ***The Elements of Statistical Learning***. Springer, New York, 2001. URL: https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf.

[10] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *arXiv preprint arXiv:1806.00069*, 2018. URL: https://arxiv.org/pdf/1806.00069.pdf.

[11] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *Journal of Computational and Graphical Statistics*, 24(1), 2015. URL: https://arxiv.org/pdf/1309.6392.pdf.

[12] Krishna M. Gopinathan, Louis S. Biafore, William M. Ferguson, Michael A. Lazarus, Anu K. Pathria, and Allen Jost. Fraud Detection using Predictive Modeling, October 6 1998. US Patent 5,819,226. URL: https://patents.google.com/patent/US5819226A.

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51(5):93, 2018. URL: https://arxiv.org/pdf/1802.01933.pdf.

[14] Patrick Hall. On the Art and Science of Machine Learning Explanations. In *JSM Proceedings, Statistical Computing Section*, pages 1781–1799. American Statistical Association, 2018. URL: https://github.com/jphall663/jsm_2018_paper.

[15] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016. URL: http://papers.nips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

[16] Linwei Hu, Jie Chen, Vijayan N. Nair, and Agus Sudjianto. Locally Interpretable Models and Effects Based on Supervised Partitioning (LIME-SUP). *arXiv preprint arXiv:1806.00663*, 2018. URL: https://arxiv.org/ftp/arxiv/papers/1806/1806.00663.pdf.

[17] Alon Keinan, Ben Sandbank, Claus C. Hilgetag, Isaac Meilijson, and Eytan Ruppin. Fair Attribution of Functional Contribution in Artificial and Biological Networks. *Neural Computation*, 16(9):1887–1915, 2004. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.436.6801&rep=rep1&type=pdf.

[18] Zachary C. Lipton. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490*, 2016. URL: https://arxiv.org/pdf/1606.03490.pdf.

[19] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent Individualized Feature Attribution for Tree Ensembles. In Been Kim, Dmitry M. Malioutov, Kush R. Varshney, and Adrian Weller, editors, *Proceedings of the 2017 ICML Workshop on Human Interpretability in Machine Learning (WHI 2017)*, pages 15–21. ICML WHI 2017, 2017. URL: https://openreview.net/pdf?id=ByTKSo-m-.

[20] Scott M. Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[21] Christoph Molnar. *Interpretable Machine Learning*. christophm.github.io/interpretable-ml-book, 2018. URL: https://christophm.github.io/interpretable-ml-book/.

[22] Arvind Narayanan. Translation Tutorial: 21 Fairness Definitions and Their Politics. In *Proceedings of the Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML), New York, USA*, 2018. URL: https://www.youtube.com/watch?v=wqamrPkF5kk.

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144. ACM, 2016. URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.

[24] Cynthia Rudin. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv preprint arXiv:1811.10154*, 2018. URL: https://arxiv.org/pdf/1811.10154.pdf.

[25] Lloyd S. Shapley, Alvin E. Roth, et al. *The Shapley value: Essays in honor of Lloyd S. Shapley*. Cambridge University Press, 1988. URL: http://www.library.fa.ru/files/Roth2.pdf.

[26] Erik Strumbelj and Igor Kononenko. An Efficient Explanation of Individual Classifications using Game Theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010. URL: http://www.jmlr.org/papers/volume11/strumbelj10a/strumbelj10a.pdf.

[27] Berk Ustun and Cynthia Rudin. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning*, 102(3):349–391, 2016. URL: https://users.cs.duke.edu/~cynthia/docs/UstunTrRuAAAI13.pdf.

[28] Joel Vaughan, Agus Sudjianto, Erind Brahimi, Jie Chen, and Vijayan N Nair. Explainable Neural Networks Based on Additive Index Models. *arXiv preprint arXiv:1806.01933*, 2018. URL: https://arxiv.org/pdf/1806.01933.pdf.

[29] Adrian Weller. Challenges for Transparency. *arXiv preprint arXiv:1708.01870*, 2017. URL: https://arxiv.org/pdf/1708.01870.pdf.

[30] Mike Williams et al. *Interpretability*. Fast Forward Labs, 2017. URL: https://www.cloudera.com/products/fast-forward-labs-research.html.

[31] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian Rule Lists. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. URL: https://arxiv.org/pdf/1602.08610.pdf.