3
Interpretability in Intelligent Systems – A New Concept?

Lars Kai Hansen(✉) and Laura Rieger

DTU Compute, Technical University of Denmark, Kgs. Lyngby, Denmark
lkai@dtu.dk

Abstract. The very active community for interpretable machine learning can learn from the rich 50+ year history of explainable AI. We here give two specific examples from this legacy that could enrich current interpretability work: First, *Explanation desiderata* were we point to the rich set of ideas developed in the 'explainable expert systems' field and, second, tools for *quantification of uncertainty* of high-dimensional feature importance maps which have been developed in the field of computational neuroimaging.

Keywords: Interpretable AI · Machine learning · Uncertainty quantification

3.1 Neural Network Interpretability

High activity research fields often develop to be somewhat myopic in their early phases - simply because the large body of published work leaves little time to follow progress in other areas or even to look back at previous research in the field. Independent component analysis is a prominent early example, for which a history of lost insights was recalled in Comon's paper [6]. Deep learning interpretability is a contemporary example: One could easily get the impression that the interpretability issue surfaced with the new wave of deep learning, however, this is not the case. While end-to-end learning has hugely accentuated the need for explanations, interpretability is an active research topic with an over 50-year history. In fact, since the early days of intelligent systems the importance and focus on interpretability has only increased [30]. From scientific contexts, where interpretability methods can assist formulation of causal hypotheses, see e.g., work in bio-medicine [43] and computational chemistry [40], to recent societal importance in the European Union's General Data Protection Regulatory, establishing the so-called *Right to explanation* as coined by Goodman and Flaxman [15].

Here we make two dives into the rich history of explainability in intelligent systems and we ask 'what can modern work learn'?

First a semantic note. The terms interpretability and explainability are often used interchangeably in the literature. However, in a recent review [13] a useful distinction is made. The more general concept is explainability which covers

W. Samek et al. (Eds.): Explainable AI, LNAI 11700, pp. 41–49, 2019.
https://doi.org/10.1007/978-3-030-28954-6_3

interpretability, i.e., to communicate machine learning function to user, and completeness, i.e., that the explanation is a close enough approximation that it can be audited. The distinction is described: '...interpretability alone is insufficient. In order for humans to trust black-box methods, we need explainability – models that are able to summarize the reasons for neural network behavior, gain the trust of users, or produce insights about the causes of their decisions. While interpretability is a substantial first step, these mechanisms need to also be complete, with the capacity to defend their actions, provide relevant responses to questions, and be audited' [13]. In Montavon et al.'s comprehensive tutorial a related distinction is made: '... interpreting the concepts learned by a model by building prototypes, and explaining the model's decisions by identifying the relevant input variables' [30].

Explainability, in this broader sense, has been a key component in several intelligent systems communities and the central tenet of this paper is that future work can learn from looking back at this history. We will focus on two specific lines of research, the first concerns the broader foundation of explainability: What are the desiderata, i.e., the salient dimensions and issues that should be addressed? Our second focus area concerns the important specific challenge of understanding the dimensions of uncertainty in machine learning models and their explanations.

Going back in time prior to the new wave of deep learning, many have stressed the importance of interpretability. Breiman's classic paper *Statistical Modeling: The Two Cultures* has a strong focus on interpretability [3]. Breiman notes: 'Occam's Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict. For instance, linear regression gives a fairly interpretable picture of the y, x relation. But its accuracy is usually less than that of the less interpretable neural nets'. As we will see below, this dilemma has been acknowledged by the explainable expert systems community many years earlier. Breiman clearly expressed his preferences: 'On interpretability, trees rate an A+', however, it was already known that trees and rule based systems have severe limitations when it comes to both implementing and comprehending function, see e.g. [29].

The interest in intelligent systems' interpretability has earlier roots. In a 1988 position paper Mitchie discussed how AI would pass different criteria from weak to ultra-strong [28]: 'The ultra-strong criterion demands that the system be capable not only of explaining how it has structured its acquired skills: it should also be able to teach them'. This ambition is still very relevant.

Going further back in early expert system history, explanation and human interaction were key issues. Expert systems in the late 60's - like 'SCHOLAR' developed for instructional support - were designed for interaction [5], such as explaining why a student's answer was wrong in a mixed initiative dialogue. Stanford's widely discussed 'MYCIN' expert system for antimicrobial selection was designed with three components: A rule based decision support component that combined MYCIN and physicians judgment, an explanation module and a learning module [4,41,42]. This rule based system had about 200 rules in 1975.

MYCIN developers held it self-evident that AI could get medical acceptance only with convincing explanations [42]. Thus, MYCIN was equipped to map its internal rules to natural language and answer both 'why' and 'how' questions. By 1983 the MYCIN system had expanded to 500 rules and the state of the art was summarized in a review in Science [10]. In a 1984 book summarizing experiences with MYCIN no less than four chapters are devoted to MYCIN's explanation mechanisms [4].

Prior to Breiman's comments, earlier work on explainability in statistics includes Good's discussion of evidence in context of belief networks [14]. Good considered three dimensions of explanations: 'What', concerning semantic explanations as in a dictionary, 'How' as in natural or manufacturing process descriptions, and finally the 'Why' type explanations - hypothesizing causal mechanisms behind an event.

3.2 Desiderata of Explainable AI

Expert systems moved on and important principles can be learned from Swartout and Moore's 1993 review of 'second generation explainable expert systems' [47] listing five general desiderata for useful explanations of AI, adding significant perspective to recent work in the field:

$\mathcal{D}_1$ Fidelity: the explanation must be a reasonable representation of what the system actually does.
$\mathcal{D}_2$ Understandability: Involves multiple usability factors including terminology, user competencies, levels of abstraction and interactivity.
$\mathcal{D}_3$ Sufficiency: Should be able to explain function and terminology and be detailed enough to justify decision.
$\mathcal{D}_4$ Low Construction Overhead: The explanation should not dominate the cost of designing AI.
$\mathcal{D}_5$ Efficiency: The explanation system should not slow down the AI significantly.

Expert systems have developed through several generations. The notion of second versus first generation AI was based on the modes of explanation. First generation systems were characterized by explanations based directly on rules applied by the AI to reach decisions. This leads to high fidelity ($\mathcal{D}_1$), but often conflicts with understandability ($\mathcal{D}_2$) because the rules used for inference may be incomprehensive for the user [47]. So-called *Explainable Expert Systems* (EES) addressed this dilemma. The XPLAIN system [46] is an example. XPLAIN was based on two key principles to enhance understandability: 'explicitly distinguishing different forms of domain knowledge present in the knowledge base and formal recording of the system development process' [34]. The evaluation of XPLAIN is anecdotal, yet quite convincing. Cases are presented in which the system is able to answer 'why' questions - and even at times resorting to 'white lies' to create a smoother learning experience [46]. Computational complexity both in construction and execution (desiderata $\mathcal{D}_4 - \mathcal{D}_5$) are not so prominent in current literature, although the most widely used methods differ significant in complexity. The so-called Local Interpretable Model-agnostic Explanation (LIME) scheme, for example, is based on image segmentation, random sampling and multiple linear model fittings, hence rather complex at explanation time [37], hence a challenge to $\mathcal{D}_5$. An approach such a 'Testing with Concept Activation Vectors' (TCAV) comes at a significant initial cost [19], hence may pose a challenge to $\mathcal{D}_4$.

Much of the EES progress was produced in the context of rule based expert systems, while AI based on machine learning - so-called connectionists' methods - more often was considered 'black box'. Interest in connectionists' methods was primarily based on performance and not interpretability, c.f., the quote from [2] '...symbolic learning techniques produce more understandable outputs but they are not as good as connectionist learning techniques in generalization'. We already noted that this view was propagated by Breiman, hence, the sparking interest in converting existing neural networks to decision tree form [1,48,49] or even learn neural networks that more readily are converted to trees see for example work by Gallant [12] and by Craven and Shavlik [7]. But trees may not deliver on $\mathcal{D}_2$, in particular, as discussed above and noted by [46] - the intuitive appeal of trees fails in practice when trees get to be complex in structure or operate in high dimensional feature spaces. These challenges were also recently noted in [33]. For domains where modern neural networks excel such as image, audio and text data, tree based explanations are challenged.

Returning to the list of desiderata, several recent papers have aimed at framing the discourse of interpretability. Presumably unaware of [47], Lipton notes that interpretability is not a well-defined concept and goes on to discuss multiple dimensions of interpretability and formulates a set of desiderata [25,26] closely related to $\mathcal{D}_1 - \mathcal{D}_3$. Lipton's desiderata read (i) 'Trust', (ii) 'Causality', (iii) 'Transferability', (iv) 'Informativeness', and (v) 'Fair and Ethical Decision-Making'. Here Lipton discusses several dimensions of (i) 'Trust' mostly covered in desiderata $\mathcal{D}_1 - \mathcal{D}_2$, (ii) 'Causality' is roughly equivalent to [47]'s $\mathcal{D}_3$, while the notion of 'Transferability' and 'Informativeness' both refer to the user's ability to gain abstract 'knowledge' from explanations. This idea also appeared in the original paper's discussion of usability $\mathcal{D}_2$, viz. the need to explain a system at different levels of abstraction. 'Fair and Ethical Decision-Making' is noted by Lipton as an area that specifically requires interpretability. In [47] such considerations are framed in a general discussion of usability ($\mathcal{D}_2$). It is also noted that an explanation systems must be able to explain from different perspectives '.. e.g., form versus function in the biological domain or safety vs. profitability in the financial domain'.

The usability dimension ($\mathcal{D}_2$) remains an important issue in contemporary interpretability papers. The question 'Interpretable to Whom?' has been raised in several papers [8,33,50] focusing on the user and addressed by human factors evaluation. In fact, Doshi-Velez and Been open their paper with the more general statement 'Unfortunately, there is little consensus on what interpretability in machine learning is and how to evaluate it for benchmarking' [8]. Their key contribution is to point out that current machine learning workflows are incomplete in the sense that they have unspecified objectives from the application domain.

This can be important issues that were not included in the machine learning objective function: '...incompleteness that produces some kind of unquantified bias, e.g. the effect of including domain knowledge in a model selection process' [8]. Seemingly unaware of the results of the EES community they focus on usability, and in case there are human users involved, a fully specified application somehow will entail human factors evaluation, immediately making [47]'s discussion of desideratum $\mathcal{D}_2$ relevant. When evaluating AI explanation systems with human subjects, we should be aware of the users' potential cognitive biases [39]. In the context of explainability, it is interesting to note that users may suffer from biases, for example the interesting phenomenon 'choice blindness' discovered by Johansson et al. [18]. Choice blindness shows up in failure to make and explain consistent decisions. In the work of [18] magicians 'fool' users to explain decisions users did or did not make with similar strengths. Yet, we note that the importance of actual usability evaluation of explanation methods also appeared early. A 1990 AAAI workshop featured work on user scenarios [9] and later work was reported in [45].

Breiman equated simplicity and interpretability. However, it is well-known that seemingly simple models can be hard to interpret. Even simple linear classification models need careful tuning to optimize stability of feature importance maps [36]. The 'filters vs. pattern' discussion that first emerged in the context of neuroimaging is another example of unexpected complexity. In this context, there is an important difference between visualizing the classification model and the corresponding (causal) generative model. The difference is induced by correlated input noise and can lead to wrong conclusions if not handled appropriately as pointed out by Haufe et al. [17]. Similar challenges appear in deep networks [20]. Further examples of the dissociation of simplicity and interpretability are discussed by Lipton [25,26], citing the work on 'Interpretable Boosted Naïve Bayes Classification' by [38]. This paper opens with a statement aligned with the Breiman's dilemma: 'Efforts to develop classifiers with strong discrimination power using voting methods have marginalized the importance of comprehensibility'. The objective of the paper is to demonstrate that the interpretation problem for voting systems can be mitigated. Specifically, Madigan et al.'s tools for interpretation in Naive Bayes classifiers [27]' is shown to be useful for complex boosting ensembles.

3.3 Quantify Similarity and Uncertainty of Feature Importance Maps Using Resampling

In certain application domains of neural networks, including scientific computing and bio-medicine, interpretation have played an important role for long and tools have been developed for explanation of neural networks' function.

In early work on mind reading based on brain scanning interpretability was naturally in focus [24,31]. The dominating analysis paradigm at the time was Friston et al.'s SPM 'statistical parametric mapping' [11]. This approach produced intuitively appealing three dimensional brain maps of voxel-wise significant activation. These maps have had significant impact in the field and it was a strong aim of neural networks visualizations to produce matching SPMs. The tools developed included 3D mapping of voxel-wise saliency [32] and sensitivity [21]. The usability ($\mathcal{D}_2$) for neuroscientists was enhanced by embedding the maps in 3D 'anatomically informed' navigation tools, see e.g., [32] for examples. The brain map visualizations were further enriched with knowledge graph information to facilitate decision making and teaching [35].

Concerning the first desideratum ($\mathcal{D}_1$) - fidelity of explanations - we need to consider the two logical fundamental properties: 'Existence' and 'uniqueness'. Considering the many constraints imposed by the desiderata, the very existence of a satisfactory interpretability scheme is a non-trivial issue. Finding such schemes is the concern of current interpretability engineering literature. Given existence, we face an equally important issue of uniqueness. Note that at least two mechanisms of uncertainty can contribute to non-uniqueness: Firstly, epistemic uncertainty, i.e., uncertainty in the explainability model, typically induced by a combination of limited data and knowledge. Epistemic uncertainty gives rise to multiple competing paradigm of explainability. The second source of non-uniqueness is the inherent randomness of a given problem domain for which noise and finite samples can conspire to create large fluctuations in solutions ('aleatory uncertainty').

Epistemic uncertainty was discussed in detail in the work of Lange et al. [23]. Nine different interpretation schemes were evaluated to explore the diversity in model space and learn similarities. The idiosyncratic scales employed by different mapping procedures is a significant challenge for quantitative comparisons of visualizations. This problem was addressed in [16] proposing a simple nonparametric approach to standardization of maps, hence, allowing maps to be meaningfully combined, e.g., by simple averaging. Such consensus based methods allow reduction of model uncertainty and quantification of inter-map differences (epistemic uncertainty).

Aleatory uncertainty in brain maps was addressed by the so-called NPAIRS framework [44]. Statistical re-sampling techniques such as split-half, can provide unbiased estimates of variance of interpretability heat maps. This allows for mapping of the local visualization 'effect size', by scaling heat maps by their standard deviation. Application of these tools include imaging pipeline optimization [22]. Outside the original domain of these methods, they have been applied for skin cancer diagnosis support [43]. We foresee that future applications of deep learning within scientific computing will call for careful and unbiased quantification of aleatory uncertainty and methods to aggregate multiple explanation hypotheses.

3.4 Concluding Remarks

Explainability is at the core of modern machine learning. The transparency made possible by effective tools for explainability can improve design and debugging for the machine learning engineer and even more importantly, our users' trust and usability in the tools we develop. It would be productive if the very active community of scientist working in this field made an even bigger effort to embrace the rich 50+ year history of explainable AI. Here we focused on two specific topics from this legacy that could enrich current interpretability work: (1) Careful definition of the task via *Explanation desiderata*. Here we pointed to a rich set of ideas developed in the 'explainable expert systems' field and (2) Careful handling of *uncertainty*. Here we pointed to the comprehensive workflows for quantification of uncertainty of high-dimensional feature importance maps, originally developed in the field of computational neuroimaging.

References

1. Andrews, R., Diederich, J., Tickle, A.B.: Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl. Based Syst. **8**(6), 373–389 (1995)
2. Boz, O.: Converting a trained neural network to a decision tree dectext-decision tree extractor (2000)
3. Breiman, L.: Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. **16**(3), 199–231 (2001)
4. Bruce, G., Buchanan, B., Shortliffe, E.: Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project. Addison-Wesley, Reading (1984)
5. Carbonell, J.R.: AI in CAI: an artificial-intelligence approach to computer-assisted instruction. IEEE Trans. Man Mach. Syst. **11**(4), 190–202 (1970)
6. Comon, P.: Independent component analysis, a new concept? Sign. Proc. **36**(3), 287–314 (1994)
7. Craven, M.W., Shavlik, J.W.: Using sampling and queries to extract rules from trained neural networks. In: Machine Learning Proceedings 1994, pp. 37–45. Elsevier (1994)
8. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint. arXiv:1702.08608 (2017)
9. Druzdzel, M.J., Henrion, M.: Using scenarios to explain probabilistic inference. In: Working notes of the AAAI-1990 Workshop on Explanation, pp. 133–141 (1990)
10. Duda, R.O., Shortliffe, E.H.: Expert systems research. Science **220**(4594), 261–268 (1983)
11. Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.P., Frith, C.D., Frackowiak, R.S.: Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. **2**(4), 189–210 (1994)
12. Gallant, S.I.: Connectionist expert systems. Commun. ACM **31**(2), 152–169 (1988)
13. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: an approach to evaluating interpretability of machine learning. arXiv preprint. arXiv:1806.00069 (2018)
14. Good, I.: Explicativity: a mathematical theory of explanation with statistical applications. Proc. R. Soc. Lond. A **354**(1678), 303–330 (1977)
15. Goodman, B., Flaxman, S.: European union regulations on algorithmic decision-making and a "right to explanation". arXiv preprint. arXiv:1606.08813 (2016)
16. Hansen, L.K., Nielsen, F.Å., Strother, S.C., Lange, N.: Consensus inference in neuroimaging. NeuroImage **13**(6), 1212–1218 (2001)
17. Haufe, S., et al.: On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage **87**, 96–110 (2014)

18. Johansson, P., Hall, L., Sikström, S., Olsson, A.: Failure to detect mismatches between intention and outcome in a simple decision task. Science **310**(5745), 116–119 (2005)
19. Kim, B., et al.: Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning, pp. 2673–2682 (2018)
20. Kindermans, P.J., et al.: Learning how to explain neural networks: PatternNet and PatternAttribution. arXiv preprint. arXiv:1705.05598 (2017)
21. Kjems, U., et al.: The quantitative evaluation of functional neuroimaging experiments: mutual information learning curves. NeuroImage **15**(4), 772–786 (2002)
22. LaConte, S., et al.: The evaluation of preprocessing choices in single-subject bold fMRI using NPAIRS performance metrics. NeuroImage **18**(1), 10–27 (2003)
23. Lange, N., et al.: Plurality and resemblance in fMRI data analysis. NeuroImage **10**(3), 282–303 (1999)
24. Lautrup, B., Hansen, L.K., Law, I., Mørch, N., Svarer, C., Strother, S.C.: Massive weight sharing: a cure for extremely ill-posed problems. In: Workshop on Supercomputing in Brain Research: From Tomography to Neural Networks, pp. 137–144 (1994)
25. Lipton, Z.C.: The mythos of model interpretability. arXiv preprint. arXiv:1606.03490 (2016)
26. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 30 (2018)
27. Madigan, D., Mosurski, K., Almond, R.G.: Graphical explanation in belief networks. J. Comput. Graph. Stat. **6**(2), 160–181 (1997)
28. Michie, D.: Machine learning in the next five years. In: Proceedings of the 3rd European Conference on European Working Session on Learning, pp. 107–122. Pitman Publishing (1988)
29. Minsky, M.L.: Logical versus analogical or symbolic versus connectionist or neat versus scruffy. AI Mag. **12**(2), 34 (1991)
30. Montavon, G., Samek, W., Müller, K.R.: Methods for interpreting and understanding deep neural networks. Digital Sign. Proces. **73**, 1–15 (2018)
31. Mørch, N., et al.: Nonlinear versus linear models in functional neuroimaging: learning curves and generalization crossover. In: Duncan, J., Gindi, G. (eds.) IPMI 1997. LNCS, vol. 1230, pp. 259–270. Springer, Heidelberg (1997). https://doi.org/10.1007/3-540-63046-5_20
32. Mørch, N.J., et al.: Visualization of neural networks using saliency maps. In: 1995 IEEE International Conference on Neural Networks. IEEE (1995)
33. Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., Doshi-Velez, F.: How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. arXiv preprint. arXiv:1802.00682 (2018)
34. Neches, R., Swartout, W.R., Moore, J.D.: Enhanced maintenance and explanation of expert systems through explicit models of their development. IEEE Trans. Softw. Eng. **11**, 1337–1351 (1985)
35. Nielsen, F.A., Hansen, L.K.: Automatic anatomical labeling of Talairach coordinates and generation of volumes of interest via the brainmap database. NeuroImage **16**(2), 2–6 (2002)
36. Rasmussen, P.M., Hansen, L.K., Madsen, K.H., Churchill, N.W., Strother, S.C.: Model sparsity and brain pattern interpretation of classification models in neuroimaging. Pattern Recogn. **45**(6), 2085–2100 (2012)

37. Ribeiro, M.T., Singh, S., Guestrin, C.: 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, pp. 1135–1144. ACM (2016)
38. Ridgeway, G., Madigan, D., Richardson, T., O'Kane, J.: Interpretable boosted Naïve Bayes classification. In: KDD, pp. 101–104 (1998)
39. Saposnik, G., Redelmeier, D., Ruff, C.C., Tobler, P.N.: Cognitive biases associated with medical decisions: a systematic review. BMC Med. Inform. Decis. Mak. **16**(1), 138 (2016)
40. Schütt, K.T., Arbabzadah, F., Chmiela, S., Müller, K.R., Tkatchenko, A.: Quantum-chemical insights from deep tensor neural networks. Nat. Commun. **8**, 13890 (2017)
41. Shortliffe, E.H., Axline, S.G., Buchanan, B.G., Merigan, T.C., Cohen, S.N.: An artificial intelligence program to advise physicians regarding antimicrobial therapy. Comput. Biomed. Res. **6**(6), 544–560 (1973)
42. Shortliffe, E., Davis, R., Axline, S., Buchanan, B., Green, C., Cohen, S.: Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput. Biomed. Res. **8**(4), 303–320 (1975)
43. Sigurdsson, S., Philipsen, P.A., Hansen, L.K., Larsen, J., Gniadecka, M., Wulf, H.C.: Detection of skin cancer by classification of Raman spectra. IEEE Trans. Biomed. Eng. **51**(10), 1784–1793 (2004)
44. Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., Rottenberg, D.: The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. NeuroImage **15**(4), 747–771 (2002)
45. Suermondt, H.J., Cooper, G.F.: An evaluation of explanations of probabilistic inference. In: Proceedings of the Annual Symposium on Computer Application in Medical Care, p. 579. American Medical Informatics Association (1992)
46. Swartout, W.R.: XPLAIN: a system for creating and explaining expert consulting programs. University of Southern California Marina del Rey Information Sciences Institute, Technical report (1983)
47. Swartout, W.R., Moore, J.D.: Explanation in second generation expert systems. In: David, J.M., Krivine, J.P., Simmons, R. (eds.) Second Generation Expert Systems, pp. 543–585. Springer, Heidelberg (1993). https://doi.org/10.1007/978-3-642-77927-5_24
48. Thrun, S.: Extracting provably correct rules from artificial neural networks. Technical report IAI-TR-93-5, Institut for Informatik III Universitat Bonn, Germany (1994)
49. Thrun, S.: Extracting rules from artificial neural networks with distributed representations. In: Advances in Neural Information Processing Systems, pp. 505–512 (1995)
50. Tomsett, R., Braines, D., Harborne, D., Preece, A., Chakraborty, S.: Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. arXiv preprint. arXiv:1806.07552 (2018)