Self-explaining AI as an alternative to interpretable AI

Daniel C. Elton[1]

(1) Imaging Biomarkers and Computer-Aided Diagnosis Laboratory, Radiology and Imaging Sciences, National Institutes of Health Clinical Center, Bethesda, MD 20892, USA daniel.elton (at) nih.gov

Abstract. The ability to explain decisions made by AI systems is highly sought after, especially in domains where human lives are at stake such as medicine or autonomous vehicles. While it is always possible to approximate the input-output relations of deep neural networks with human-understandable rules or a post-hoc model, the discovery of the double descent phenomena suggests that no such approximation will ever map onto the actual mechanistic functioning of deep neural networks. Double descent indicates that deep neural networks typically operate by smoothly interpolating between data points rather than by extracting a few high level rules. As a result neural networks trained on complex real world data are inherently hard to interpret and prone to failure if used outside their domain of applicability (ie, for extrapolation). To show how we might be able to trust AI despite these problems, we introduce the concept of self-explaining AI. Self-explaining AIs are capable of providing a human-understandable explanation of each decision along with confidence levels for both the decision and explanation. Some difficulties to this approach along with possible solutions are sketched. Finally, we argue it is also important that AI systems warn their user when they are asked to perform outside their domain of applicability.

Keywords: Interpretability · explainability · explainable artificial intelligence · XAI · trust · deep learning

1 Introduction

There is growing interest in developing methods to explain deep neural network function, especially in high risk areas such as medicine and driverless cars. Such explanations would be useful to ensure that deep neural networks follow known rules and when troubleshooting failures. Somewhat controversially, the European Union's 2016 General Data Protection Regulation says that companies must be able to provide an explanation to consumers about decisions made by artificial intelligences [25], which has helped bolster growing interest on explainable AI and methods for interpreting deep neural network function. Despite the development of numerous techniques for interpreting deep neural networks, all such techniques have flaws, and there is confusion regarding how to properly "interpret an interpetation" [42,33]. Perhaps more troubling, though, is that a new

understanding is emerging that deep neural networks function through the interpolation of data points, rather than extrapolation [24]. This calls into question long-held narratives about deep neural networks "extracting" high level features and rules, and also indicates that all current methods of explanation do not capture failure modes that occur from extrapolation.

In response to difficulties raised by explaining black box models, Rudin argues for developing better interpretable models instead, arguing that the "interpetability-accuracy" trade-off is a myth. While it is true that the notion of such a trade-off is not rigorously grounded, empirically in many domains the state-of-the art systems are all deep neural networks. For instance, most state-of-art AI systems for computer vision are not interpretable in the sense required of Rudin. Even highly distilled and/or compressed models which achieve good performance on ImageNet require at least 100,000 free parameters [31]. Moreover, the human brain also appears to be an overfit "black box" which performs interpolation, which means that how we understand brain function also needs to change [24]. If evolution settled on a model (the brain) which is uninterpretable, then we expect advanced AIs to also be of that type. Interestingly, although the human brain is a "black box", we are able to trust each other. Part of this trust comes from our ability to "explain" our decision making in terms which make sense to us. Crucially, for trust to occur we must believe that a person is not being deliberately deceptive, and that their verbal explanations actually maps onto the processes used in their brain to arrive at their decisions.

Motivated by how trust works between humans, in this work we explore the idea of self-explaining AIs. Self-explaining AIs yield two outputs - the decision and an explanation of that decision. This idea is not new, and it is something which was pursued in expert systems research in the 1980s [48]. More recently Kulesza et al. introduced a model which offers explanations and studied how such models allow for "explainable debugging" and iterative refinement [28]. However, in their work they restrict themselves to a simple interpretable model (a multinomial naive Bayes classifier). Alvarez-Melis & Jaakkola introduce a "self-explaining" neural network which makes predictions using a number of human interpretable concepts or prototypes [4]. In a somewhat similar vein, Chen et al. [14] proposed a "This looks like That" network. Unlike previous works, in this work we explore how to create trustworthy self-explaining AI for networks and agents of arbitrary complexity. We also seek for a more rigorous way to make sure the explanation given is actually explaining an aspect of the mechanism used for prediction. Therefore, unlike previous works this work makes contact with the field of AI safety, including AI safety for artificial general intelligences (AGIs).

After defining key terms, we discuss the challenge of interpreting deep neural networks raised by recent studies on interpolation and generalization in deep neural networks. Then, we discuss how self-explaining AIs might be built. We argue that they should include at least three components - a measure of mutual information between the explanation and the decision, an uncertainty on both the explanation and decision, and a "warning system" which warns the user

when the decision falls outside the domain of applicability of the system. We hope this work will inspire further work in this area which will ultimately lead to more trustworthy AI.

1.1 Interpretation, explanation, and self-explanation

As has been discussed at length elsewhere, different practitioners understand the term "intepretability" in different ways, leading to a lack of clarity (for detailed reviews, see[33,2,36,5]). The related term "explainability" is typically used in a synonymous fashion [42], although some have tried to draw a distinction between the two terms [29]. Here we take explanation/explainability and interpretation/interpretability to be synonymous. Murdoch et al. define an **explanation** as a verbal account of neural network function which is descriptively accurate and relevant [36]. By "descriptively accurate" they mean that the interpretation reproduces a large number of the input-output mappings of the model. The explanation may or may not map onto how the model works internally. Additionally, any explanation will be an approximation, and the degree of approximation which is deemed acceptable may vary depending on application. By "relevance", what counts as a "relevant explanation" is domain specific – it must be cast in terminology that is both understandable and relevant to users. For deep neural networks, the two desiderata of accuracy and relevance appear to be in tension - as we try to accurately explain the details of how a deep neural network interpolates, we move further from what may be considered relevant to the user.

This definition of explanation in terms of capturing input-output mappings in a human understandable way contrasts with a second meaning of the term explanation which we may call **mechanistic explanation**. Mechanistic explanations abstract faithfully (but approximately) the actual data transformations occurring in the model. To consider why mechanistic explanations can be useful, consider a deep learning model we trained recently to segment the L1 vertebra [16]. The way a radiologist identifies the L1 vertebra is by scanning down from the top of the body and finding the last vertebra that has ribs attached to it, which is T12. L1 is directly below T12. In our experience our models for identifying L1 tend to be brittle, indicating they probably use a different approach. For instance, they may do something like "locate the bright object in the middle of the image" or "locate the bright object which is just above the kidneys". These techniques would not be as robust as the technique used by radiologists. If a self-explaining AI or AGI had a model of human anatomy and could couch its explanations with reference to standard anatomical concepts, that would go a long way towards engendering trust. In general, the "Rashomon Effect", first described by Leo Brieman [13], says that for any set of noisy data, there are a multitude of models of equivalent accuracy, but which differ significantly in their internal mechanism. As a real-world example of the Rashomon Effect, when detecting Alzheimer's disease in brain MRI using a CNN the visualized interpretations for models trained on different train-test folds differed significantly, even though the models were of equivalent accuracy [47]. Even more troubling,

the visualizations differed between different runs on the same fold, with the only difference being in the random initialization of the network [47]. Finally, interpretations can vary between test examples [8], and in many works only a few examples (sometimes cherry-picked) are given, rather than attempting an analysis of the interpretation method on the entire test set. In summary, in deep neural networks generally the mechanism of prediction can differ greatly between models of equivalent accuracy, even when the models all have the same architecture, due to peculiarities of the training data and initialization used. Additionally, the specific details of the mechanism may vary wildly within a given model across different test cases.

There is another type of explanation we wish to discuss which we may call **meta-level explanation**. Richard P. Feynman said "What I cannot create, I do not understand". Since we can create deep neural networks, we do understand them, in the sense of Feynman, and therefore we can explain them in terms of how we build them. More specifically, we can explain neural network function in terms of four components necessary for creating them - data, network architecture, learning rules, and objective functions [40]. The way one explains deep neural network function from data, architecture, and training is analogous to how one explains animal behaviour using the theory of evolution. The evolution of architectures by "graduate student descent" and the explicit addition of inductive biases mirrors the evolution of organisms. Similarly, the training of architectures mirrors classical conditioning of animals as they get older. The explanation of animal behaviour in terms of meta-level theories like evolution and classical conditioning has proven to be enormously successful and stands in contrast to attempts to seek detailed mechanistic accounts.

Finally, the oft-used term **black box** also warrants discussion. The term is technically a misnomer since the precise workings of deep networks are fully transparent from their source code and network weights, and therefore for sake of rigor should not be used. A further point is that even if we did not have access to the source code or weights (for instance for intellectual property reasons, or because the relevant technical expertise is missing), it is likely that a large amount of information about the network's function could be gleaned through careful study of the its input-output relations. Developing mathematically rigorous techniques for "shining lights" into "black boxes" was a popular topic in early cybernetics research [6], and this subject is attracting renewed interest in the era of deep learning. As an example of what is achievable, recently it has been shown that weights can be inferred for ReLU networks through careful analysis of input-output relations [41]. One way of designing a "self-explaining AI" would be to imbue the AI with the power to probe its own input-output relations so it can warn its user when it may be making an error and (ideally) also distill its functioning into a human-understandable format.

1.2 Why deep neural networks are generally non-interpretable

Many methods for interpretation of deep neural networks have been developed, such as sensitivity analysis (saliency maps, occlusion maps, etc.), iterative map-

ping [11], "distilling" a neural network into a simpler model [18], exploring failure modes and adversarial examples [20,22], visualizing filters in CNNs [51], activation maximization based visualizations [17], influence functions [27], Shapley values [34], Local Interpretable Model-agnostic Explanations (LIME) [39], DeepLIFT [45], explanatory graphs [53], and layerwise relevance propagation [7]. Yet, all of these methods capture only particular aspects of neural network function, and the outputs of these methods are very easy to misinterpret [42,30,50]. Often the output of interpretability methods vary largely between test cases, but only a few "representative" cases (often hand picked) are shown in papers. Moreover, it has been shown that popular methods such as LIME [4], Shapley values [4], and saliency maps [15,50,1] are not robust to small changes in the image such as Gaussian noise.

As we discussed before, we do not expect the current push towards more interpretable models led by Rudin and others to be successful in general - deep neural networks are here to stay, and they will become even more complex and inscrutable as time goes on. Lillicrap & Kording [31] note that attempts to compress deep neural networks into a simpler interpretable models with equivalent accuracy typically fail when working with complex real world data such as images or human language. If the world is messy and complex, then neural networks trained on real world data will also be messy and complex. Leo Breiman, who equates interpretability with simplicity, has made a similar point in the context of random forest models [13]. In many domains, the reason machine learning is applied is because of the failure of simple models or because of the infeasibility of physics-based simulation. While we agree with Rudin that the interpretability-accuracy trade-off is not based on any rigorous quantitative analysis, we see much evidence to support it, and in some limiting cases (for example superintelligent AGIs or brain emulations, etc) the truth of such a trade-off becomes clear.

On top of these issues, there is a more fundamental reason to believe it will be hard to give mechanistic explanations for deep neural network function. For some years now it has been noted that deep neural networks have enormous capacity and seem to be vastly underdetermined, yet they still generalize. This was shown very starkly in 2016 when in Zhang et al. showed how deep neural networks can memorize random labels on ImageNet images [52]. More recently it has been shown that deep neural networks operate in a regime where the bias-variance trade-off no-longer applies [9]. As network capacity increases, test error first bottoms out and then starts to increase, but then (surprisingly) starts to decrease after a particular capacity threshold is reached. Belkin et al. call this the "double descent phenomena" [9] and it was also noted in an earlier paper by Sprigler et al [46], who argue the phenomena is analogous to the "jamming transition" found in the physics of granular materials. The phenomena of "double descent" appears to be universal to all machine learning [9,10], although its presence can be masked by common practices such as early stopping [9,37], which may explain why it took so long to be discovered.

In the regime where deep neural networks operate, they not only interpolate each training data point, but do so in a "direct" or "robust" way [24]. This means that the interpolation does not exhibit the overshoot or undershoot which is typical of overfit models, rather it is almost a piecewise interpolation. Interpolation brings brings with it a corollary - the inability to extrapolate. The fact that deep neural networks cannot extrapolate calls into question popular ideas that deep neural networks "extract" high level features and "discover" regularities in the world. Actually, deep neural networks are "dumb" - any regularities that they appear to have captured internally are solely due to the data that was fed to them, rather than a self-directed "regularity extraction" process.

1.3 How can we trust a self-explaining AI's explanation?

In his landmark 2014 book *Superintelligence: Paths, Dangers, Strategies*, Nick Bostrom notes that highly advanced AIs may be incentivized to deceive their creators until a point where they exhibit a "treacherous turn" against them [12]. In the case of superintelligent or otherwise highly advanced AI, the possibility of deception appears to be a highly non-trivial concern. Here however, we suggest some methods by which we can trust the explanations given by present day deep neural networks, such as typical convolutional neural networks or transformer language models. Whether these methods will still have utility when it comes to future AI & AGI systems is an open question.

To show how we might create trust, we focus on an explicit and relatively simple example. Shen et al. [44] and later LaLonde et al. [29] have both proposed deep neural networks for lung nodule classification which offer "explanations". Both authors make use of a dataset where clinicians have labeled lung nodules not only by severity (cancerous vs. non-cancerous) but also quantified them (on a scale of 1-5) in terms of five visual attributes which are deemed relevant for diagnosis (subtlety, sphericity, margin, lobulation, spiculation, and texture). While the details of the proposed networks vary, both output predictions for severity and scores for each of the visual attributes. Both authors claim that the visual attribute predictions "explain" the diagnostic prediction, since the diagnostic branch and visual attribute prediction branch(es) are connected near the base of the network. However, no evidence is presented that the visual attribute prediction is in any way related to the diagnosis prediction. While it may seem intuitive that the two output branches must be related, this must be rigorously shown for trustworthiness to hold.[1] Additionally, even if the visual attributes were used, no weights ("relevances") are provided for the importance of each attribute to the prediction, and there may be other attributes of equal or greater importance that are used but not among those outputted (this point is admitted and discussed by Shen et al. [44]).

Therefore, we would like to determine the degree to which the attributes in the explanation branch are responsible for the prediction in the diagnosis branch.

[1] Non-intuitive behaviours have repeatably been demonstrated in deep neural networks, for instance it has been shown networks based on rectified linear units contain unexpectedly large "linear regions" [23].

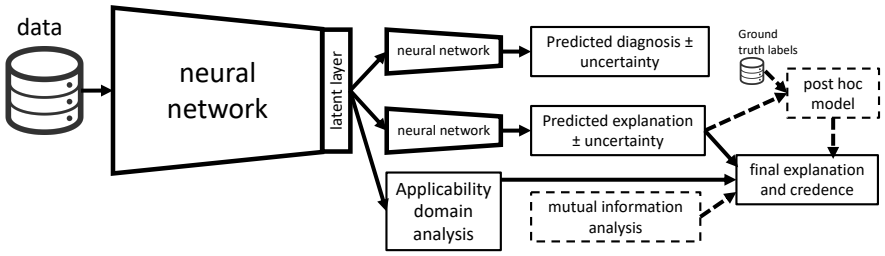

Fig. 1. Sketch of a simple self-explaining AI system. Optional components are shown with dashed lines.

We focus on the layer where the diagnosis and explanation branch diverge and look at how the output of each branch relates to activations in that layer. There are many ways of quantifying the relatedness of two variables, the Pearson correlation being one of the simplest, but also one of the least useful in this context since it is only sensitive to linear relationships. A measure which is sensitive to non-linear relationships and which has nice theoretical interpretation is the mutual information. For two random variables X and Y it is defined as:

$$\begin{aligned} \mathrm{MI}(X,Y) &\equiv \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &= H(x,y) - H(x) - H(y) \end{aligned} \tag{1}$$

Where $H(x)$ is the Shannon entropy. One can also define a mutual information correlation coefficient:[32]

$$r^{\mathrm{MI}}(X,Y) = \sqrt{1 - e^{-2\ \mathrm{MI(X,Y)}}} \tag{2}$$

This coefficient has the nice property that it reduces to the Pearson correlation in the case that $P(x,y)$ is a Gaussian function with non-zero covariance. The chief difficulty of applying mutual information is that the underlying probability distributions $P(x,y)$, $P(x)$, and $P(y)$ all have to be estimated. Various techniques exist for doing this however, such as by using kernel density estimation with Parzen windows [49].[2]

Suppose the latent vector is denoted by $\boldsymbol{L}$ and has length N. Denote the diagnosis of the network as D and the vector of attributes $\boldsymbol{A}$. Then for a particular attribute A_j in our explanation word set we calculate the following to obtain a

[2] Note that this sort of approach should not be taken as quantifying "information flow" in the network. In fact, since the output of units is continuous, the amount of information which can flow through the network is infinite (for discussion and how to recover the concept of "information flow" in neural networks see [21]). What we propose to measure is the the mutual information over the data distribution used.

"relatedness" score between the two:

$$R(A_j) = \sum_{i}^{N} \mathrm{MI}(L_i, D)\mathrm{MI}(L_i, A_j) \tag{3}$$

An alternative (an perhaps complimentary) method is to train a "post-hoc" model to try to predict the diagnosis from the attributes (also shown in figure 1). While this cannot tell us much about mechanism of the main model (due to the Rashomon effect) we can learn a bit from it. Namely, if the post-hoc model is not as accurate as the diagnosis branch of the main model, then we know the main model is using additional features.

1.4 Ensuring robustness through applicability domain and uncertainty analysis

The concept of an "applicability domain", or the domain where a model makes good predictions, is well known in the area of molecular modeling known as quantitative structure property relationships (QSPR), and a number of techniques have been developed (for a review, see [43] or [38]). However, the practice of quantifying the applicability domain of models hasn't become widespread in other areas where machine learning is applied. A simple way of defining the applicability domain is to calculate the convex hull of the latent vectors for all training data points. If the latent vector of a test data point falls on or outside the convex hull, then the model should send an alert saying that the test point falls outside the domain it was trained for. Applicability domain analysis is a relatively simple form of AI self-awareness, which is thought to be an important component of AI safety for advanced AIs and AGIs [3].

Finally, models should contain measures of uncertainty for both their decisions and their explanations. Ideally, this should be done in a Bayesian way using a Bayesian neural network [26]. With the continued progress of Moore's law, training Bayesian CNNs [35] is now becoming feasible and in our view this is a worthwhile use of additional CPU/GPU cycles. There are also approximate methods - for instance it has been shown that random dropout during inference can be used to estimate uncertainties at little extra computational cost [19]. Just as including experimental error bars is standard in all of science, and just as we wouldn't trust a doctor who could not also give a confidence level in his diagnosis, uncertainty quantification should be standard practice in AI research.

1.5 Conclusion

We argued that deep neural networks trained on complex real world data are very difficult to interpret due to their power arising from brute-force interpolation over big data rather than through the extraction of high level generalizable rules. Motivated by this and by the need for trust in AI systems we introduced the concept of self-explaining AI and described how a simple self-explaining AI would function for diagnosing medical images. To build trust, we showed how a

mutual information metric can be used to verify that the explanation given is related to the diagnostic output. Crucially, in addition to an explanation, self-explaining AI outputs confidence levels for both the decision and explanation, further aiding our ability to gauge the trustworthiness of any given diagnosis or decision. Finally, an applicability domain analysis should be done for AI systems where robustness and trust are important, so that systems can alert their user if they are asked work outside their domain of applicability.

2 Funding & disclaimer

No funding sources were used in the creation of this work. The author (Dr. Daniel C. Elton) wrote this article in his personal capacity. The opinions expressed in this article are the author's own and do not reflect the view of the National Institutes of Health, the Department of Health and Human Services, or the United States government.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 9525–9536. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
2. Ahmad, M.A., Eckert, C., Teredesai, A.: Interpretable machine learning in healthcare. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18. ACM Press (2018). https://doi.org/10.1145/3233547.3233667
3. Aliman, N.M., Kester, L.: Hybrid strategies towards safe self-aware superintelligent systems. In: Artificial General Intelligence, pp. 1–11. Springer International Publishing (2018)
4. Alvarez-Melis, D., Jaakkola, T.S.: Towards robust interpretability with self-explaining neural networks. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. p. 7786–7795. NIPS'18, Curran Associates Inc., Red Hook, NY, USA (2018)
5. Arya, V., Bellamy, R.K.E., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K.R., Wei, D., Zhang, Y.: One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques (2019)
6. Ashby, W.R.: An introduction to cybernetics. London : Chapman & Hall (1956)
7. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLOS ONE **10**(7), e0130140 (Jul 2015). https://doi.org/10.1371/journal.pone.0130140
8. Barnes, B.C., Elton, D.C., Boukouvalas, Z., Taylor, D.E., Mattson, W.D., Fuge, M.D., Chung, P.W.: Machine learning of energetic material properties (2018)
9. Belkin, M., Hsu, D., Ma, S., Mandal, S.: Reconciling modern machine-learning practice and the classical bias–variance trade-off. Proceedings

of the National Academy of Sciences **116**(32), 15849–15854 (Jul 2019). https://doi.org/10.1073/pnas.1903070116
10. Belkin, M., Hsu, D., Xu, J.: Two models of double descent for weak features. arXiv eprints: 1903.07571 (2019)
11. Bordes, F., Berthier, T., Jorio, L.D., Vincent, P., Bengio, Y.: Iteratively unveiling new regions of interest in deep learning models. In: Medical Imaging with Deep Learning (MIDL) (2018)
12. Bostrom, N.: Superintelligence: Paths, Dangers, Strategies. Oxford University Press, Inc., USA, 1st edn. (2014)
13. Breiman, L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science **16**(3), 199–231 (Aug 2001). https://doi.org/10.1214/ss/1009213726
14. Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., Su, J.: This looks like that: Deep learning for interpretable image recognition. In: Wallach, H.M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E.B., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. pp. 8928–8939 (2019)
15. Dombrowski, A.K., Alber, M., Anders, C.J., Ackermann, M., Müller, K.R., Kessel, P.: Explanations can be manipulated and geometry is to blame (2019)
16. Elton, D.C., Sandfort, V., Pickhardt, P.J., Summers, R.M.: Accurately identifying vertebral levels in large datasets. In: In Proceedings of the SPIE: Medical Imaging. (2020)
17. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. Tech. Rep. 1341, University of Montreal (2009), also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.
18. Frosst, N., Hinton, G.: Distilling a neural network into a soft decision tree. arXiv eprintss: 1711.09784 (2017)
19. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016)
20. Goertzel, B.: Are there deep reasons underlying the pathologies of today's deep learning algorithms? In: Artificial General Intelligence, pp. 70–79. Springer International Publishing (2015)
21. Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B., Polyanskiy, Y.: Estimating information flow in deep neural networks. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 2299–2308. PMLR, Long Beach, California, USA (09–15 Jun 2019)
22. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv eprintss: 1412.6572 (2014)
23. Hanin, B., Rolnick, D.: Deep ReLU networks have surprisingly few activation patterns. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 359–368. Curran Associates, Inc. (2019)
24. Hasson, U., Nastase, S.A., Goldstein, A.: Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. Neuron **105**(3), 416–434 (Feb 2020). https://doi.org/10.1016/j.neuron.2019.12.002

25. Kaminski, M.E.: The right to explanation, explained. SSRN Electronic Journal (2018). https://doi.org/10.2139/ssrn.3196985, https://doi.org/10.2139/ssrn.3196985
26. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 5580–5590. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
27. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning - Volume 70. p. 1885–1894. ICML'17, JMLR.org (2017)
28. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI '15. ACM Press (2015)
29. LaLonde, R., Torigian, D., Bagci, U.: Encoding Visual Attributes in Capsules for Explainable Medical Diagnoses. arXiv e-prints: 1909.05926 (Sep 2019)
30. Lie, C.: Relevance in the eye of the beholder: Diagnosing classifications based on visualised layerwise relevance propagation. Master's thesis, Lund Unversity, Sweden (2019)
31. Lillicrap, T.P., Kording, K.P.: What does it mean to understand a neural network? arXiv eprints: 1907.06374 (2019)
32. Linfoot, E.: An informational measure of correlation. Information and Control **1**(1), 85 – 89 (1957). https://doi.org/10.1016/S0019-9958(57)90116-X
33. Lipton, Z.C.: The mythos of model interpretability. arXiv eprints: 1606.03490 (2016)
34. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
35. McClure, P., Rho, N., Lee, J.A., Kaczmarzyk, J.R., Zheng, C.Y., Ghosh, S.S., Nielson, D.M., Thomas, A.G., Bandettini, P., Pereira, F.: Knowing what you know in brain segmentation using Bayesian deep neural networks. Frontiers in Neuroinformatics **13** (Oct 2019). https://doi.org/10.3389/fninf.2019.00067
36. Murdoch, W.J., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B.: Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences **116**(44), 22071–22080 (Oct 2019). https://doi.org/10.1073/pnas.1900654116
37. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., Sutskever, I.: Deep double descent: Where bigger models and more data hurt. arXiv eprints: 1912.02292 (2019)
38. Netzeva, T.I., Worth, A.P., Aldenberg, T., Benigni, R., Cronin, M.T., Gramatica, P., Jaworska, J.S., Kahn, S., Klopman, G., Marchant, C.A., Myatt, G., Nikolova-Jeliazkova, N., Patlewicz, G.Y., Perkins, R., Roberts, D.W., Schultz, T.W., Stanton, D.T., van de Sandt, J.J., Tong, W., Veith, G., Yang, C.: Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. Alternatives to Laboratory Animals **33**(2), 155–173 (Apr 2005). https://doi.org/10.1177/026119290503300209
39. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you? In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD. ACM Press (2016). https://doi.org/10.1145/2939672.2939778

40. Richards, B.A., Lillicrap, T.P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R.P., de Berker, A., Ganguli, S., Gillon, C.J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G.W., Miller, K.D., Naud, R., Pack, C.C., Poirazi, P., Roelfsema, P., Sacramento, J., Saxe, A., Scellier, B., Schapiro, A.C., Senn, W., Wayne, G., Yamins, D., Zenke, F., Zylberberg, J., Therien, D., Kording, K.P.: A deep learning framework for neuroscience. Nature Neuroscience **22**(11), 1761–1770 (Oct 2019). https://doi.org/10.1038/s41593-019-0520-2
41. Rolnick, D., Kording, K.P.: Identifying weights and architectures of unknown relu networks. arXiv eprintss: 1910.00744 (2019)
42. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence **1**(5), 206–215 (May 2019). https://doi.org/10.1038/s42256-019-0048-x
43. Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., Consonni, V., Todeschini, R.: Comparison of different approaches to define the applicability domain of QSAR models. Molecules **17**(5), 4791–4810 (Apr 2012). https://doi.org/10.3390/molecules17054791
44. Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W.: An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Systems with Applications **128**, 84–95 (Aug 2019). https://doi.org/10.1016/j.eswa.2019.01.048
45. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. arXiv eprintss: 1704.02685 (2017)
46. Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., Wyart, M.: A jamming transition from under- to over-parametrization affects generalization in deep learning. Journal of Physics A: Mathematical and Theoretical **52**(47), 474001 (Oct 2019). https://doi.org/10.1088/1751-8121/ab4c8b
47. Sutre, E.T., Colliot, O., Dormont, D., Burgos, N.: Visualization approach to assess the robustness of neural networks for medical image classification. In: In Proceedings of the SPIE: Medical Imaging. (2020)
48. Swartout, W.R.: XPLAIN: a system for creating and explaining expert consulting programs. Artificial Intelligence **21**(3), 285–325 (Sep 1983). https://doi.org/10.1016/s0004-3702(83)80014-9
49. Torkkola, K.: Feature extraction by non-parametric mutual information maximization. Journal of Machine Learning Research **3**, 1415–1438 (2003)
50. Yeh, C.K., Hsieh, C.Y., Suggala, A.S., Inouye, D.I., Ravikumar, P.: On the (in)fidelity and sensitivity for explanations. arXiv eprints: 1901.09392 (2019)
51. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision – ECCV 2014, pp. 818–833. Springer International Publishing (2014)
52. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. arXiv eprints: 1611.03530 (2016)
53. Zhang, Q., Cao, R., Shi, F., Wu, Y.N., Zhu, S.C.: Interpreting cnn knowledge via an explanatory graph. In: McIlraith, S.A., Weinberger, K.Q. (eds.) AAAI. pp. 4454–4463. AAAI Press (2018)