The Mythos of Model Interpretability

Article *in* Communications of the ACM · October 2016

DOI: 10.1145/3233231

1 author:

Zachary Chase Lipton
Carnegie Mellon University

The Mythos of Model Interpretability

Zachary C. Lipton ZLIPTON@CS.UCSD.EDU
University of California, San Diego 9500 Gilman Drive, La Jolla, CA 92093 USA

Abstract

Supervised machine learning models boast remarkable predictive capabilities. But can you trust your model? Will it work in deployment? What else can it tell you about the world? We want models to be not only good, but interpretable. And yet the task of *interpretation* appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable. Despite this ambiguity, many papers proclaim interpretability axiomatically, absent further explanation. In this paper, we seek to refine the discourse on interpretability. First, we examine the motivations underlying interest in interpretability, finding them to be diverse and occasionally discordant. Then, we address model properties and techniques thought to confer interpretability, identifying transparency to humans and post-hoc explanations as competing notions. Throughout, we discuss the feasibility and desirability of different notions, and question the oft-made assertions that linear models are interpretable and that deep neural networks are not.

1. Introduction

As machine learning models penetrate critical areas like medicine, the criminal justice system, and financial markets, the inability of humans to understand these models seems problematic (Caruana et al., 2015; Kim, 2015). Some suggest *model interpretability* as a remedy, but few articulate precisely *what* interpretability means or *why* it is important. Despite the absence of a definition, papers frequently make claims about the interpretability of various models. From this, we might conclude either that the definition of interpretability is universally agreed upon or that the term *interpretability* is ill-defined. Our investigation of the literature suggests the latter to be the case. Both the motives for interpretability and the technical descriptions of interpretable models are diverse and occasionally discordant, suggesting that *interpretability* refers to more than one concept. In this paper, we seek to clarify both, suggesting that *interpretability* is not a monolithic concept, but in fact reflects several distinct ideas. We hope, through this critical analysis, to bring focus to the dialog.

Here, we consider supervised learning but not other machine learning paradigms, such as reinforcement learning and interactive learning. This scope derives from our original interest in the oft-made claim that linear models are preferable to deep neural networks on account of their interpretability (Lou et al., 2012). To gain conceptual clarity, we ask the refining questions: *What is interpretability and why is it important?* Broadening the scope of discussion seems counterproductive with respect to our aims. For research investigating interpretability in the context of reinforcement learning, we point to (Dragan et al., 2013) which studies the human interpretability of robot actions.

To contextualize any definition of interpretability, we first consider the motives that it addresses (expanded in §2). Many papers motivate interpretability as a means to engender trust (Kim, 2015; Ridgeway et al., 1998). But what precisely is trust? Some equate trust with understanding while others equate trust with confidence in a model's accuracy (Ribeiro et al., 2016). If trust connotes understanding, we still must ask, to what aspect of machine learning does it apply? Do we seek understandable features, parameters, models, or training algorithms?

Often, our machine learning problem formulations are imperfect matches for the real-life tasks they are meant to solve. This can happen when simplified optimization objectives fail to capture our more complex real-life goals. Consider medical research with longitudinal data. Our real goal may be to discover potentially causal associations, as with smoking and cancer (Wang et al., 1999). But the optimization objective for most supervised learning models is simply to minimize error.

Another such divergence of real-life and machine learning problem formulations emerges when the off-line training

2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, USA.

data for a supervised learner is not perfectly representative of the likely deployment environment. For example, the environment is typically not stationary. This is the case for product recommendation, as new products are introduced and preferences for some items shift daily. In more extreme cases, actions influenced by a model may alter the environment, invalidating future predictions.

Discussions of interpretability sometimes suggest that human decision-makers are themselves interpretable because they can explain their actions (Ridgeway et al., 1998). But precisely what notion of interpretability do these explanations satisfy? They seem unlikely to clarify the mechanisms or the precise algorithms by which brains work. Nevertheless, the information conferred by an interpretation may be useful. Thus one purpose of interpretations may be to convey useful information of any kind.

We then consider what properties of models might render them interpretable (expanded in §3). Some equate interpretability with *understandability* or *intelligibility* (Lou et al., 2013), i.e., that we can grasp *how the models work*. In these papers, understandable models are sometimes called *transparent*, while incomprehensible models are called *black boxes*. But what constitutes transparency? We might look to the algorithm itself. Will it converge? Does it produce a unique solution? Or we might look to its parameters: do we understand what each represents? Alternatively, we could consider the model's complexity. Is it simple enough to be examined all at once by a human?

Other papers investigate so-called post-hoc interpretations. These interpretations might *explain* predictions without elucidating the mechanisms by which models work. Examples of post-hoc interpretations include the verbal explanations produced by people or the saliency maps used to analyze deep neural networks. Thus, humans possess *post-hoc interpretability* despite being *black boxes* (non-transparent), revealing a contradiction between two popular notions of interpretability.

2. Motives for Interpretability

The desire for an *interpretation* presupposes that predictions alone do not suffice. This implies a discrepancy between our real-world objectives and the simple objectives optimized by most machine learning models. Typically, we train models to achieve strong predictive power, optimizing simple objectives such as accuracy or AUC. However, these objectives can be weak surrogates for the real-world goals of machine learning practitioners. This tension can arise when our real world objectives are difficult to encode as simple real-valued functions. For example, an algorithm for making hiring decisions should simultaneously optimize productivity, ethics, and legality. But ethics and legality don't admit direct optimization. The problem can also arise when the dynamics of the deployment environment differ from the training environment. Thus *interpretations* serve those objectives that we deem important but struggle to model formally.

Trust: Some papers motivate interpretability by suggesting it to be prerequisite for *trust* (Kim, 2015; Ribeiro et al., 2016). But what is trust? Is it simply confidence that a model will perform well? If so, a sufficiently accurate model should be demonstrably trustworthy and interpretability would serve no purpose. Trust might also be defined subjectively. For example, a person might feel more at ease with a well-understood model, even if this understanding served no obvious purpose. Alternatively, when the training and deployment objectives diverge, trust might denote confidence that the model will perform well with respect to the real objectives and scenarios. For example, consider a model used to allocate police officers. We may trust the model to predict accurately but not to respect ethics or legality. Or in the case of self-driving cars, we might not trust them to respond appropriately in unforeseen situations.

Causality: Although supervised learning models are only optimized directly to make associations, we might use them to infer properties or generate hypotheses about the natural world. For example, a simple regression model could reveal a strong association between thalidomide use and birth defects or smoking and lung cancer (Wang et al., 1999). Any association discovered by supervised models might not be causal. There could always exist unobserved causes responsible for both associated variables. However, interpretations yielded by supervised learning models can yield hypotheses that scientists could then test experimentally.

Transferability: Typically we choose training and test data by randomly partitioning examples from the same distribution. We then judge a model's generalization error by the gap between its performance on train and test data. However, humans exhibit a far richer capacity to generalize, transferring learned skills to unfamiliar situations. We already use machine learning algorithms in situations where such abilities are required, such as when the environment is non-stationary. We also deploy models in settings where their use might alter the environment, invalidating their future predictions. Along these lines, Caruana et al. (2015) describe a model trained to predict probability of death from pneumonia that assigned less risk to patients if they also had asthma. In fact, asthma was predictive of lower risk of death. This owed to the more aggressive treatment these patients received. But if the model were deployed to aid in triage, these patients would then receive less aggressive treatment, invalidating the model. Even worse, we could imagine situations, like machine learning for security, where the environment might be actively adversarial. Consider the recently discovered susceptibility of convolutional neural networks (CNNs) to adversarial examples. The CNNs were made to misclassify images that were imperceptibly (to a human) perturbed (Szegedy et al., 2013). Of course, this isn't overfitting in the classical sense. The results achieved on training data generalize well to i.i.d. test data. But these are mistakes a human wouldn't make and we would prefer models not to make these mistakes either.

Informativeness: Sometimes machine learning doesn't directly take actions, but instead provides assistance to human decision makers, a setting considered by Kim et al. (2015); Huysmans et al. (2011). While the machine learning objective might be to reduce error, the real-world purpose is to provide useful evidence. An interpretation may satisfy this purpose, even without shedding light on the model's inner workings. For example, a diagnosis model might provide intuition to a human decision-maker by pointing to similar cases in support of a diagnostic decision. In some cases, we train a supervised learning model, but our real task more closely resembles unsupervised learning. That is, our real goal is to explore the data and the objective might serve only as *weak supervision*.

3. Properties of Interpretable Models

We now consider model properties proposed either to enable or to comprise *interpretations*. These broadly fall into two categories. The first relates to *transparency*, i.e., *how does the model work?* The second consists of *post-hoc interpretations*, i.e., *what else can the model tell me?*

3.1. Transparency

Informally, *transparency* is the opposite of *opacity* or *blackbox-ness*. and connotes the *understandability* of the model. We consider transparency at the level of the entire model, at the level of individual components (e.g. parameters), and at the level of the training algorithm.

Simulatability: In the strictest sense, we might call a model transparent if a person can contemplate the entire model at once. In other words, a human could take the input data together with the parameters of the model and in *reasonable* time step through every calculation required to produce a prediction. While the quantity denoted by *reasonable* seems subjective, this ambiguity can only span a couple orders of magnitude. Few would suggest that a human could contemplate a thousand parameters at once, but most could simulate a linear model with 20 parameters. Thus, simulatability connotes low computational complexity. In this light, neither linear models nor rules, nor decision trees are intrinsically interpretable. Sufficiently high-dimensional models, unwieldy rule lists, and deep decision trees are no more interpretable than deep neural networks.

Decomposability: A second and less strict notion of transparency might be that each part of the model - each input, parameter, and calculation - admits an intuitive explanation. This accords with the property of *intelligibility* as described by (Lou et al., 2012). For example, each node in a decision tree might correspond to a plain text description (e.g. *all patients with diastolic blood pressure over 150*). Similarly, the parameters of a linear model could be described as representing strengths of association between each feature and the label. This notion of interpretability requires that inputs themselves be individually interpretable, disqualifying some models with highly engineered or anonymous features. While this notion is popular, we shouldn't accept it blindly. The weights of a linear model might seem intuitive, but they can be fragile with respect to feature selection. For example, associations between flu risk and vaccination can be positive or negative depending on whether the feature set includes indicators of old age, infancy, and immunodeficiency.

Algorithmic Transparency: A final notion of transparency might apply at the level of the algorithm, even absent the ability to mentally simulate a model or to intuit the meaning of its components. For example, in the case of linear models, we understand the shape of the error surface. We can prove that training will converge to a unique solution, even for previously unseen datasets. On the other hand, modern deep learning methods lack algorithmic transparency. While the heuristic optimization procedures for neural networks are demonstrably powerful, we don't understand how they work, and at present cannot guarantee a priori that they will work on new problems. Note, however, that humans exhibit none of these forms of transparency.

3.2. Post-hoc Interpretability

A distinct model-based notion of *interpretability*, post-hoc interpretations consist of explanations that need not elucidate the exact process by which models work. These interpretations include natural language explanations, visualizations of learned representations or models, and explanations by example (e.g. *this tumor is classified as malignant because to the model it looks a lot like these other tumors*). To the extent that we might consider humans to be interpretable, it is this sort of interpretability that applies. For all we know, the processes by which we humans make decisions and those by which we explain them may be distinct. One advantage of this concept of interpretability is that we can interpret opaque models after-the-fact, without sacrificing predictive performance.

Text Explanations: Humans often justify decisions verbally. Similarly we might imagine training one model to generate predictions and a separate model, such as a recurrent neural network language model, to generate an explanation. Already, the representations learned by neural network image classifiers have been co-opted to generate captions. These captions might be regarded as interpretations that accompany classifications. In a less anthropomorphic approach, McAuley & Leskovec (2013) use text to explain the decisions of a latent factor model by training topic models on product reviews using normalized latent factors as topic distributions. The predictions are then explained via the top words in the topics corresponding to the latent factors signaling user-item compatibility. Note that this interpretation is not fully post-hoc because the topic model is incorporated into the training procedure. Also, note that the practice of interpreting topic models by presenting the top words is itself a post-hoc interpretation that has invited scrutiny (Chang et al., 2009).

Visualization: Another common approach to generating post-hoc interpretations is to render visualizations, hinting at what a model has learned. A popular approach for deep neural nets, saliency maps compute gradients of predictions with respect to inputs, rendering the gradient as an image (Simonyan et al., 2013). While this doesn't say precisely how a model works, it conveys which image regions the current classification depends upon most heavily. Another popular approach is to visualize high-dimensional distributed representations with t-SNE (Van der Maaten & Hinton, 2008), a technique that renders 2D visualizations in which nearby data points are likely to appear close together. Mordvintsev et al. (2015) attempt to explain what an image classification network has learned by altering the input through gradient descent to enhance the activations of certain nodes selected from the hidden layers. An inspection of the perturbed inputs can give clues to what the model has learned.

Explanation by Example: One post-hoc mechanism for explaining the decisions of a model might be to report (in addition to predictions) which other examples the model considers to be most similar, a method suggested by Caruana et al. (1999). After training a deep neural network or latent variable model for a discriminative task, we then have access not only to predictions but also to the learned representations. Then, for any example, in addition to generating a prediction, we can use the activations of the hidden layers to identify the k-nearest neighbors based on the proximity in the space learned by the model. This sort of explanation by example has precedent in how humans sometimes justify actions by analogy. Doctors for example, often refer back to case studies.

4. Discussion

The concept of interpretability appears simultaneously important and slippery. Earlier, we analyzed both the motivations for interpretability and the attempts of the research community to confer it. In this discussion, we consider the implications of our analysis and offer several takeaways to the reader.

Linear models are not strictly more interpretable than deep neural networks: Despite this claims' enduring popularity, its truth content varies depending on what notion of interpretability we employ. With respect to *algorithmic transparency*, this claim seems uncontroversial, but given high dimensional or heavily engineered features, linear models lose *simulatability* or *decomposability* respectively. Considering *simulatability*, large-scale linear models may be just as opaque as deep neural networks. However, for post-hoc interpretability, deep neural networks exhibit a clear advantage, learning rich representations that could be visualized, verbalized, or used for clustering. Considering the motivations for interpretability, linear models appear to have a better track record for studying the natural world and for identifying weaknesses in training data, but we do not know of a theoretical reason why this must be so. Conceivably, post-hoc interpretations could prove useful in both scenarios.

Claims about interpretability must be qualified: As demonstrated above, the term does not reference a monolithic concept. To be meaningful, any assertion regarding interpretability should fix a specific definition. If the model satisfies a form of transparency, this can be shown directly. For post-hoc interpretability, papers ought to fix clear a motivation and demonstrate evidence that the offered form of interpretation satisfies it.

In some cases, transparency may be at odds with the broader objectives of AI: Arguments against *black-box* algorithms might preclude any model that could match or surpass our abilities on complex tasks. As a concrete example, the short-term goal of building trust with doctors by developing transparent models might clash with the longer-term goal of improving health care. We should be careful when giving up predictive power, that the desire for transparency is justified and isn't simply a concession to institutional biases against new methods.

Post-hoc interpretations can potentially mislead: We also caution against blindly embracing post-hoc notions of interpretability, especially when optimized to placate subjective demands. In such cases, one might inadvertently optimize an algorithm to present misleading but plausible explanations. As humans, we are known to engage in this behavior, as evidenced in hiring practices and college admissions. Several journalists and social scientists have demonstrated that acceptance decisions attributed to virtues like *leadership* or *originality* often disguise racial or gender discrimination (Mounk, 2014). In the rush to gain acceptance for machine learning and to emulate human intelligence, we should be careful not to reproduce pathological behavior at scale.

Future Work: We see several promising directions for future work. First, for some problems, the discrepancy between real-life and machine learning objectives could be mitigated by developing richer objectives and performance metrics. Exemplars of this direction include research on sparsity-inducing regularizers and cost-sensitive learning. Second, we can expand this analysis to other ML paradigms such as reinforcement learning. Reinforcement learners can address some (but not all) motivations for interpretability by directly modeling interaction between models and environments. However, this capability may come at the cost of allowing models to experiment in the world, incurring real consequences.

5. Acknowledgements

Thanks to Charles Elkan, Julian McAuley, Maggie Makar, David Kale, Been Kim, Lihong Li, Rich Caruana, Sepp Hochreiter, Daniel Fried, Jack Berkowitz, and others for helpful conversations and critical feedback.

References

Caruana, Rich, Kangarloo, Hooshang, Dionisio, JD, Sinha, Usha, and Johnson, David. Case-based explanation of non-case-based learning methods. In *Proceedings of the AMIA Symposium*, pp. 212. American Medical Informatics Association, 1999.

Caruana, Rich, Lou, Yin, Gehrke, Johannes, Koch, Paul, Sturm, Marc, and Elhadad, Noémie. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. 2015.

Chang, Jonathan, Gerrish, Sean, Wang, Chong, Boyd-Graber, Jordan L, and Blei, David M. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pp. 288–296, 2009.

Dragan, Anca D, Lee, Kenton CT, and Srinivasa, Siddhartha S. Legibility and predictability of robot motion. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*, pp. 301–308. IEEE, 2013.

Huysmans, Johan, Dejaeger, Karel, Mues, Christophe, Vanthienen, Jan, and Baesens, Bart. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1):141–154, 2011.

Kim, Been. *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology, 2015.

Kim, Been, Glassman, Elena, Johnson, Brittney, and Shah, Julie. ibcm: Interactive bayesian case model empowering humans via intuitive interaction. 2015.

Lou, Yin, Caruana, Rich, and Gehrke, Johannes. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 150–158. ACM, 2012.

Lou, Yin, Caruana, Rich, Gehrke, Johannes, and Hooker, Giles. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 623–631. ACM, 2013.

McAuley, Julian and Leskovec, Jure. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pp. 165–172. ACM, 2013.

Mordvintsev, Alexander, Olah, Christopher, and Tyka, Mike. Inceptionism: Going deeper into neural networks. *Google Research Blog. Retrieved June*, 20, 2015.

Mounk, Yascha. Is Harvard unfair to asian-americans?, 2014. URL `http://www.nytimes.com/2014/11/25/opinion/is-harvard-unfair-to-asian-americans.html?_r=0`.

Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. " why should i trust you?": Explaining the predictions of any classifier. *arXiv preprint arXiv:1602.04938*, 2016.

Ridgeway, Greg, Madigan, David, Richardson, Thomas, and O'Kane, John. Interpretable boosted naïve bayes classification. In *KDD*, pp. 101–104, 1998.

Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Szegedy, Christian, Zaremba, Wojciech, Sutskever, Ilya, Bruna, Joan, Erhan, Dumitru, Goodfellow, Ian, and Fergus, Rob. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Van der Maaten, Laurens and Hinton, Geoffrey. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

Wang, Hui-Xin, Fratiglioni, Laura, Frisoni, Giovanni B, Viitanen, Matti, and Winblad, Bengt. Smoking and the occurence of alzheimer's disease: Cross-sectional and longitudinal data in a population-based study. *American journal of epidemiology*, 149(7):640–644, 1999.