



# The Mythos of Model Interpretability

**IN MACHINE LEARNING, THE  
CONCEPT OF INTERPRETABILITY IS  
BOTH IMPORTANT AND SLIPPERY.**

ZACHARY C. LIPTON

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? Models should be not only good, but also interpretable, yet the task of interpretation appears underspecified. The academic literature has provided diverse and sometimes non-overlapping motivations for interpretability and has offered myriad techniques for rendering interpretable models. Despite this ambiguity, many authors proclaim their models to be interpretable axiomatically, absent further argument. Problematically, it is not clear what common properties unite these techniques.

This article seeks to refine the discourse on interpretability. First it examines the objectives of previous

papers addressing interpretability, finding them to be diverse and occasionally discordant. Then, it explores model properties and techniques thought to confer interpretability, identifying transparency to humans and post hoc explanations as competing concepts. Throughout, the feasibility and desirability of different notions of interpretability are discussed. The article questions the oft-made assertions that linear models are interpretable and that deep neural networks are not.

## INTRODUCTION

Until recently, humans had a monopoly on agency in society. If you applied for a job, loan, or bail, a human decided your fate. If you went to the hospital, a human would attempt to categorize your malady and recommend treatment. For consequential decisions such as these, you might demand an explanation from the decision-making agent.

If your loan application is denied, for example, you might want to understand the agent's reasoning in a bid to strengthen your next application. If the decision was based on a flawed premise, you might contest this premise in the hope of overturning the decision. In the hospital, a doctor's explanation might educate you about your condition.

In societal contexts, the *reasons* for a decision often matter. For example, intentionally causing death (murder) vs. unintentionally (manslaughter) are distinct crimes. Similarly, a hiring decision being based (directly or indirectly) on a protected characteristic such as race has a bearing on its legality. However, today's predictive models are not capable of reasoning at all.

**I**n the academic literature, few authors articulate precisely what interpretability means or precisely how their proposed solution is useful.

Over the past 20 years, rapid progress in ML (machine learning) has led to the deployment of automatic decision processes. Most ML-based decision making in practical use works in the following way: the ML algorithm is trained to take some input and predict the corresponding output. For example, given a set of attributes characterizing a financial transaction, an ML algorithm can predict the long-term return on investment. Given images from a CT scan, the algorithm can assign a probability that the scan depicts a cancerous tumor. The ML algorithm takes in a large corpus of (input, output) pairs, and outputs a *model* that can predict the output corresponding to a previously unseen input. Formally, researchers call this problem setting *supervised learning*. Then, to automate decisions fully, one feeds the model's output into some decision rule. For example, spam filters programmatically discard emails predicted to be spam with a level of confidence exceeding some threshold.

Thus, ML-based systems do not know why a given input should receive some label, only that certain inputs are correlated with that label. For example, shown a dataset in which the only orange objects are basketballs, an image classifier might learn to classify all orange objects as basketballs.

This model would achieve high accuracy even on held out images, despite failing to grasp the difference that actually makes a difference.

As ML penetrates critical areas such as medicine, the criminal justice system, and financial markets, the inability of humans to understand these models seems problematic. Some suggest *model interpretability* as a remedy, but in the

academic literature, few authors articulate precisely what interpretability means or precisely how their proposed solution is useful.

Despite the lack of a definition, a growing body of literature proposes purportedly interpretable algorithms. From this, you might conclude that either: (1) the definition of *interpretability* is universally agreed upon, but no one has bothered to set it in writing; or (2) the term *interpretability* is ill-defined, and, thus, claims regarding interpretability of various models exhibit a quasi-scientific character. An investigation of the literature suggests the latter. Both the objectives and methods put forth in the literature investigating interpretability are diverse, suggesting that interpretability is not a monolithic concept but several distinct ideas that must be disentangled before any progress can be made.

This article focuses on supervised learning rather than other ML paradigms such as reinforcement learning and interactive learning. This scope derives from the current primacy of supervised learning in real-world applications and an interest in the common claim that linear models are interpretable while deep neural networks are not.<sup>15</sup> To gain conceptual clarity, consider these refining questions: What is interpretability? Why is it important?

Let's address the second question first (expanded in the section, "Desiderata of Interpretability Research"). Many authors have proposed interpretability as a means to engender trust.<sup>9,24</sup> This leads to a similarly vexing epistemological question: What is trust? Does it refer to faith that a model will perform well? Or does interpretability simply mean a low-level mechanistic

understanding of models? Is trust defined subjectively?

Other authors suggest that an interpretable model is desirable because it might help uncover causal structure in observational data.<sup>1</sup> The legal notion of a *right to explanation* offers yet another lens on interpretability. One goal of interpretability might simply be to get more useful information from the model.

While the discussed desiderata, or objectives of interpretability, are diverse, they typically speak to situations where standard ML problem formulations, e.g. maximizing accuracy on a set of hold-out data for which the training data is perfectly representative, are imperfectly matched to the complex real-life tasks they are meant to solve. Consider medical research with longitudinal data. The real goal may be to discover potentially causal associations that can guide interventions, as with smoking and cancer.<sup>29</sup> The optimization objective for most supervised learning models, however, is simply to minimize error, a feat that might be achieved in a purely correlative fashion.

Another example of such a mismatch is that available training data imperfectly represents the likely deployment environment. Real environments often have changing dynamics. Imagine training a product recommender for an online store, where new products are periodically introduced, and customer preferences can change over time. In more extreme cases, actions from an ML-based system may alter the environment, invalidating future predictions.

After addressing the desiderata of interpretability, this article considers which properties of models might

render them interpretable (expanded in the section, “Properties of Interpretable Models”). Some papers equate interpretability with understandability or intelligibility,<sup>16</sup> (i.e., you 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? You might look to the algorithm itself: Will it converge? Does it produce a unique solution? Or you might look to its parameters: Do you understand what each represents? Alternatively, you could consider the model’s complexity: Is it simple enough to be examined all at once by a human?

Other work has investigated so-called post hoc interpretations. These interpretations might explain predictions without elucidating the mechanisms by which models work. Examples include the verbal explanations produced by people or the saliency maps used to analyze deep neural networks. Thus, human decisions might admit post hoc interpretability despite the black-box nature of human brains, revealing a contradiction between two popular notions of interpretability.

## DESIDERATA OF INTERPRETABILITY RESEARCH

This section spells out the various desiderata of interpretability research. The demand for interpretability arises when a mismatch occurs between the formal objectives of supervised learning (test set predictive performance) and the real-world costs in a deployment setting.

Typically, evaluation metrics require only predictions and ground-truth labels. When stakeholders additionally

demand interpretability, you might infer the existence of objectives that cannot be captured in this fashion. Consider that most common evaluation metrics for supervised learning require only predictions, together with ground truth, to produce a score. So, the very desire for an interpretation suggests that sometimes predictions alone and metrics calculated on them do not suffice to characterize the model. You should then ask, What are these other objectives and under what circumstances are they sought?

Often, real-world objectives are difficult to encode as simple mathematical functions. Otherwise, they might just be incorporated into the objective function and the problem would be considered solved. For example, an algorithm for making hiring decisions should simultaneously optimize productivity, ethics, and legality. But how would you go about writing a function that measures ethics or legality? The problem can also arise when you desire robustness to changes in the dynamics between the training and deployment environments.

### Trust

Some authors suggest interpretability is a prerequisite for *trust*.<sup>9,23</sup> Again, 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 serves 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 the growing use of ML models to forecast crime rates for purposes of allocating police officers. The model may be trusted to make accurate predictions but not to account for racial biases in the training data for the model's own effect in perpetuating a cycle of incarceration by over-policing some neighborhoods.

Another sense in which an end user might be said to trust an ML model might be if they are comfortable with relinquishing control to it. Through this lens, you might care not only about *how often* a model is right, but also *for which examples* it is right. If the model tends to make mistakes on only those kinds of inputs where humans also make mistakes, and thus is typically accurate whenever humans are accurate, then you might trust the model owing to the absence of any expected cost of relinquishing control. If a model tends to make mistakes for inputs that humans classify accurately, however, then there may always be an advantage to maintaining human supervision of the algorithms.

## Causality

Although supervised learning models are only optimized directly to make associations, researchers often use them in the hope of inferring properties of the natural world. For example, a simple regression model might reveal a strong association between thalidomide use and birth defects, or between smoking and lung cancer.<sup>29</sup>

The associations learned by supervised learning



**W**hile the machine-learning objective might be to reduce error, the real-world purpose is to provide useful information.

algorithms are not guaranteed to reflect causal relationships. There could always be unobserved causes responsible for both associated variables. You might hope, however, that by interpreting supervised learning models, you could generate hypotheses that scientists could then test. Liu et al.,<sup>14</sup> for example, emphasize regression trees and Bayesian neural networks, suggesting that these models are interpretable and thus better able to provide clues about the causal relationships between physiologic signals and affective states. The task of inferring causal relationships from observational data has been extensively studied.<sup>22</sup> Causal inference methods, however, tend to rely on strong assumptions and are not widely used by practitioners, especially on large, complex data sets.

### Transferability

Typically, training and test data are chosen by randomly partitioning examples from the same distribution. A model's generalization error is then judged by the gap between its performance on training and test data. Humans exhibit a far richer capacity to generalize, however, transferring learned skills to unfamiliar situations. ML algorithms are already used in these situations, such as when the environment is nonstationary. Models are also deployed in settings where their use might alter the environment, invalidating their future predictions. Along these lines, Caruana et al.<sup>3</sup> describe a model trained to predict probability of death from pneumonia that assigned less risk to patients if they also had asthma. Presumably, asthma was predictive of a lower risk of death because of the more aggressive treatment these patients

received. If the model were deployed to aid in triage, these patients might then receive less aggressive treatment, invalidating the model.

Even worse, there are situations, such as machine learning for security, where the environment might be actively adversarial. Consider the recently discovered susceptibility of CNNs (convolutional neural networks). The CNNs were made to misclassify images that were imperceptibly (to a human) perturbed.<sup>26</sup> Of course, this isn't overfitting in the classical sense. The models both achieve strong results on training data and generalize well when used to classify held out test data. The crucial distinction is that these images have been altered in ways that, while subtle to human observers, the models never encountered during training. However, these are mistakes a human wouldn't make, and it would be preferable that models not make these mistakes, either. Already, supervised learning models are regularly subject to such adversarial manipulation. Consider the models used to generate credit ratings; higher scores should signify a higher probability that an individual repays a loan. According to its own technical report, FICO trains credit models using logistic regression,<sup>6</sup> specifically citing interpretability as a motivation for the choice of model. Features include dummy variables representing binned values for average age of accounts, debt ratio, the number of late payments, and the number of accounts in good standing.

Several of these factors can be manipulated at will by credit-seekers. For example, one's debt ratio can be improved simply by requesting periodic increases to credit lines while keeping spending patterns constant.

Similarly, the total number of accounts can be increased by simply applying for new accounts when the probability of acceptance is reasonably high. Indeed, FICO and Experian both acknowledge that credit ratings can be manipulated, even suggesting guides for improving one's credit rating. These rating-improvement strategies may fundamentally change one's underlying ability to pay a debt. The fact that individuals actively and successfully game the rating system may invalidate its predictive power.

### Informativeness

Sometimes, decision theory is applied to the outputs of supervised models to take actions in the real world. In another common use paradigm, however, the supervised model is used instead to provide information to human decision-makers, a setting considered by Kim et al.<sup>11</sup> and Huysmans et al.<sup>8</sup> While the machine-learning objective might be to reduce error, the real-world purpose is to provide useful information. The most obvious way that a model conveys information is via its outputs, but it may be possible via some procedure to convey additional information to the human decision-maker.

An interpretation may prove informative even without shedding light on a 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, a supervised learning model is trained when the real task more closely resembles unsupervised learning. The real goal might be to explore the underlying structure of the data, and the labeling objective serves only as weak supervision.

### Fair and ethical decision making

At present, politicians, journalists, and researchers have expressed concern that interpretations must be produced for assessing whether decisions produced automatically by algorithms conform to ethical standards.<sup>7</sup> Recidivism predictions are already used to determine whom to release and whom to detain, raising ethical concerns. How can you be sure that predictions do not discriminate on the basis of race? Conventional evaluation metrics such as accuracy or AUC (area under the curve) offer little assurance that ML-based decisions will behave acceptably. Thus, demands for fairness often lead to demands for interpretable models.

#### THE TRANSPARENCY NOTION OF INTERPRETABILITY

Let's now consider the techniques and model properties that are proposed to confer interpretability. These fall broadly into two categories. The first relates to transparency (i.e., how does the model work?). The second consists of post hoc explanations (i.e., what else can the model tell me?)

Informally, transparency is the opposite of opacity or “black-boxness.” It connotes some sense of understanding the mechanism by which the model works. Transparency is considered here at the level of the entire model (*simulatability*), at the level of individual components such as parameters (*decomposability*), and at the level of the training algorithm (*algorithmic transparency*).

#### *Simulatability*

In the strictest sense, a model might be called transparent

if a person can contemplate the entire model at once. This definition suggests that an interpretable model is a simple model. For example, for a model to be fully understood, a human should be able to take the input data together with the parameters of the model and in reasonable time step through every calculation required to produce a prediction. This accords with the common claim that sparse linear models, as produced by lasso regression,<sup>27</sup> are more interpretable than dense linear models learned on the same inputs. Ribeiro et al.<sup>23</sup> also adopt this notion of interpretability, suggesting that an interpretable model is one that “can be readily presented to the user with visual or textual artifacts.”

The tradeoffs between model size and computation to apply a single prediction varies across models. For example, in some models, such as decision trees, the size of the model (total number of nodes) may grow quite large compared to the time required to perform inference (length of pass from root to leaf). This suggests that simulatability may admit two subtypes: one based on the size of the model and another based on the computation required to perform inference.

Fixing a notion of simulatability, the quantity denoted by *reasonable* is subjective. Clearly, however, given the limited capacity of human cognition, this ambiguity might span only several orders of magnitude. In this light, neither linear models, rule-based systems, nor decision trees are intrinsically interpretable. Sufficiently high-dimensional models, unwieldy rule lists, and deep decision trees could all be considered less transparent than comparatively compact neural networks.

**T**he weights of a linear model might seem intuitive, but they can be fragile with respect to feature selection and preprocessing.

### *Decomposability*

A second notion of transparency might be that each part of the model—input, parameter, and calculation—admits an intuitive explanation. This accords with the property of intelligibility as described by Lou et al.<sup>15</sup> 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.

Note that 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, it shouldn't be accepted blindly. The weights of a linear model might seem intuitive, but they can be fragile with respect to feature selection and preprocessing. For example, the coefficient corresponding to the association between flu risk and vaccination might be positive or negative, depending on whether the feature set includes indicators of old age, infancy, or immunodeficiency.

### *Algorithmic transparency*

A final notion of transparency might apply at the level of the learning algorithm itself. In the case of linear models, you may understand the shape of the error surface. You can prove that training will converge to a unique solution, even for previously unseen data sets. This might provide some confidence that the model will behave in an online setting requiring programmatic retraining on previously

unseen data. On the other hand, modern deep learning methods lack this sort of 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.

### Post hoc interpretability

Post hoc interpretability represents a distinct approach to extracting information from learned models. While post hoc interpretations often do not elucidate precisely how a model works, they may nonetheless confer useful information for practitioners and end users of machine learning. Some common approaches to post hoc 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, this is the sort of interpretability that applies. For all we know, the processes by which humans make decisions and those by which they explain them may be distinct. One advantage of this concept of interpretability is that opaque models can be interpreted after the fact, without sacrificing predictive performance.

### *Text explanations*

Humans often justify decisions verbally. Similarly, one model might be trained to generate predictions, and a separate model, such as a recurrent neural network

language model, to generate an explanation. Such an approach is taken in a line of work by Krening et al.<sup>12</sup> They propose a system in which one model (a reinforcement learner) chooses actions to optimize cumulative discounted return. They train another model to map a model's state representation onto verbal explanations of strategy. These explanations are trained to maximize the likelihood of previously observed ground-truth explanations from human players and may not faithfully describe the agent's decisions, however plausible they appear. A connection exists between this approach and recent work on neural image captioning in which the representations learned by a discriminative CNN (trained for image classification) are coopted by a second model to generate captions. These captions might be regarded as interpretations that accompany classifications.

In work on recommender systems, McAuley and Leskovec<sup>18</sup> use text to explain the decisions of a latent factor model. Their method consists of simultaneously training a latent factor model for rating prediction and a topic model for product reviews. During training they alternate between decreasing the squared error on rating prediction and increasing the likelihood of review text. The models are connected because they use normalized latent factors as topic distributions. In other words, latent factors are regularized such that they are also good at explaining the topic distributions in review text. The authors then explain user-item compatibility by examining the top words in the topics corresponding to matching components of their latent factors. Note that the practice of interpreting topic models by presenting the top words is



itself a post hoc interpretation technique that has invited scrutiny.<sup>4</sup> Moreover note that here we have only spoken to the form factor of an explanation (that it consists of natural language), but not what precisely constitutes correctness. So far, the literature has dodged the issue of correctness, sometimes punting the issue by embracing a subjective view of the problem and asking people what they prefer.

### *Visualization*

Another common approach to generating post hoc interpretations is to render visualizations in the hope of determining qualitatively what a model has learned. One popular method is to visualize high-dimensional distributed representations with t-SNE (t-distributed stochastic neighbor embedding),<sup>28</sup> a technique that renders 2D visualizations in which nearby data points are likely to appear close together.

Mordvintsev et al.<sup>20</sup> 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. Likely because the model was trained on a large corpus of animal images, they observed that enhancing some nodes caused certain dog faces to appear throughout the input image.

In the computer vision community, similar approaches have been explored to investigate what information is retained at various layers of a neural network. Mahendran and Vedaldi<sup>17</sup> pass an image through a discriminative CNN to generate a representation. They then demonstrate that

the original image can be recovered with high fidelity even from reasonably high-level representations (level 6 of an AlexNet) by performing gradient descent on randomly initialized pixels. As before with text, discussions of visualization focus on form factor and appeal, but we still lack a rigorous standard of correctness.

### *Local explanations*

While it may be difficult to describe succinctly the full mapping learned by a neural network, some of the literature focuses instead on explaining what a neural network depends on locally. One popular approach for deep neural nets is to compute a saliency map. Typically, they take the gradient of the output corresponding to the correct class with respect to a given input vector. For images, this gradient can be applied as a mask, highlighting regions of the input that, if changed, would most influence the output.<sup>25,30</sup>

Note that these explanations of what a model is focusing on may be misleading. The saliency map is a local explanation only. Once you move a single pixel, you may get a very different saliency map. This contrasts with linear models, which model global relationships between inputs and outputs.

Another attempt at local explanations is made by Ribeiro et al.<sup>23</sup> In this work, the authors explain the decisions of any model in a local region near a particular point by learning a separate sparse linear model to explain the decisions of the first. Strangely, although the method's appeal over saliency maps owes to its ability to provide explanations for non-differentiable models, it is more

often used when the model subject to interpretation is in fact differentiable. In this case, what is provided, besides a noisy estimate of the gradient, remains unclear. In this paper, the explanation is offered in terms of a set of superpixels. Whether or not this is more informative than a plain gradient may depend strongly on how one chooses the superpixels. Moreover, absent a rigorously defined objective, who is to say which hyper-parameters are correct?

### *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 are most similar with respect to the model, a method suggested by Caruana et al.<sup>2</sup> Training a deep neural network or latent variable model for a discriminative task provides access to not only predictions but also the learned representations. Then, for any example, in addition to generating a prediction, you 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. For example, doctors often refer to case studies to support a planned treatment protocol.

In the neural network literature, Mikolov et al.<sup>19</sup> use such an approach to examine the learned representations of words after training the word2vec model. While their model is trained for discriminative skip-gram prediction, to examine which relationships the model has learned they enumerate nearest neighbors of words based on

**L**inear models lose simulatability or decomposability, respectively.

distances calculated in the latent space. Kim et al.<sup>10</sup> and Doshi-Velez et al.<sup>5</sup> have done related work in Bayesian methods, investigating case-based reasoning approaches for interpreting generative models.

## DISCUSSION

The concept of interpretability appears simultaneously important and slippery. Earlier, this article analyzed both the motivations for interpretability and some attempts by the research community to confer it. Now let's consider the implications of this analysis and offer several takeaways.

- ➔ Linear models are not strictly more interpretable than deep neural networks. Despite this claim's enduring popularity, its truth value depends on which notion of interpretability is employed. With respect to algorithmic transparency, this claim seems uncontroversial, but given high-dimensional or heavily engineered features, linear models lose simulatability or decomposability, respectively.

When choosing between linear and deep models, you must often make a tradeoff between algorithmic transparency and decomposability. This is because deep neural networks tend to operate on raw or lightly processed features. So, if nothing else, the features are intuitively meaningful, and post hoc reasoning is sensible. To get comparable performance, however, linear models often must operate on heavily hand-engineered features. Lipton et al.<sup>13</sup> demonstrate such a case where linear models can approach the performance of recurrent neural networks (RNNs) only at the cost of decomposability.

For some kinds of post hoc interpretation, deep

neural networks exhibit a clear advantage. They learn rich representations that can be visualized, verbalized, or used for clustering. Considering the desiderata for interpretability, linear models appear to have a better track record for studying the natural world, but there seems to be no theoretical reason why this must be so. Conceivably, post hoc interpretations could prove useful in similar scenarios.

- ➔ Claims about interpretability must be qualified.

As demonstrated here, the term interpretability 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, work in this field should fix a clear objective and demonstrate evidence that the offered form of interpretation achieves it.

- ➔ In some cases, transparency may be at odds with the broader objectives of AI (artificial intelligence). Some arguments against black-box algorithms appear to preclude any model that could match or surpass human 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. Be careful when giving up predictive power that the desire for transparency is justified and not simply a concession to institutional biases against new methods.

- ➔ Post hoc interpretations can potentially mislead. Beware of blindly embracing post hoc notions of interpretability, especially when optimized to placate

subjective demands. In such cases, one might—deliberately or not—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 such as leadership or originality often disguise racial or gender discrimination.<sup>21</sup> In the

rush to gain acceptance for machine learning and to emulate human intelligence, we should all be careful not to reproduce pathological behavior at scale.

#### FUTURE WORK

There are 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 loss functions and performance metrics. Exemplars of this direction include research on sparsity-inducing regularizers and cost-sensitive learning. Second, this analysis can be expanded to other ML paradigms such as reinforcement learning.

Reinforcement learners can address some (but not all) of the objectives of interpretability research by directly

#### Related articles

➡ Accountability in Algorithmic Decision-making

Nicholas Diakopoulos

A view from computational journalism

<https://queue.acm.org/detail.cfm?id=2886105>

➡ Black Box Debugging

James A. Whittaker, Herbert H. Thompson

It's all about what takes place at the boundary of an application.

<https://queue.acm.org/detail.cfm?id=966807>

➡ Hazy: Making it Easier to Build and

Maintain Big-data Analytics

Arun Kumar, Feng Niu, and Christopher Ré

Racing to unleash the full potential of big data with the latest statistical and machine-learning techniques.

<https://queue.acm.org/detail.cfm?id=2431055>

modeling interaction between models and environments. This capability, however, may come at the cost of allowing models to experiment in the world, incurring real consequences.

Notably, reinforcement learners are able to learn causal relationships between their actions and real-world impacts. Like supervised learning, however, reinforcement learning relies on a well-defined scalar objective. For problems such as fairness, where we struggle to verbalize precise definitions of success, a shift of the ML paradigm is unlikely to eliminate the problems we face.

## References

1. Athey, S., Imbens, G. W. 2015 Machine-learning methods <https://arxiv.org/abs/1504.01132v1> (see also ref. 7).
2. Caruana, R., Kangarloo, H., Dionisio, J. D., Sinha, U., Johnson, D. 1999. Case-based explanation of non-case-based learning methods. In *Proceedings of the American Medical Informatics Association (AMIA) Symposium*: 212-215.
3. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21<sup>st</sup> Annual SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721-1730.
4. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., Blei, D. M. 2009. Reading tea leaves: how humans interpret topic models. In *Proceedings of the 22<sup>nd</sup> International Conference on Neural Information Processing Systems (NIPS)*, 288-296.

5. Doshi-Velez, F., Wallace, B., Adams, R. 2015. Graph-sparse LDA: a topic model with structured sparsity. In *Proceedings of the 29<sup>th</sup> Association for the Advancement of Artificial Intelligence (AAAI) Conference*, 2575-2581.
6. FICO (Fair Isaac Corporation). 2011. Introduction to model builder scorecard; <http://www.fico.com/en/latest-thinking/white-papers/introduction-to-model-builder-scorecard>.
7. Goodman, B., Flaxman, S. 2016. European Union regulations on algorithmic decision-making and a “right to explanation.” <https://arxiv.org/abs/1606.08813v3>.
8. Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., Baesens, B. 2011. An empirical evaluation of the comprehensibility of decision table, tree- and rule-based predictive models. *Journal of Decision Support Systems* 51(1), 141-154.
9. Kim, B. 2015. Interactive and interpretable machine-learning models for human-machine collaboration. Ph.D. thesis. Massachusetts Institute of Technology.
10. Kim, B., Rudin, C., Shah, J. A. 2014. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Proceedings of the 27<sup>th</sup> International Conference on Neural Information Processing Systems (NIPS)*, volume 2, 1952-1960.
11. Kim, B., Glassman, E., Johnson, B., Shah, J. 2015. iBCM: Interactive Bayesian case model empowering humans via intuitive interaction. Massachusetts Institute of Technology, Cambridge, MA.
12. Krening, S., Harrison, B., Feigh, K., Isbell, C., Riedl, M., Thomaz, A. 2017. Learning from explanations using sentiment and advice in RL. *IEEE Transactions on*



- Cognitive and Developmental Systems* 9(1), 41-55.
13. Lipton, Z. C., Kale, D. C., Wetzel, R. 2016. Modeling missing data in clinical time series with RNNs. In *Proceedings of Machine Learning for Healthcare*.
  14. Liu, C., Rani, P., Sarkar, N. 2006. An empirical study of machine-learning techniques for affect recognition in human-robot interaction. *Pattern Analysis and Applications* 9(1): 58-69.
  15. Lou, Y., Caruana, R., Gehrke, J. 2012. Intelligible models for classification and regression. In *Proceedings of the 18<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 150-158.
  16. Lou, Y., Caruana, R., Gehrke, J., Hooker, G. 2013. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19<sup>th</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 623-631.
  17. Mahendran, A., Vedaldi, A. 2015. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9.
  18. McAuley, J., Leskovec, J. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7<sup>th</sup> ACM Conference on Recommender Systems*, 165-172.
  19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26<sup>th</sup> International Conference on Neural Information Processing Systems (NIPS)*, volume 2, 3111–3119.
  20. Mordvintsev, A., Olah, C., Tyka, M. 2015. Inceptionism:

- going deeper into neural networks. Google AI Blog; <https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html>.
21. Mounk, Y. 2014. Is Harvard unfair to Asian-Americans? *New York Times* (Nov. 24); <http://www.nytimes.com/2014/11/25/opinion/is-harvard-unfair-to-asian-americans.html>.
  22. Pearl, J. 2009. *Causality*. Cambridge University Press.
  23. Ribeiro, M. T., Singh, S., Guestrin, C. 2016. “Why should I trust you?": explaining the predictions of any classifier. In *Proceedings of the 22<sup>nd</sup> SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135-1144.
  24. Ridgeway, G., Madigan, D., Richardson, T., O’Kane, J. 1998. Interpretable boosted naïve Bayes classification. In *Proceedings of the 4<sup>th</sup> International Conference on Knowledge Discovery and Data Mining*: 101-104.
  25. Simonyan, K., Vedaldi, A., Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. <https://arxiv.org/abs/1312.6034> [see notes to refs 1, 7].
  26. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R. 2013. Intriguing properties of neural networks. <https://arxiv.org/abs/1312.6199> [see refs 1, 7, 25].
  27. Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58(1), 267-288.
  28. Van der Maaten, L., Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 2579-2605.

29. Wang, H.-X., Fratiglioni, L., Frisoni, G. B., Viitanen, M., Winblad, B. 1999. Smoking and the occurrence of Alzheimer's disease: cross-sectional and longitudinal data in a population-based study. *American Journal of Epidemiology* 149(7), 640-644.
30. Wang, Z., Freitas, N., Lanctot, M. 2016. Dueling network architectures for deep reinforcement learning. *Proceedings of the 33<sup>rd</sup> International Conference on Machine Learning* 48, 1995-2003.

**Zachary Chase Lipton** is an assistant professor at Carnegie Mellon University. His research spans both core machine-learning methods and their social impact, concentrating on deep learning for time series data and sequential decision making. This work addresses diverse application areas, including medical diagnosis, dialogue systems, and product recommendation. He is the founding editor of the *Approximately Correct* blog and the lead author of *Deep Learning – The Straight Dope*, an open-source interactive book teaching deep learning through Jupyter Notebook. Find him on Twitter (@zacharylipton) or GitHub (@zackchase).

