phg foundation
making science work for health

Black box medicine and transparency

Interpretable machine learning

A PHG Foundation report for the Wellcome Trust

UNIVERSITY OF CAMBRIDGE

Authors

Johan Ordish and Alison Hall

Acknowledgements

The *Black Box Medicine and Transparency* project was funded by the Wellcome Trust as a part of their 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z]. We thank the Wellcome Trust for their support.

The series of reports is informed and underpinned by a series of roundtables and interviews. These roundtables and interviews are detailed in the Report of Roundtables and Interviews. Further, highlights from both are seeded throughout all reports, being found in 'A Salient Feature' boxes.

Disclaimer

URLs in this report were correct as of February 2020

This report is available from www.phgfoundation.org

Published by PHG Foundation 2 Worts Causeway, Cambridge, CB1 8RN, UK

+44 (0)1223 761900

February 2020

Correspondence to: intelligence@phgfoundation.org

How to reference this report:

Ordish J, Hall A. *Black Box Medicine and Transparency: Interpretable Machine Learning*. PHG Foundation. 2020.

Contents

1. Interpretable machine learning

Machine learning is often thought to represent a black box - the exact functioning of models being either hidden from or uninterpretable to observers. This report unpacks the black box concept and the related concept of human interpretability. It provides a general summary of what interpretability is with respect to machine learning, the methods available for rendering machine learning interpretable, and how to evaluate the interpretability of machine learning.

To recapitulate the Machine Learning Landscape report, we use the following terms in the following ways:

> *Machine learning* is a programming paradigm that differs from classical programming in that machine learning systems are trained rather than explicitly programmed.[1]

> *Features* 'are the input variables to a machine learning model. For example, when developing a model predicting stroke risk, a feature would be a patient's height or weight. Features can be processed before they are entered into a model, such as combining height and weight into a body mass index. For an image, a feature may be some component of the image, such as an eye or a nose, when developing a facial recognition machine learning system.'[2]

> *Machine learning algorithm* dictates how the model is trained, how the features are structured with respect to the task. For instance, convolutional neural networks in the context of image classification dictate how the model learns edges and distinguishes between classes of images.[3]

> *Machine learning model* is 'produced as the output of a machine learning algorithm applied to training data.'[4] Sometimes also called 'the trained model,' the model here has been trained according to the machine learning algorithm on a training set of data.

2. Black box models

Black box models are models 'whose internal workings are either unknown to the observer or known but uninterpretable to humans.'[5] In order to understand the scope of this definition, there is a need to distinguish between the following examples.

First, a model and its internal working may be perfectly interpretable, comprehensible, or explainable yet simply unavailable to an observer. Perhaps details about the model - the algorithm that trained the model, the data used to train as well as test it, and any graphical representation of the model - are restricted. In Burrell's (2016) terms, this is opacity that stems from intentional corporate or state secrecy.[6] That is, in principle the model is comprehensible but in practice, the information necessary to interpret the model's function is unavailable to the observer. Accordingly, a model may be a black box, not because of any intrinsic complexity, but because factors like trade secrets have led to restricted access to information that would otherwise allow an observer to interpret the model.

A Salient Feature | Interviews

Multiple interviewees emphasised that a model can be a black box due to its intrinsic properties or because restrictions upon information required to interpret the model are in place.

Second, an observer may have access to a model's internal workings but out-of-the-box the model may be uninterpretable, incomprehensible, or unexplainable. In Burrell's (2016) typology, this is opacity that arises from the characteristics of machine learning algorithms.[7] Perhaps the observer is told the algorithm used to train the model, has access to training and test data, but is unable to understand the relationship between the model's inputs and outputs. In this sense, the model is a black box due to some characteristics of the model or the algorithm that trained the model.

It is important to emphasise that even where the model is out-of-the-box interpretable to an observer, it may be uninterpretable to another observer given the same information. Indeed, in Burrell's typology, opacity may arise because of technical illiteracy.[8] In this regard, interpretability is also partially contingent upon the observer's capacity to interpret any given machine learning model.[9] Interpretability beyond a back-of-the-envelope estimation likely requires different approaches for different users in different contexts.[10] That said, some types of model may be more comprehensible than others.

Experiments comparing the subjective comprehensibility of types of models to users found general agreement that decision trees, classification rules, or decision tables tend to be comprehensible.[11] However, these experiments relied upon intuitive judgments of users, reporting rather than testing the perceptions of users. Nevertheless, it is important to note that assessing the comprehensibility of any given machine learning model is likely more complicated than merely assessing its complexity. That is, complexity is a poor measure of comprehensibility.[12] Accordingly, we provide a fuller description of comprehensibility and associated terms below.

A Salient Feature | Roundtable 3

Some participants were uncomfortable with the term 'black box,' these participants noting that all models are in principle interpretable so long as the observer has sufficient expertise and information available to them.

Section 2 key messages:

- **The term 'black box' includes opacity due to intentional restriction of information, and opacity due to the uninterpretability of the model, even when given unrestricted access to the model.**
- **Given the same information, a model may be interpretable to one observer and uninterpretable to another; technical literacy and understanding underpins the ability to find models interpretable.**
- **Interpretability of models is not just a matter of the complexity or simplicity of the model. We address the elements of interpretability in Section 4 below.**

3. Black box medicine

Price defines *black box medicine* as 'the use of opaque computational models to make decisions related to healthcare.'[13] Price distinguishes between two (rough) forms of black-box medicine:

I. *Literal black box medicine* where 'relationships are totally hidden, even though the machine learning process is known.'[14] That is, the mechanism underpinning the model is opaque to everyone, including the original programmer.
II. *Practical black box medicine* where the 'machine learning algorithm can examine data, determine a relationship, and state it, but the underlying biological relationship is too complex to be amenable to scientific understanding or clinical trials.'[15]

In Price's terms, opacity derives from the model (in Price's terms 'black box algorithm') itself or the underlying complexity of the subject matter (or possibly both). It is unclear how these definitions fit with the methods we discuss in Section 4(c)(iv) below, under the umbrella of *explainable artificial intelligence* (XAI). For example, in regards to literal black box medicine, it is uncertain how hidden the relationships in the machine learning process would be when a black box predictor like LIME is applied.[16] Further, in regards to practical black box medicine, it is also unclear how efforts to reduce dimensionality to produce a simplified explanation or efforts to 'algorithmise' causal inference might erode this definition.[17] We address such complications later in this report.

What is clear is that the concept of black box medicine reflects a broader concern about the use of opaque computational models in healthcare. Indeed, beyond securing interpretability for the purposes of debugging or to provide assurances in regards to safety and effectiveness, we should reflect upon how opaque models change the practice of medicine and research. If healthcare is merely the practice of reporting accurate diagnoses, prognosis, and prescribing the correct regimen, perhaps this shift ought not concern us. However, if healthcare is more than this, an important consideration is how the implementation of black box medicine might enrich yet also impoverish the practice of medicine.

Section 3 key messages:

- **The idea of 'black box medicine' captures wider concerns about how opaque forms of computational modelling might change the practice of medicine.**
- **Price distinguishes between 'literal black box medicine,' where 'relationships are totally hidden, even though the machine learning process is known' and 'practical black box medicine,' where the 'machine learning algorithm can examine data, determine a relationship, and state it, but the underlying biological relationship is too complex to be amenable to scientific understanding or clinical trials.'**
- **It is unclear how Price's distinction between these two kinds of black box medicine fits with the burgeoning literature on explainable machine learning.**

4. Interpretability

A model may be a black box because information about the model is restricted or because the model is relatively human uninterpretable. This begs the question: what is interpretability? What are its dimensions? How should we evaluate the interpretability of any given model? What methods exist to render an otherwise uninterpretable model interpretable? This section considers each of these questions.

a. Interpretability et al

There is no one term that captures all the literature on the interpretability of machine learning. For instance, 'comprehensibility' often references approximately the same concept as 'interpretability,' authors defining each as follows:[18]

> Interpretability 'as the ability to explain or present in understandable terms to a human.'[19]

> Comprehensibility demands 'that the user is able to understand the logic behind a prediction of the model.'[20]

However, even these two terms combined miss much of the literature that references similar concepts, for instance:

> Intelligibility 'users can understand the contribution of individual features in the model.'[21] [i]

> Decomposability 'each part of the model - each input, parameter, and calculation - admits an intuitive explanation.'[22]

> Simulatability considers a model to be transparent 'if a person can contemplate the entire model at once.'[23]

> Legibility 'is concerned with making data and analytics algorithms both transparent and comprehensible to the people the data and processing concerns.'[24]

> An explanation 'is the collection of features of the interpretable domain, that have contributed to a given example to produce a decision.'[25]

[i] N.B. Lou et al use both 'interpretability' and 'intelligibility' interchangeably in their paper.

> Interpreting 'is the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of.'[26]

> Explicability (in the context of agent behaviour) 'measures how close a plan is to the expectations of the observer, given a known goal.'[27]

Our preference is to use the term 'interpretability' as defined by Doshi-Velez et al (2017) above, this definition gaining significant traction and consensus in the field.

b. Dimensions of interpretability

There are many ways in which machine learning can be human interpretable or made human interpretable. The dimensions of interpretability, the contours of the concept as it relates to machine learning are outlined below with consideration of their applications, strengths, and weaknesses.

When considering interpretability of machine learning there should be clarity about what you are trying to render interpretable. For instance, do you want to understand how a model learns from a dataset? If so, algorithm transparency at Section 4(b)(i) may be the best solution. Alternatively, do you want to understand what features the trained model generally finds significant? If so, then global interpretability of the model discussed at Section 4(b)(ii) may best fit that description. Perhaps you have little interest in general function but have a strong desire to understand what features were significant for a particular instance of processing, that is, what the model found significant in a particular case. If so, local interpretability discussed at Section 4(b)(iii) is probably what you seek. Of course, perhaps all three elements are important to you. Nevertheless, having a firm idea of what you want to be interpretable is important as different dimensions of interpretability require different tools and some can be radically more difficult to implement than others.

i. Algorithm transparency

'Algorithm transparency' is defined as information 'about how the algorithm learns a model from the data and what kind of relationships it can learn.'[28] As noted in the Machine Learning Landscape report , machine learning models are trained rather than explicitly programmed. In the case of 'algorithm transparency,' we are given information about how the algorithm creates the model but *not* information on the specific model that is learned and how individual predictions are made.[29] For example, in the case of a convolutional neural network to classify images, algorithmic transparency might constitute an explanation of how the algorithm learns edge detectors and filters the lowest layers.[30] In this way, explanations of the algorithm give a sense of how the model is trained, how images will be segmented, but not what any given model produced will find significant. There are three main points to keep in mind when considering algorithm transparency.

First, in many circumstances, algorithmic transparency may be less demanding than transparency or interpretability of the trained model itself. For example, in regards to linear

models, it can be proven that trained models will converge to a unique solution, even for unseen datasets.[31] Accordingly, from the outset we have some understanding of how retrained models will function on unseen datasets. In some circumstances then, algorithm transparency provides some transparency to the models it might produce.

Second, some algorithms are less transparent than others. Indeed, modern deep learning algorithms are not as well understood as linear models and so the same guarantees cannot be offered at the level of the algorithm.[32] That is, the less we understand the algorithm, the less we can predict its future behaviour, hence algorithmic transparency will provide less information on how trained models will function.

Third, in the context of adaptive machine learning, algorithm transparency may be more important. Adaptive machine learning methods retrain 'every now and then' (in batches) or continuously (online).[33] Consequently, understanding how retrained models behave becomes more important if the trained model changes frequently.

Algorithm transparency is one tool that facilitates interpretability that may prove useful in a number of circumstances. One advantage of algorithm transparency is that it may provide a general understanding of how and what the model learns without disclosing further commercially sensitive details. Algorithm transparency may be of interest to a variety of audiences. For instance, other developers are likely interested in the method developed, an understanding of how a model is trained is likely key to assessing the safety and effectiveness of many systems, and algorithm transparency may be one way to provide a generalisable and relatively high-level explanation to users.

ii. Global interpretability

Global interpretability concerns the ability to 'understand the whole logic of a model and follow the entire reasoning leading to all different possible outcomes.'[34] Global interpretability can be demanding. Lipton (2016) emphasises that true global interpretability (simulatability) requires the observer to be able to 'contemplate the entire model at once.'[35] Given the immense complexity and size of many machine learning models, simulatability is often an impossible task, humans having difficulty comprehending even relatively simple models with uncluttered feature spaces, and simple weighting.[36] For example, consider Molnar's (2019) description of what would have to be kept in mind to render a Naive Bayes model simulatable:[37]

- The model may have many hundreds of features - keeping each of these in mind would be a feat of memory.
- We would then have to consider the weighting of each feature; even if we did this, we would not be able to make predictions for new data quickly.
- We would then have to contemplate the joint distribution of all features to consider the significance of each feature and the prediction average.

All things considered, Molnar notes that this is an impossible task. While simulatability may be impossible what can be done to at least grasp the 'whole logic of the model'?

Simulatability may be out of reach but modular understanding of some models is possible. While keeping an entire model in one's mind is impossible, there are interpretable parts of some models. For instance, following Molnar (2019):[38]

- In regards to linear models, the interpretable parts are the weightings
- In regards to decision trees, the interpretable parts are the splits and leaf node predictions

However even this modularity comes with caveats, Molnar (2019) cautions that interpretation of each weight (when considering linear models) is interlocked with all other weights. Consequently, the 'interpretation of a single weight always comes with the footnote that the other features remain at the same value, which is not the case with many real applications.'[39] For example, consider Molnar's instructive example:

> 'A linear model that predicts the value of a house, that takes into account both the size of the house and the number of rooms, can have a negative weight for the room feature. It can happen because there is already the highly correlated house size feature. In a market where people prefer larger rooms, a house with fewer rooms could be worth more than a house with more rooms if both have the same size. The weights only make sense in the context of the other features in the model.'[40]

While Molnar's example is not from healthcare or health-related research, we can easily imagine a similar scenario in this context. For example, in the context of health, instead of predicting house prices, we might want to predict the risk of dying from pneumonia. Instead of the market making a difference, a difference in training sets between hospitals might be relevant. Indeed, the Caruana (2015) set of models to do exactly this might illustrate the context-sensitive nature of the weighting of features.[41] The convolutional neural network in Caruana's paper appeared to learn the rule 'HasAsthma(x) ⇒ LowerRisk(x).' In this way, asthma was thought to be protective of dying from pneumonia, an obviously counterintuitive conclusion. However, Caruana notes that this rule reflects the truth in the dataset - those who had asthma were typically admitted straight to the emergency room for targeted care over and above those without asthma, thereby generating the counterintuitive conclusion that those with existing respiratory conditions being at less risk.[42] However, we can imagine contexts in which the rule would reverse - perhaps a hospital in a less developed country might not have such effective care for asthmatics, meaning asthma becomes a risk rather than a protective factor.

Global interpretability as simulatability is demanding. Modular understanding of some models may confer some understanding but any interpretation is limited by *ceteris paribus* - other things remaining the same. Even so, this modular understanding is often beyond many modern, complex machine learning algorithms such as deep neural networks. Given this, tools to assist with the global interpretability of machine learning models may be necessary. We consider some of these tools below at Section 4(c) below.

iii. Local interpretability

Local interpretability concerns 'only the reasons for a specific decision.'[43] In other words, we can 'zoom in on a single instance and examine what the model predicts for this input, and explain why.'[44] Following the example of a model to predict risk of death from pneumonia, local interpretability, instead of telling what the model generally finds significant, tells us what the model found significant for *that particular patient*.

As with global interpretability, there is still the issue of how to interpret a complex web of features with different weightings. However, as Molnar notes, local interpretations may behave better - some of the complexity in global interpretability falling away when we consider a particular prediction or local group of predictions.

There is no categorical separation between local interpretability and global interpretability.[45] For instance, we can take global methods, applying them to a subset of instances to ask what the model finds significant. Alternatively, we can take local methods, aggregating the results for an entire dataset.

c. Methods for interpretability

Interpretability has dimensions: we might seek to explain the algorithm that trains the model, explain the general function of the model (global interpretability), or a particular instance of processing of that model (local interpretability). In addition to this complexity, there are also various ways in which machine learning models might be interpretable and methods to render otherwise opaque models interpretable. Broadly, these methods exist on a spectrum from model-specific to model-agnostic.

> *Model-specific* means the method to render the model interpretable is specific to that model.[46]

> *Model-agnostic* means the method to render the model interpretable is, in principle, applicable to any machine learning model.[47]

The paragon of model-specific interpretability is for the model to be intrinsically interpretable. Indeed, the interpretability of the model in this case is by definition, specific to that model. On the other hand, a model-agnostic method can theoretically be applied to any given machine learning model, limited only by practicalities restricting its ability to interpret the underlying machine model. These methods intersect with the dimensions of interpretability: global and local interpretability (see Table 1 below).

Table 1: Dimensions of interpretability x methods of interpretability	**Global interpretability**	**Local interpretability**
Model-agnostic	These explanations explain the general function of (in principle) any given machine learning model.	These explanations explain specific instances of processing and (in principle) can be applied to any given machine learning model.
Model-specific	These explanations explain the general function of the model but are specific to only this model.	These explanations explain specific instances of processing but are specific to only this model.

We consider a selection of methods across this spectrum below.

i. Intrinsically interpretable

A machine learning model may be intrinsically (out-of-the-box) interpretable. There are a number of methods that are generally recognised as being intrinsically interpretable, namely: decision trees, rules, and linear models.[48]

The judgment that these methods are in fact interpretable is often made on the basis of the intuitive judgment of authors rather than by any empirical evidence. Regardless, empirical work examining the subjective understandability of these methods demonstrates that decision trees, rules, and linear models at least appear interpretable according to self-reported findings.[49] Other authors such as Molnar (2019) list interpretable methods as being: linear regression, logistic regression, GLM/GAM, decision trees, decision rules and RuleFit.[50]

ii. Global surrogate models

In the Machine Learning Landscape report , we noted that algorithmic modelling differs from data modelling by treating the underlying process as a black box. Consequently, instances of algorithmic modelling such as machine learning may have exceptional predictive accuracy but often do not illuminate the underlying process. One way to approximate the underlying rules behind such processes and retain predictive accuracy is to train two models.[51] First, a machine learning model that is uninterpretable but has high predictive accuracy. Second, a rule-based model using the same data that mimics the behaviour of the machine learning model. This method is known as training a 'global surrogate model.'[52] Training models in parallel like this may give some insight into the rules that structure the otherwise black box machine learning model. Indeed, this was the method by which Caruana et al found that their deep neural network to predict readmission risk had learned the rule: 'HasAsthma(x) ⇒ LowerRisk(x).'[53]

This method of interpretability has evolved, spawning more complex but related methods we consider below in Section 4(c)(iv).[54]

iii. Visualisation

Interpretability is sometimes a matter of effective representation of the model. Sometimes this problem includes reducing the complexity of a model to a digestible level. In this way, interpretability is often a matter of effective communication of what the model finds significant. The importance of visualisation and careful explanation of what is already interpretable should not be underestimated. Moreover, with respect to some models, visualisation may be one of the only efficient means to meaningfully convey some of the model's function. For instance, semantic maps (heat maps) can graphically demonstrate what an image classification found significant or indicate how it segmented the image. For example, a heat map applied to a model to identify fracture in x-ray images would highlight those elements of the image that the model found to be indicative of a fracture.

Visualisation as the sole method of interpretability can have shortcomings. For instance, semantic maps do demonstrate what the model found significant in the image but this can often be a thin explanation of model - lacking *why* the model found those areas significant.[55] Further, image classification models can be vulnerable to attacks known as 'adversarial examples.' In these cases, the model is tricked by inputting an imperceptibly small vector to an image, resulting in the misclassification of that image.[56] While semantic maps may highlight the imperceptible addition, we may be left with little idea of what caused such a misclassification.

iv. Post hoc methods

Suppose our model is not intrinsically interpretable. Suppose further that we wish to render it interpretable but have convincing reasons - perhaps in the form of accuracy trade-offs - to continue with the same uninterpretable underlying model. Models that are otherwise uninterpretable can be rendered somewhat human interpretable by using methods such as post hoc explainers. Parallel models and visualisation may indeed count as post hoc explainers. However, there is also a subset of XAI tools designed specifically to render otherwise uninterpretable machine learning interpretable. Notably, there are multiple potential benefits to separating the model from the explanation, namely: the developer is then free to pick a model regardless of its interpretability, the explanations produced by post hoc methods can be more flexible being easily modified, and if the post hoc explainer is model-agnostic, it can be applied to other models.[57]

There are multiple different kinds of post hoc explainers. For instance, consider the following methods:

> Partial dependence plots 'calculates and graphically represents the marginal effect of one or two input features on the output of an opaque model by probing the dependency relation between the inputs variable(s) of interest and the predicted outcome across the dataset, while averaging out the effect of all other features in the model.'[58]

> LIME 'works by fitting an interpretable model to a specific prediction or classification produced by an opaque system. It does this by sampling data points at random around the target prediction or classification and then using them to build a local approximation of the decision boundary that can account for the features which figure prominently in the specific prediction or classification under scrutiny.'[59]

> Counterfactual explanations 'offer information about how specific factors that influenced an algorithmic decision can be changed so that better alternatives can be realised by the recipient of a particular decision or outcome.'[60]

For a fuller consideration of a broad spectrum of tools, see ICO's *Project Explain: Explaining decision made with AI: Part 2: Explaining AI in practice*.[61] For our purposes, it is sufficient to note that each of these tools has strengths and weaknesses and can explain different elements of a model subject to different limitations. For instance, counterfactual explanations are example-based explanations. Notably, these explanations may do well explaining the smallest change to make to achieve the desired outcome and so are useful in terms of local interpretability but may lead to over-generalisation and misunderstanding in relation to global interpretability.[62] We consider further limitations of post hoc explainers in the next section.

1. Limitations of post hoc explainers

Using post hoc explainers to interpret black box models is promising but has a number of limitations:

I. Fidelity. Post hoc explainers often approximate the underlying machine learning model to explain its contents. Since these explainers estimate the underlying model they may provide inaccurate answers, especially if these explainers are highly localised and taken outside their local context.[63]
II. Partial explanations. Even if the post hoc explanation generated is correct, it may be incomplete and (potentially) instil a false sense of confidence.[64] For example, saliency maps provide a heat map overlay of an image, demonstrating what part of the image the model found relevant. However, knowing where the model is looking does not tell us what the model is doing with that part of the image.
III. Calibration of machine learning models. If the underlying machine learning model is a black box, it is difficult to calibrate the model in light of external information not input into the model.[65] For instance, suppose contextual information tells us that we have racial bias in the training set that underpins our convolutional neural network model. Our ability to manually adjust for this discrepancy without removing data points will be limited. In short, while post hoc explainers may help us diagnose the problem, this does not guarantee that there is a feasible solution.

Following these three weaknesses, authors like Rudin (2019) emphasise that the gains in interpretability by using intrinsically interpretable machine learning often exceed the cost of reduced accuracy.[66] That is, while the accuracy loss in choosing an intrinsically interpretable model is low, the gain that interpretability brings usually outweighs this loss. This underlines

the point that post hoc explanations for black boxes are not a shortcut to rendering models interpretable - they are imperfect and inappropriate in some circumstances.

d. Interpretability or accuracy?

What is the relationship between accuracy and interpretability? It has often been assumed that there is an inverse relationship between the accuracy of a machine learning model - the number of data correctly classified - and the interpretability of the model - the human understandability of that model.[67] In this way, gains for interpretability when choosing machine learning methods are necessarily losses in terms of accuracy. However, this adage has been questioned by others. Rudin (2018) notes that the accuracy-interpretability trade-off amounts to a 'blind belief' conjured rather than being generated from data.[68] Warning against the trade-off, Rudin notes that in her experience, the opposite is often true: the iterative improvement that comes with interpretable machine learning typically leads to better performance.[69] The tenor of Rudin's thoughts on this point is to question the often implicit assumption that interpretability necessarily comes at the cost of accuracy. Rudin does not say that there is never a trade-off between interpretability and accuracy. Rather, Rudin notes that often there is no trade-off to be had, and, where there is, an interpretable method is generally preferable. Consequently, the most interpretable machine learning model is not necessarily the least accurate and vice versa.

e. Why be interpretable?

We examine the question of 'why be interpretable' more fully in the Ethics of Transparency and Regulating Transparency reports where we explore the ethical and legal/regulatory rationale for interpretability. For the purposes of this report, Fox et al outline three main motivations that underpin the need for interpretable machine learning:[70]

I. The need for trust
II. The need for transparency
III. The need for interaction

In regards to trust, Fox et al note that if a clinician is to use a neural network to make a diagnosis, they need to be confident that there is a clear rationale behind that neural network's output.[71] Trust is discussed in Section 3 of the Ethics of Transparency report.

In regards to transparency, Fox et al note that the use of machine learning increasingly has legal consequences.[72] Accordingly, where a machine learning system makes the 'wrong' decision or where there is disagreement between the system and the human in the loop, understanding of how the decision was made may be necessary. We discuss some of the legal requirements that attach to machine learning in the Regulating Transparency report.

In regards to interaction, Fox et al tell us that if a machine learning system works together with humans, interpretability may be necessary to facilitate successful interaction between the machine learning and human elements. Indeed, if a healthcare professional must input some

value or contextualise the model's output, interpretability may facilitate this interaction. We discuss the importance of interpretability to facilitate interaction in the next section.

i. Working together

As described in the Machine Learning Landscape report, many near-use machine learning systems are assistive, that is, they augment, complement, and strengthen human clinicians, researchers, or consumers. A part of why many of these systems will not be fully automated is that humans and machine learning methods often have complementary strengths that mitigate against their respective weaknesses.[73] The hope is that by partnering the two together, the team will exceed the sum of its parts.[74] This teamwork requires interaction between the machine learning system and the human. Accordingly, many machine learning tools will involve *AI-advised human decision making* where the user - the healthcare professional, researcher, or health consumer in this case - 'considers the recommendation and, based on previous experience with the system, decides to accept the suggested action or take a different action.'[75] In this way, it is important to consider not just the machine learning model itself but how the human will interpret the model.[76]

There is a burgeoning literature that considers how humans interpret and interact with computers. The following distil some of the key points from this expanding literature. First, humans intuitively build 'mental models' of systems.[77] These mental models may accurately reflect the machine learning model or they may not. Second, mental models are often anchored in how the system performed in the past, in particular the errors observed.[78] That is, the observer will build a mental picture of a machine learning model based on its past performance. If the performance of the machine learning model changes, perhaps because the model is retrained, this update in performance can disrupt and impair the human-AI team's performance where the change is not communicated.[79] Further, as errors are better recalled than instances of correct decision making, errors may lead to a distorted view of the machine learning model. Third and critically, accurate understandings of a model facilitate users' willingness to rely upon the automated aid.[80] If an observer understands the machine learning model, they are more likely to trust its outputs in context and be able to operate the system to their satisfaction.[81] The likely conclusion of this is as follows: if the human-AI team is to reap the benefits of its partnership, interpretability and explanation of the machine learning tool's outputs will be important to facilitate successful interaction.

ii. Trust and interpretability

Arguably, the primary need for interpretability arises because most machine learning problems are necessarily incomplete.[82] If a problem was complete, there would be little room for interpretability - we would simply state the accuracy of the model and be content with that metric. To explain, conceptually, incompleteness is distinct from uncertainty in the following way:

> *Uncertainty* is quantified variance that can be formalised. For instance, we can calculate some forms of uncertainty using false positive rates and confidence intervals.

> *Incompleteness* references those elements that have not been formalised or quantified. For instance, we are yet to formalise and fully quantify ethical principles or other concepts such as safety.

In the foreseeable future, many, indeed, most machine learning problems will be incomplete. That is, problem formulations will be imperfect and some features relevant to the problem will go unformalized and unquantified. For instance, models to predict readmission risk often include many quantified features relevant to predicting risk of readmission.[83] However, given the variety of information that could be possibly relevant, these models will inevitably leave some features out. This is no indictment on these models as they may still be far more accurate than their human counterparts. However, it does give us good reason to look beyond accuracy and beyond out of sample error reporting to consider interpretability.

To be sure, an incomplete machine learning problem does not demand interpretability. As we note in the Interpretability by Design Framework, the work that interpretability will do depends on the context of the machine learning system. In this respect, there are three points to note.

A. If the consequences of the machine learning system are trivial, the ability to properly contextualise and interpret its outputs may be a matter of usability but not a matter of safety.
B. Even if the outcome of a machine learning system is consequential, machine learning systems often do not stand on their own. Indeed, if the area in which the system operates is well-studied and well-understood, we might rely on our contextual understanding of the area to contextualise and check the system's recommendations.[84]
C. Even if a machine learning system has consequential outcomes and there is little evidence to second guess the system, there are other strategies apart from interpretability to ensure the safety and effectiveness of systems. For instance, implementing alarms, inputting the ability to recognise out of sample errors, and so on.

While interpretability may not be necessary for every machine learning system, it often pays dividends. Indeed, Rudin argues that, in some circumstances, if we must choose between a more accurate model or a more interpretable model, we should generally prefer the more interpretable model.[85]

f. What is the bar for interpretability?

One reason to be content with relatively opaque models is that they are no more opaque than healthcare professionals and their 'black box brains.' This speaks to the standard of interpretability we require (if we require interpretability at all) of machine learning models. Where do we set the bar for interpretability for machine learning models?

i. Black box brains

A common defence of opacity in machine learning is that machine learning models are as transparent, or perhaps even more transparent, than the human alternative.[86] This defence of

opacity in machine learning relies on two premises. First, that machine learning is in some sense as, or more interpretable, than its human equivalent. Second, is the often implicit inference that we should hold machine learning to the same bar of interpretability to which we hold their human equivalents. Both premises may be questioned.

It is true that human reasoning is often capricious. Where reasons for actions are given, often these reasons are post hoc rationalisations, inaccurate, or laden with bias.[87] Indeed, clinical decision-making is not immune from imperfect reasoning, the field of medical heuristics noting that clinical judgments are often underpinned by personal theories, assumptions, experiences, traditions and lore.'[88] However, as Pearl notes, we can forgive our lack of understanding of how the human brain works because 'our brains work the same way, and that enables us to communicate with other humans, learn from them, instruct them, and motivate them in our own native language.'[89] If we cannot communicate with or interpret machine learning models, we lack this method of interrogation.

It is not clear that we should be satisfied with machine learning that merely has the same opacity as the equivalent healthcare professional. Consider an explicitly programmed model - the features selected and the weighting of each feature tailored by hand - in this way the model's output will reflect the designer's understanding of the relative importance of each feature. As a consequence, we may request that the model be evidenced with ground truth - scientific and clinical evidence to underpin the selection and weighting of features. If a machine learning model is opaque - we do not know what features the model found significant - evidencing the predictive accuracy of the model to ground truth seems an intractable task (see the Interpretability by Design Framework for more). Of course, we should retain the ability to be surprised by machine learning models; we should be willing to challenge scientific and clinical orthodoxy with inferences made from even relatively opaque machine learning models. Nevertheless, extraordinary conclusions require extraordinary evidence and use of models in healthcare require comparatively strong evidence to ensure the model is safe and meets its intended use. In the context of debugging and discovering confounding factors then, there may be reason to more fully interrogate an uninterpretable machine learning model. All of this begs the question, how should we evaluate interpretability of machine learning?

ii. Evaluating interpretability

Interpretability divorced from a task or domain is often thought nonsensical.[90] That is, evaluation of interpretability is domain and possibly task-specific. This aside, suppose we wish to evaluate interpretability of any given model with respect to a task - what experiments should we run to assess the model's interpretability? Doshi-Velez et al (2017) provide a taxonomy of evaluation approaches for interpretability, distinguishing between functionally-grounded evaluation, human-grounded evaluation, and application-grounded evaluation (see Table 2 below).[91]

Table 2: Evaluating interpretability		
Evaluation method	**Definition**	**Example**
Functionally-grounded evaluation	'Requires no human experiments; instead, it uses some formal definition of interpretability as a proxy for explanation quality'[92]	If we are confident that certain classes of machine learning model are generally interpretable, for instance, decision trees, these assurances may evidence the interpretability of our decision tree model[93]
Human-grounded evaluation	'Is about conducting simpler human-subject experiments that maintain the essence of the target application.'[94] This kind of evaluation involves human experiments with simplified tasks	Display different explanations for users, having them pick their favoured explanation[95]
Application-grounded evaluation	'Involves conducting human experiments with a real application.'[96] This kind of evaluation includes real human experiments with real tasks	Often takes the form of user studies with a focus on ensuring the system meets its intended use in its intended context[97] In the context of radiological image segmentation, a reader study ('a diagnostic accuracy study aiming to assess clinical performance of one technology versus another, on the basis of image interpretation by a group of human readers') might constitute an example of application-grounded evaluation[98]

The three evaluation methods represent a range of options to evaluate the interpretability of machine learning models. It is important to emphasise that each method of evaluation has its place - each method is likely part of the answer and appropriate for machine learning systems at different stages. For instance, application-grounded evaluation is likely to be experimentally demanding and expensive. However, even so, where there is a concrete application, for instance, assisting with the diagnosis of patients, the best way to show that the system meets

its intended use is to ensure that the assessment is as close as possible to real conditions, using actual patients and clinicians.[99] Still, functionally-grounded evaluation is likely important where models are relatively new and where experiments with real humans may be unethical.[100]

A Salient Feature | Roundtable 1

Participants questioned what the bar for interpretability should be - should we ask machine learning models to be more interpretable than human healthcare professionals? Should we ask for parity? Should the standard of interpretability for machine learning models be higher than their human equivalents?

Section 4 key messages:

- **Interpretability of machine learning is complex, there being no settled term to represent the literature, the concept having multiple dimensions. There are many methods to render machine learning interpretable, and varying views on whether interpretability necessarily comes at the cost of accuracy and what the standard for considering a model 'interpretable' should be.**
- **The literature encompassing explainable machine learning includes many related terms, for example: interpretability, comprehensibility, intelligibility, decomposability, simulatability, and so on. Sometimes these terms are synonymous, sometimes they mark out distinct but related concepts.**
- **Interpretability of machine learning has multiple dimensions. These dimensions emphasise different elements of machine learning and are useful for different purposes.**
- **Algorithm (untrained model) transparency includes information about 'how the algorithm learns a model from the data and what kind of relationships it can learn.' Algorithm transparency may be a useful and, to an extent, less demanding form of transparency compared to interpretability of the machine learning model itself.**
- **Global interpretability of machine learning models concerns the ability to 'understand the whole logic of a model and follow the entire reasoning leading to all different possible outcomes.' Following this strict definition, global interpretability is demanding indeed, especially where models are complex, including thousands of nodes, all of which are potentially contingent upon one another.**
- **Local interpretability of machine learning models illuminates 'only the reasons for a specific decision.' That is, local interpretability explains the outcome for a particular instance of processing. In healthcare, this will typically be the output for a particular patient or consumer.**
- **There are different ways in which machine learning might be interpretable or be rendered somewhat interpretable. Each of these methods have their own strengths and weaknesses - no method suits all purposes for which we might request a model to be interpretable.**
- **Machine learning models may be intrinsically susceptible to interpretation, or out of the box human interpretable.**
- **Global surrogate models train interpretable models alongside uninterpretable models to infer an explanation. The advantage of using a post hoc method such as this being that the developer may pick a model without regard for its interpretability, training a parallel model to infer an explanation, thereby avoiding an accuracy to interpretability trade off.**

- **Visualisation involves using visual representations of machine learning models to illuminate their function. For instance, semantic maps may be used to overlay a heat map onto images processed by image segmentation models to highlight what the model found significant.**
- **There is a suite of tools also under the umbrella of 'post hoc explainers' that seek to render otherwise uninterpretable models somewhat interpretable. Broadly, post hoc methods have their advantages, freeing developers to pick a model regardless of how interpretable the model is, bolting on a post hoc explanation instead. However, post hoc explainers also have weaknesses and risks. Notably, in terms of fidelity, these explainers are approximations. Further, even if accurate, post hoc explainers sometimes only offer a partial explanation of the model or output of the model. Finally, even where the explainer is accurate and presents a sufficient explanation, the ability to manually calibrate uninterpretable machine learning models is often limited.**
- **There is not necessarily an inverse relationship between interpretability and accuracy of machine learning models. Nevertheless, there may still be trade-offs between the two: the most accurate model may not be the most interpretable model.**
- **There are multiple practical reasons to render a machine learning model interpretable that emerge from the explainable AI (and related) literature. These different purposes may require different methods and emphasise different dimensions of interpretability.**
- **Interpretability of a machine learning model may be necessary to facilitate successful interaction between the human and machine learning system.**
- **Interpretability may be necessary for users to rely on and contextualise the model's outputs. Indeed, most models are necessarily 'incomplete', meaning the model does not include or formalise all the relevant features relevant to the decision or task the model purports to assist with. Consequently, because most models are incomplete, it is important that their outputs be interpretable so their outputs can be contextualised.**
- **What is the standard for interpretability? Notably, human healthcare professionals can be black boxes - human reasoning being subject to bias and impulse. Nevertheless, we might forgive some of this opacity because we can interrogate and question human professionals. We ought to ensure that we have similar tools to examine machine learning, especially if the system is safety critical.**
- **There are three broad methods to evaluate the interpretability of machine learning models. Namely, functionally-grounded evaluation, human-grounded evaluation, and application-grounded evaluation. Each method has its appropriate place: application-grounded evaluation being the most thorough but the most difficult to perform, functionally-grounded evaluation being the easier of the three to perform but also being the least rigorous.**

5. Interpreting interpretability

The term 'black box' has entered common parlance to describe opacity in machine learning. Notably, this concept rolls together opacity due to intentional restriction of information and opacity due to the inherent interpretability of some machine learning algorithms. Interpretability is also contingent upon the observer. The same information may render a model interpretable for one observer, doing little to illuminate the model for another - interpretability is partially contingent upon the technical literacy of the observer.

Interpretability of machine learning is complex: the concept has multiple dimensions and there are many methods by which machine learning might be interpretable or rendered interpretable. Notably, different methods emphasise different dimensions of interpretability. For instance, example-based explanations best capture local interpretability, explaining the output of a particular instance of processing, whereas, for example, partial dependence plots may give insight into how the model generally functions. Accordingly, different dimensions may be more important depending on the purpose for which we seek interpretability or an explanation. As a consequence, different methods may suit the different purposes for rendering a model interpretable - for instance, interpretability to facilitate interaction, interpretability to generate reliability, and interpretability to comply with ethical or legal requirements. It is to these ethical and legal requirements that we now turn.

References

[1] Chollet F. *Deep Learning with Python*. Version 6. New York: Manning Publications; 2017: 2-3.

[2] Liu L, Chen PHC, Krause J, et al. How to Read Articles that Use Machine Learning: Users' Guides to the Medical Literature. *JAMA*. 2019; 322(18): 1808.

[3] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 23 February 2020].

[4] Flach P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge: Cambridge University Press; 2012: 13.

[5] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5): 5.

[6] Burrell J. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*. 2016; 3(1): 3-4.

[7] Ibid, 4-5.

[8] Ibid, 4.

[9] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5): 6.

[10] Lahav O, Mastronarde N, van der Schaar M. What is Interpretable? Using Machine Learning to Design Interpretable Decision-Support Systems. *arXiv*. 2018: 2.

[11] Freitas A. Comprehensible Classification Models: A Position Paper. *SIGKDD Explor*. 2014; 15(1): 1-10.

[12] Piltaver R, Lustrek M, Gams M, et al. Comprehensibility of Classification Trees - Survey Design. *17th International Information Society*. 2014.

[13] Price W.N. Black-Box Medicine. *Harvard Journal of Law & Technology*. 2015; 28: 421.

[14] Ibid, 433.

[15] Ibid, 433.

[16] Ribeiro MT, Singh S, Guestrin C. Model-Agnostic Interpretability of Machine Learning. *ICML Workshop on Human Interpretability in Machine Learning*. 2016.

[17] Pearl J. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution. *arXiv*. 2018: 1-8.

Pearl J. The limitations of opaque learning machines. In: Brockman J, (eds.) *Possible Minds: 25 ways of looking at AI*. London: Penguin Press; 2019: 13-16.

[18] Piltaver R, Lustrek M, Gams M, et al. What makes classification trees comprehensible? *Expert Systems with Applications*. 2016; 62(15): 333.

[19] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv*. 2017: 2.

[20] Martens D, Vanthienen J, Verbeke W, et al. Performance of classification models from a user perspective. *Decision Support Systems*. 2011; 51(4): 782.

[21] Lou Y, Caruana R, Gehrke J. Intelligible Models for Classification and Regression. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2012: 150.

[22] Lipton ZC. The Mythos of Model Interpretability. *ICML Workshop on Human Interpretability in Machine Learning*. 2016: 5.

[23] Ibid, 4.

[24] Mortier R, Haddadi H, Henderson T, et al. Human-Data Interaction: The Human Face of the Data-Driven Society. *SSRN*. 2015.

[25] Montavon G, Samek W, Muller K. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing.* 2018; 73: 2.

[26] Ibid.

[27] Chakraborti T, Sreedharan S, Kambhampati S. Balancing Explicability and Explanation in Human-Aware Planning. *arXiv*. 2017.

[28] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 23 February 2020].

[29] Ibid.

[30] Ibid.

[31] Lipton ZC. The Mythos of Model Interpretability. *ICML Workshop on Human Interpretability in Machine Learning*. 2016: 5.

[32] Ibid.

[33] Gepperth A, Hammer B. Incremental learning algorithms and applications. *European Symposium on Artificial Neural Networks*. 2016: 1-12.

[34] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5): 6.

[35] Lipton ZC. The Mythos of Model Interpretability. *ICML Workshop on Human Interpretability in Machine Learning*. 2016: 4.

[36] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 23 February 2020].

[37] Ibid.

[38] Ibid.

[39] Ibid.

[40] Ibid.

[41] Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015: 1721-1730.

[42] Ibid.

[43] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5): 6.

[44] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 23 February 2020].

[45] Ibid.

[46] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5).

[47] Ibid.

[48] Freitas A. Comprehensible Classification Models: A Position Paper. *SIGKDD Explor*. 2014; 15(1): 1-10.

[49] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5): 7-9.

[50] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/simple.html [Accessed 23 February 2020].

[51] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 23 February 2020].

[52] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/global.html [Accessed 23 February 2020].

[53] Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015: 1721-1730.

[54] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/lime.html [Accessed 23 February 2020].

[55] Rudin C. Please stop explaining black box models for high stakes decisions. *arXiv*. 2018.

[56] Weld DS, Bansal G. The challenge of crafting intelligible intelligence. *Communications of the ACM*. 2019; 62(6): 70-79.

[57] Ribeiro MT, Singh S, Guestrin C. Model-Agnostic Interpretability of Machine Learning. *ICML Workshop on Human Interpretability in Machine Learning*. 2016: 91-93.

[58] The Information Commissioner's Office, The Alan Turing Institute. *Explaining decisions made with AI: Draft Guidance for Consultation: Part 2: Explaining AI in Practice*. 2019: 54.

[59] Ibid, 61-62.

[60] Ibid, 65-66.

[61] Ibid.

[62] Kim B, Khanna R, Koyejo OO. Examples are not enough, learn to criticize!. *Advances in Neural Information Processing*. 2016: 2280-2288.

[63] Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Computing Surveys*. 2018; 51(5).

[64] Rudin C. Please stop explaining black box models for high stakes decisions. *arXiv*. 2018: 4.

[65] Ibid, 5.

[66] Ibid, 3.

[67] Gunning D. *Explainable Artificial Intelligence (XAI)*. DARPA. 2017: 14.

[68] Rudin C. Please stop explaining black box models for high stakes decisions. *arXiv*. 2018: 2.

[69] Ibid, 3.

[70] Fox M, Long D, Magazzeni D. Explainable planning. *IJCAI-17 Workshop on Explainable AI*. 2017: 1.

[71] Ibid.

[72] Ibid.

[73] Wang D, Khosla A, Gargeya R, et al. Deep Learning for Identifying Metastatic Breast Cancer. *arXiv*. 2016.

Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American Journal of Surgical Pathology*. 2018; 42(12): 1636.

[74] Kamer E. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. *IJCAI*. 2016: 4070-4073.

[75] Bansal G, Nushi B, Kamer E, et al. Updates in Human-AI Teams. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019: 2429-2437.

[76] Zhang Y, Sreedharan S, Kulkarni A, et al. Plan Explicability and Predictability for Robot Task Planning. *arXiv*. 2016.

[77] Kulesza T, Stumpf S, Burnett MM, et al. Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. *ACM SIGCHI Conference on Human Factors in Computing Systems.* 2012: 1.

[78] Bansal G, Nushi B, Kamer E, et al. Updates in Human-AI Teams. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019: 2429-2437.

Dzindolet MT, Peterson SA, Pomranky RA, et al. The role of trust in automation reliance. *International Journal of Human Computer Studies*. 2006; 58(6): 697–718.

[79] Bansal G, Nushi B, Kamer E, et al. Updates in Human-AI Teams. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2019: 2429-2437.

[80] Dzindolet MT, Peterson SA, Pomranky RA, et al. The role of trust in automation reliance. *International Journal of Human Computer Studies*. 2006; 58(6): 697–718.

[81] Kulesza T, Stumpf S, Burnett MM, et al. Tell Me More? The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. *ACM SIGCHI Conference on Human Factors in Computing Systems.* 2012: 1.

[82] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv*. 2017: 1-13.

[83] Caruana R, Lou Y, Gehrke J, et al. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015: 1721-1730.

[84] Rudin C. Please stop explaining black box models for high stakes decisions. *arXiv*. 2018.

[85] Rudin C. Please stop explaining black box models for high stakes decisions. *arXiv*. 2018.

[86] Lipton ZC. The Mythos of Model Interpretability. *ICML Workshop on Human Interpretability in Machine Learning*. 2016: 2.

[87] Kahneman D. *Thinking, Fast and Slow*. London: Penguin; 2012.

[88] McDonald CJ. Medical Heuristics: The Silent Adjudicators of Clinical Practice. *Annals of Internal Medicine*. 1996; 1(124): 56-62.

[89] Pearl J. The limitations of opaque learning machines. In: Brockman J, (eds.) *Possible Minds: 25 ways of looking at AI*. London: Penguin Press; 2019: 14.

[90] Karim A, Mishra A, Newton MH, et al. Machine Learning Interpretability: A Science rather than a tool. *arXiv*. 2018: 3-5.

[91] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv*. 2017: 2.

[92] Ibid, 5-6.

[93] Ibid.

[94] Ibid, 5.

[95] Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Learnpub; 2019. Available from: https://christophm.github.io/interpretable-ml-book/scope-of-interpretability.html [Accessed 23 February 2020].

[96] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv*. 2017: 4-5.

[97] Ibid.

[98] Gennaro G. The 'Perfect' Reader Study. *European Journal of Radiology*. 2018; 103(March): 139-146.

[99] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv*. 2017: 4.

[100] Ibid.

The Black box medicine and transparency report was funded by the Wellcome Trust as part of the 2018 Seed Awards in Humanities and Social Sciences [Grant Number: 213623/Z/18/Z].

We thank the Wellcome Trust for their support.

The PHG Foundation is a non-profit think tank with a special focus on how genomics and other emerging health technologies can provide more effective, personalised healthcare and deliver improvements in health for patients and citizens.

For more information contact:

intelligence@phgfoundation.org

UNIVERSITY OF CAMBRIDGE

phg foundation

making science work for health