Canadian Journal of Philosophy (2023), 53: 3, 224–239
doi:10.1017/can.2024.1

ARTICLE

On the Opacity of Deep Neural Networks

Anders Søgaard

Center for Philosophy of AI, University of Copenhagen, Copenhagen, Denmark
Email: soegaard@di.ku.dk

Abstract
Deep neural networks are said to be opaque, impeding the development of safe and trustworthy artificial intelligence, but where this opacity stems from is less clear. What are the sufficient properties for neural network opacity? Here, I discuss five common properties of deep neural networks and two different kinds of opacity. Which of these properties are sufficient for what type of opacity? I show how each kind of opacity stems from only one of these five properties, and then discuss to what extent the two kinds of opacity can be mitigated by explainability methods.

Keywords: deep neural networks; opacity; explainability; model size; mitigation

1. Two Kinds of Opacity

Deep neural networks (DNNs) are, by now, familiar to most scientists. They are used across all fields of science, for example, to screen for anthrax spores (Jo et al., 2017), to calculate phenotypic distances from butterfly photographs (Cuthill et al., 2019), to improve breast cancer diagnostics (Witowski et al., 2022), to monitor urban earthquakes (Yang et al., 2022), in materials science (Zhong et al., 2022), and so on. DNNs come in many different flavors. Our discussion will not depend on the specifics of the different architectures employed in the wild, but on five common, almost-across-the-board properties of DNNs: DNNs are big (have overwhelmingly many parameters), continuous (propagate continuous values), nonlinear (have nonlinear decision boundaries), instrumental (rely on parameters whose semantics is undefined), and trained in an incremental fashion, that is, on ordered sequences of randomly sampled batches of data. Conventional machine learning models—for example, customer risk estimation models or spam filters—had hundreds or thousands of parameters. Today, neural language models or computer vision models have millions or billions of parameters. Earlier generations of language models were symbolic, with manually defined rules over discrete vocabularies, combining these in mostly linear ways, whereas modern DNNs operate in continuous space and employ nonlinear activation functions. DNNs are instrumental in that they rely on the induction of latent variables that do not track input variables directly, and they are incremental in that parameters are induced by updating the weights for small batches of data at a time, presented in a particular order, and while often doing multiple passes over the data before arriving in a final model state, training is often extremely sensitive to the order in which data happened to be presented.[1]

[1]This, for example, is why some DNNs are trained with training curricula, arranging the training data from easy (on some metrics) to harder.

DNNs are also often said to be *opaque*.[2] Opacity of DNNs causes moral problems, depriving users of full agency (Vaassen, 2022) and obfuscating developer responsibility (Goetze, 2022), but philosophers of science have also discussed the implications of opacity of DNNs. Here, the focus has been whether opacity challenges the application of DNNs to science (Babic et al., 2021; Gunning et al., 2019; Price, 2018; Rudin, 2019; Sullivan, 2022b).

Few have considered *why* DNNs are opaque, or *where* their opacity stems from. That is:

What properties of DNNs are sufficient conditions for opacity?

This is an incredibly important problem, however. If we want to find good trade-offs between transparency and performance (Dziugaite et al., 2020; Johansson et al., 2011), or between transparency and fairness or differential privacy (Rust & Søgaard, 2023), we need to know what causes opacity in DNNs. If, for example, continuity is *not* a source of opacity, binarizing DNNs (Tang et al., 2017) will not make them transparent.

In the literature, each of the five properties mentioned above has been accused of being the source of opacity in DNNs—size, continuity, nonlinearity, instrumentality, or incrementality—but are they really *all* sufficient conditions of opacity? This is the question addressed here, shedding new light on the challenges presented by the opacity of DNNs, as well as on what can be done to mitigate the opacity of DNNs.

For this, it is helpful to distinguish between two kinds of opacity:

Definition 1.1 (Inference-opacity). A DNN is said to be inference-opaque when expert humans cannot, upon inspection, say why, in general, it predicts an output y as a result of an input x.

Definition 1.2 (Training-opacity). A DNN is said to be training-opaque when expert humans cannot, upon inspection, say how, in general, the parameters of the DNN were induced as a result of its training data.

The two notions of opacity correspond to two ways in which the predictions of DNNs on a data point x_0 can be explained: either in terms of features of the input data x_0 that the DNN is making inferences about—or in terms of past training data x_1, x_2, and x_3 that influenced predictions on x_0. This corresponds to how a doctor can provide two types of rationales for a cancer diagnosis: The first is of the form: "Look! The patch in the upper corner of this x-ray scan looks like a tumor, doesn't it?" The second is of the form: "Your symptoms are identical to past patients who had cancer." In the DNN literature, explanations of the second form are often referred to as training data influence attributions.

Training data influence attributions seem to intuitively answer important questions about why DNNs predict what they do, but they are generally unreliable and hard to evaluate (Karthikeyan & Søgaard, 2021): If I train two randomly initialized networks on the same data (in the same order), they will be influenced by different data points. Why? The answer is simple: What data points incur loss, will depend on the network's current state. Similarly, two exact copies of an initialized network trained on the same data, but in different order, will also be influenced by different data points.

We will, in other words, be interested in the following question:

What properties of DNNs are sufficient conditions for inference-opacity, and what properties of DNNs are sufficient for training-opacity?

[2]Opacity means different things across different scientific disciplines, for example, the term is used in linguistics to denote coreferential senses licensing different inferences. Here, opacity refers to a lack of transparency, as in Burrell (2016): "[DNNs] are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs." I will discuss how appropriate the term is in §9; in a sense, DNNs are not opaque at all.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

1.a Scope

Creel (2020) introduced a distinction between three notions of opacity: functional (algorithmic), structural (software), and run (hardware) opacity. Training-opacity and inference-opacity are both forms of functional (algorithmic) opacity, and I will have nothing to say here about structural and run opacity. Similarly, Sullivan (2022a) introduced a distinction between external and internal opacity, where external opacity concerns how DNNs model the data, and internal opacity concerns their internal dynamics. I will only consider internal opacity here.

The distinction between inference-opacity and training-opacity is loosely related to the distinction in Boge (2022) between what-opacity and how-opacity. Consider first Boge's (very informal) definition of how-opacity:

> H(ow)-opacity concerns the way in which a DNN automatically alters the instantiated function in response to data. (p. 59)

DNNs are function estimators, so what Boge is saying here, is simply that h-opacity concerns how DNNs are induced from data. What-opacity—or w-opacity—concerns "what was learned" (p. 61) and is said to be non-reducible to how-opacity. Boge also claims that what-opacity is unique to DNNs. If this (empirical) statement is true, however, what-opacity must differ subtly from inference-opacity, which is commonly attributed to other learning algorithms, such as kernel SVMs or stacked ensembles. I will not try to characterize the difference between the two dichotomies any further.

In this work, I will instead explore the 5×2 grid of the above properties (size, continuity, nonlinearity, instrumentality, incrementality) and the two definitions just cited (inference-opacity, training-opacity) to see *what* properties are sufficient for *what* type of opacity. First, however, I will discuss how, given a property ϕ and a definition of opacity, we can decide whether ϕ is a sufficient condition for opacity.

1.b Testing for opacity

First, a methodological concern: What is a good test of opacity? One way to think about opacity goes as follows: A source of opacity is a property ϕ of DNNs such that DNNs with property ϕ are opaque, and DNNs *without* property ϕ are not opaque, that is, transparent. Such thinking, however, assumes that there is only one source of opacity in DNNs—but this is an empirical question to be settled (below). If both ϕ and ψ are properties of a DNN δ and sources of opacity, removing ϕ is insufficient to make δ transparent. By analogy, if a book is hard to read because of its topic and its average sentence length, reducing the average sentence length is insufficient to make it easy to read. So instead, we need to consider the following: A source of opacity is a property ϕ of DNNs such that otherwise transparent, noncomplex machine learning models become opaque by adding ϕ. I therefore suggest the following practical test for sources of opacity:[3]

> **Test of ϕ as a source of opacity:** Take a non-opaque (transparent) machine learning model, e.g., a small decision tree or a small Naive Bayes classifier, and add ϕ to test if it becomes opaque.

The test explores the space between DNNs and easy-to-understand machine learning models such as small decision trees and Naive Bayes. We take these transparent models and modify them by adding ϕ in the most natural way. If the models turn opaque from adding ϕ, ϕ is a sufficient condition for opacity, a source of opacity. Let us first remind ourselves what decision trees and Naive Bayes models are:

Decision tree is a flowchart-like structure, that is, a nested if-else statement learned from data. Simple decision trees can be very intuitive, for example, a product review is positive if it contains the

[3]As pointed out by one of the reviewers, this is an application of Mill's method of difference.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

word "good," or if it contains the word "bad" and "bad" is prefixed by the word "not." Decision trees are nonlinear and non-incremental, that is, they rely on global statistics rather than statistics of small random batches of data, but are generally considered fully transparent (Mittelstadt et al., 2018; Pedapati et al., 2020).[4] The following decision tree, for example, classifies all red cylinders and blue boxes as "1," and all red boxes and blue cylinders as "0": **if** x_{color} = "red" **then** (**if** x_{shape} = "cylindrical" **then** y = "1" **else** y = "0") **else** (**if** x_{shape} = "cylindrical" **then** y="0" **else** y = "1"). Call this model for Model 1. Model 1 is inherently inference-transparent. This should be easy to see. Since the training is order-independent and, for most decision trees, computable by hand, decision trees are also training-transparent.

Naive Bayes is a linear classifier (if the likelihood factors are from exponential families) in which the individual weights are easy-to-grasp maximum likelihood estimates (Marques-Silva et al., 2020). The dynamics of the final model is thus similar to other linear models, for example, linear regression or perceptrons (see below), but learning is non-incremental and therefore insensitive to the ordering of training data. Naive Bayes is often said to be inherently transparent (Askari et al., 2020; Marques-Silva et al., 2020; Mittelstadt et al., 2018).[5] Since Naive Bayes is linear, it cannot express the logic expressed by the above decision tree, but a Naive Bayes model with a prior probability of 0.5 for "1," but likelihoods $P(x_{shape} = 'box'|1) = 0.9$, $P(x_{shape} = 'cylindrical'|1) = 0.1$, $P(x_{shape} = 'box'|0) = 0.5$, $P(x_{shape} = 'cylindrical'|5) = 0.1$, would make cylinders "0" and boxes "1." Call this model 2. Model 2 is inherently inference-transparent. This should again be easy to see. It is also training-transparent, because the maximum likelihood estimates can be computed by hand from the training data, and the process is not sensitive to the order in which the training data are seen.

Decision trees and Naive Bayes models are thus said to be inherently transparent. Both machine learning algorithms are easy to apply by hand and easy to compute by hand, as long as the model size is sufficiently small. This means that they are *neither* inference-opaque nor training-opaque by Definitions 1.1 and 1.2. The fact that binary decision trees can be transparent in spite of their nonlinearity, and that Naive Bayes models can be transparent in spite of their continuity,[6] already suggest that continuity and nonlinearity are *not* sufficient sources of opacity.

I have introduced two notions of opacity, inference-opacity and training-opacity, as well as a simple test for evaluating potential sources of opacity. Over the next five sections, I will consider each possible source of opacity in more detail.

2. Size

One possible source of both inference-opacity and training-opacity is model size, that is, the number of model parameters. DNNs have overwhelmingly many parameters, for example, modern large language models have tens or hundreds of billions of parameters. Many researchers have highlighted

[4]Formally, a decision tree can be defined as a directed acyclic graph that consists of three types of nodes: A **root node** that represents the entire dataset or the starting point of the decision tree. Several **internal nodes** that implement decision rules based on thresholding an attribute value, thereby splitting the dataset into subsets. Finally, **leaf nodes** represent class labels or prediction outcomes. Leaf nodes do not have outgoing branches. The decision tree learning algorithm recursively partitions the dataset based on the selected features and decision rules until a stopping criterion is met. This process aims to maximize the predictive accuracy or minimize the impurity of the resulting subsets at each internal node.

[5]Formally, Naive Bayes can be defined as follows: Let $x_1,\ldots,x_n$ be a set of n independent features or attributes, and let $y = \{0,1\}$ be a class label. The goal is to predict the class label y given the feature values $x_1,\ldots,x_n$. The Naive Bayes algorithm calculates the conditional probability of the class label y given the feature values using Bayes' theorem, but under the above independence assumption and because the marginal likelihood of $x_1,\ldots,x_n$ does not affect the ranking of the labels by their conditional probability, we can simply compute $\arg\max_y \Pi_i P(y|x_i)$ to find y. The resulting classifier is a linear classifier of the form $b + \mathbf{w}^T\mathbf{x}$ with $b = \log P(y)$ and $\mathbf{w}^T\mathbf{x} = \Sigma_i x_i \cdot \log p(x_i|y)$.

[6]There is an ambiguity here. Naive Bayes come in two flavors: for continuous and discrete input. But both exhibit continuous inference dynamics and reasoning with probabilities.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

this source of opacity (e.g., Shrestha et al., 2021, §5.2), who point out how DNNs use "millions of neurons, which makes them opaque." The intuition is this: Humans cannot be expected to familiarize themselves with n parameters for sufficiently high n, and they obviously cannot hold this many parameters in short-term memory in an attempt to reconstruct inference dynamics. Even deciding whether two DNNs have induced the same function can be hard. For a single layer with n parameters, there is $n!$ equivalent ways of permuting its weights—a number higher than 3.5 million for $n = 10$. Humans cannot, upon inspection, say *why* models with overwhelmingly many parameters predict y as a result of an input x. Large DNNs are therefore inference-opaque by Definition 1.1.

To see whether the number of parameters qualify as a source of inference-opacity, let us apply our test from above. Is a decision tree or a Naive Bayes model with a million or a billion parameters inference-opaque? The knee jerk response of many machine learning researchers will be to say "no." In Naive Bayes, for example, all conditional probabilities are statistically independent, so unlike in DNNs, we can consider each input dimension in isolation. So, the argument would run, can we not just iterate through the parameters, one by one, and identify the predictive variables this way? Clearly, the interdependencies in dense DNNs add to the complexity of these models, but Naive Bayes models with millions or billions of parameters are also opaque. To see this, we need to remember how not all problems have sparse solutions. If sparse solutions exist, they can be computed in polynomial time (Marques-Silva et al., 2020), but often they simply do not exist. Consider, for example, a classifier used to distinguish positive from negative movie reviews and let the input movie reviews be of increasing length. A 50-word positive review may, for example, include 4 positive words and 2 negative ones. A 100-word positive review may include 8 positive words and 4 negative ones. And so on. While we can (very crudely) explain such decisions by saying "these eight words outweigh these four words," this becomes impossible *in the limit* (as the length grows to infinity), since the number of predictive variables will quickly surpass the number of predictive variables that our short-term memory capacity admits. Imagine a decision tree with thousands, or even millions, of nested if-else statements. Such nested statements would be lost on most of us. This is in line with Humphreys, for whom "a process is said to be epistemically opaque when it is impossible for a scientist to know all of the factors that are epistemically relevant to licensing claims on the basis of that process" (Humphreys, 2009).

Size is, in other words, sufficient for inference-opacity, but how about training-opacity? Is a decision tree or a Naive Bayes model with a million or a billion parameters also training-opaque?

This does not necessarily follow from its inference-opacity. By Definition 1.2, a model is opaque if it is generally not feasible to explain the individual parameters as a result of the training data. Even for large decision trees and Naive Bayes, this is fairly easy, since individual parameters reflect global statistics that are easy to compute. It may not be easy to predict the entire model if sufficiently big, or its posterior distribution, but the value of a single parameter, for example, a likelihood $P(x_i|y)$, is easy to compute.

In sum, size is sufficient for inference-opacity, but not for training-opacity.

3. Continuity

Humans are arguably better at reasoning with discrete variables than with continuous ones. If this is true, symbolic representations and logical reasoning are more intuitive to us than vector representations and probabilities. Many philosophers and computer scientists have pointed to the continuous nature of DNNs as a source of both inference-opacity and training-opacity, for example, claiming how "discrete representations have the advantage of being readily interpretable" (Cartuyvels et al., 2021, §2.1). Many approaches to "explaining" DNNs have discretized their inference dynamics—by uptraining rule-based models to simulate DNNs or by jointly providing discrete explanations, for example, by text generation or concept activation. Other approaches, however, do not, including uptraining of linear models and saliency maps, suggesting that

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

continuity itself is not a source of inference-opacity. Note also how many scientific explanations in the natural sciences are continuous.

I will apply our opacity test, but obviously, we cannot add continuity to Naive Bayes, since the inference dynamics of Naive Bayes classifiers are inherently continuous. Naive Bayes classifiers are, in their own right, an argument for why continuity is not sufficient for inference-opacity or training-opacity. That we consider small Naive Bayes models inference-transparent and training-transparent, is proof that continuity is insufficient for both types of opacity.

That said, maybe probabilities are an exception to this rule? If the weights did not have a probabilistic interpretation, would Naive Bayes still be transparent (on a small scale)? Our reply is as follows: Our interpretation of Naive Bayes models may built on our every-day intuitions about probabilities, but so may our interpretation of other discriminative, linear models. Perceptrons are single-layered neural networks and discriminative Naive Bayes models, for example, but if we add sigmoids, we obtain logistic regression models, which admit for probabilistic interpretations Rosenblatt (1958). So even if ϕ is said to mean "being continuous and non-probabilistic," continuity still seems insufficient to generate opacity. Our final example would be a simple decision tree over continuous input features, for example, "if you are taller than 6 five feet and older than 15 years, then y_1, otherwise y_2." Such a tree seems very easy to understand, and if so, this is further evidence that continuity is insufficient for inference-opacity and training-opacity. One way to add continuity to decision trees will give you decision tree regressors. Such regression models are also widely considered inference-transparent and training-transparent.

In sum, continuity is neither sufficient for inference-opacity nor for training-opacity.

4. Nonlinearity

Many authors have seen the reliance on nonlinear activation functions as a source of both of inference-opacity and training-opacity, referencing DNN's "nonlinear structure which makes them opaque" (Joneidy & Ayadurai, 2021; Sachan et al., 2020). Why are nonlinearities supposedly hard to grasp? Clearly, nonlinear functions are generally harder to grasp than linear functions, as evidenced by mathematics textbooks and online lectures saying things such as "in order to understand what a nonlinear function is, it is essential to understand what a linear function is."[7] But are nonlinearities sufficient for opacity? I do not think so. The most common nonlinear functions in DNNs seem easy to understand: ReLU functions, for example, are thresholded, linear activation functions, for example, "$f(x) = ax + b$ unless $x < c$." We are familiar with such functions in our daily lives. This is how simple taxation systems work, for example. Sigmoid functions are arguably somewhat harder to compute with by hand, but they express natural relationships, for example, that between crop yield and soil salinity, or those between titrant volume and pH. A second reason to question that nonlinearity is the source of opacity in DNNs, is that DNNs often induce near-linear functions in practice and can be approximated very closely by shallow, linear networks (Ba & Caruana, 2014). They nevertheless come across as both inference-opaque and training-opaque.

Let us apply our test from Section 1. We cannot add nonlinearities to decision trees, for decision trees are already nonlinear. They are as such an argument for why nonlinearity is insufficient for (both types of) opacity. Small nonlinear small decision trees are considered transparent; see Section 1 for details. But they cannot be nonlinear and transparent, if nonlinearity is a source of (both types of) opacity.

If we instead add ϕ (nonlinearity) to a Naive Bayes classifier, we may, for example, get classifiers that associate low- and high-probability instances to the same class. Consider a model *trained* to

[7] https://study.com/academy/lesson/nonlinear-function-definition-examples.html.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

predict positive product reviews, but *used* to detect opinionated reviews. Opinionated reviews will tend to be the ones for which P(positive|$\mathbf{x}$) is either very high or very low. Such a compound model, albeit nonlinear, seems perfectly intelligible (fathomable) to humans, even if the nonlinearity adds a bit an extra step to the inference procedure. It is likely that the cognitive load of nonlinear activations is higher than for linear activations, but what our test shows is that nonlinearity alone is insufficient for inference-opacity or training-opacity.

One may of course reply that just because I can understand a single ReLU activation does not mean I can understand a sequence of them, especially with high-dimensional inputs. I would reply that high dimensionality is a reference to size and itself a source of opacity. If a sequence of ReLUs with low-dimensional input would lead to inference-opacity, this would show that nonlinearity is a sufficient source of inference-opacity, but I would argue that this is only the case for sufficiently long sequences of ReLUs—a recourse, again, to model size. To see this, consider the following sequence of ReLUs:

$$\text{if } x > 30 \text{ then } f(x) = 3x + 4 \text{ else } f(x) = x; \text{ if } f(x) > 15 \text{ then } g(x) = 4f(x) + 3 \text{ else } g(x) = x.$$

It is fairly easy to see that this function for small enough values of x ($x \leq 15$) will return x, and that for large enough values of x will return $12x + 19$. For mid-range values of x, it will return $4x + 3$. A sequence of three or more ReLUs will of course be more complex to trace by hand, but note how the complexity is comparable to that seen in decision trees, typically considered transparent machine learning models.

4.a Zerilli's fathomability/intelligibility distinction

Zerilli (2022) distinguishes between two kinds of inference-opacity, fathomability and intelligibility:

> Linearity and low dimensionality together can be said to make a system and/or its operations *fathomable*, in the sense that "a person can contemplate the entire model at once" […] But when dimensionality increases considerably, even a linear model will cease to be fathomable. Dense linear models, extensive rule lists and large ensemble methods are really of intermediate complexity, and in many cases too cumbersome for a person to work through step-wise in real time without losing their footing. Nonetheless, so long as linearity is a feature of the model, it will still be *intelligible*—i.e., inspectable—without necessarily being fathomable—i.e., inspectable all at once.

Note that fathomability implies intelligibility. The distinction between intelligibility and fathomability based on linearity seems inconsistent with our basic intuitions, though. To see this, consider a model which is both fathomable and intelligible, say, a small Naive Bayes model (e.g., model 2 in Section 1), producing a posterior probability $P(y|x)$ for any input data point x. The following decision tree is nonlinear: **if** $P(y|x) > 0.8 \vee P(y|x) < 0.2$ **return** "confident" **else** "not confident." The decision tree basically just tells us when Model 2 is confident or not, but the model is nonlinear, so the decision tree—in spite of its intuitive simplicity—is not intelligible—and hence, also not fathomable.

Nonlinearity is simply not a source of inference-opacity. Decision tree rules, even those that are easy for a person to contemplate or fathom, can be nonlinear. Think of the colored cylinder-box rule in the above (model 1), for example. Zerilli (2022) runs into a second problem, also, when declaring that DNNs are neither intelligible nor fathomable. Intelligible models, to Zerilli, are inspectable, but it is simply not clear in what sense DNNs are less inspectable than decision trees. Their internal computations are certainly no more complicated; in fact, DNNs can generally be translated into decision trees (Aytekin, 2022). Hence, if DNNs are always unintelligible and unfathomable, as Zerilli would have it, the same should hold for decision trees. (Note how only size will resolve this paradox.)

In sum, non-linearity is neither sufficient for inference-opacity nor for training-opacity.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

5. Instrumentality

The fourth contender for a sufficient condition for inference-opacity or training-opacity is instrumentality: DNNs are opaque because they are instrumental and their elements are ungrounded. This implication has been floated by several philosophers, complaining, for example, that DNNs "are opaque because there is no inherent meaning to the hidden nodes," or that they do not "work in accordance with the ways humans themselves assign meaning to the reality that surrounds them" (Landgrebe & Smith, 2021, p. 2070). This is also the intuition behind Boge's what-opacity, it seems.

The argument here is not just that DNNs have derived intentionality in Searle's sense (Searle, 1983). Decision trees, Naive Bayes models, and ball frames all have derived intentionality. The argument is that the hidden layers of DNNs have less than that. In a decision tree, a node is said to correspond to an input feature, which in turn corresponds to a property of an object or an event that the input example represents. In Naive Bayes, the conditional probabilities represent relations between such properties. The balls on a ball frame correspond to numbers, which may in turn represent objects or events. What about the parameters in the hidden layers of DNNs?

One way to think about hidden nodes is as if they were latent variables or weighted mixtures of other input variables.[8] Latent variables are common in science. Scientists use latent variables to prove the existence of—or determine the causal effects of—everything from black holes in the universe to bias in clinical diagnosis. And lumping several variables together as one is something we do all the time. When computing momentum, for example, we multiply mass and velocity. Mass, however, is itself the product of two input variables, namely volume and density. In a DNN with a hidden layer and two input variables, the hidden nodes are no more complex than using *mass* as a shorthand for the product of volume and density.

Let us apply our test. Is a small, binary decision tree still transparent if one or more of its nodes correspond to a compound input variable? Does a transparent Naive Bayes model remain transparent after adding a small set of latent variables? I think the answer is "yes" in both cases. A decision tree, in which a node corresponds to the compound input variable "mass," does not in itself seem opaque. Sufficiently small Bayesian models with latent variables, for example, Gaussian mixtures or topic models, also seem transparent. Note also that clustering methods, such as *k*-means or agglomerative clustering, induce categories that do not have direct real-world correspondences. These are also traditionally classified as transparent machine learning methods.

A possible reply, again, is that just because we can describe *some* compound variables as the product of two simple ones, does not mean I can understand more complex combinations of more input variables. However, this is a reference to model size, not to instrumentality itself. Instrumentality in itself is not a problem, but if a variable x is a (sufficiently complex) function of a sufficiently large number of input variables, the meaning of x can be hard to understand. True, but this is exactly the argument made here: We can only wrap our heads around sufficiently small variable combinations.[9]

Some methods for inducing the values of latent variables, for example, Monte Carlo or Gibbs procedures,[10] may not be training-opaque, but instrumentality is not on its own a sufficient source

[8]Some researchers have argued that these latent variables encode concepts through superpositioning (Elhage et al., 2022).

[9]Obviously, not all parameters are alike. It is easy to design decision problems that would require an exponential number of parameters, for example, $2^{N-1}+1$, in a single hidden layer, but which would require only a linear set of parameters in a network of $2\log_2(N)$ layers. So depth adds to complexity, but because you have more interactions in such networks. Generally, parameters can play more or less complex roles. A product of phase-shifted sine waves is arguably harder to wrap your head around than the summation of integers. Still, only the *number* of waves and integers are sufficient sources of opacity. So, we can say without asking the question of model opacity: Some architectures may take 1,000,000 parameters to express what other architectures express in 1,000 parameters, but across such architectures, opacity results from size.

[10]Monte Carlo and Gibbs sampling are techniques used in statistics and computational statistics for generating samples from a probability distribution.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

of training-opacity. Clustering methods that are both instrumental *and* incremental, for example, incremental *k*-means (Halkidi et al., 2012), are also training-opaque.[11] See the next section for a discussion of incrementality.

In sum, instrumentality is neither sufficient for inference-opacity nor for training-opacity.

6. Incrementality

Seeing incrementality as a source of opacity generally seems to have gone somewhat under the radar. Some authors have alluded to training history as a general source of opacity, for example, "learning algorithms are even more opaque because they do not rely on pre-specified instructions, but on evolving weights and networks of connections that get refined with each additional data point" (Faraj et al., 2018, §2.1), but such examples are rare.

Why is training history important? Clearly, how a model ended up in its final state is orthogonal to its inference-opacity in that state. Some neural network learning algorithms only update parameters when networks make incorrect classifications, but are randomly initialized. If a network ends up in state A after seeing N training data points, it is possible to arrive at an identical network B without seeing any data at all, simply by drawing B from the process of random initialization. The probability of this happening is low, but nonzero. Since A and B must be equally inference-opaque, inference-opacity is independent of the training history of A and B. However, how a model ended up in its final state is *crucial to training-opacity*. Since DNNs converge to local (not global) optima during training, training histories are important for understanding *how* the DNN came about. This includes what training data were influential in this process. Two identically initialized DNNs trained on different sequences of training data will have very different (and often uncorrelated) influence functions.

The non-incrementality of (standard) decision trees and Naive Bayes models means that they are training-transparent, that is, *not* training-opaque. Conversely, consider negligible DNNs with two parameters. Such models remain training-opaque because they are trained on random sequences of training data, the order of which is not stored—or deleted once training is complete. Even if we only have two parameters, say, to keep track of, each parameter is affected by hundreds, thousands, or millions of training instances, some of which may be seen multiple times during training, and the final value of each parameter is potentially sensitive to the order in which training data are seen.

Consider what Naive Bayes would look like if it were sensitive to its training history. Or simply think of Perceptrons, a linear model that is sensitive to the ordering of training data. On non-separable problems,[12] the final model can be very sensitive to training history, but even on separable problems, margins can be quite different across models trained on different sequences of randomly sampled batches of data. Perceptrons (and by extension, DNNs) are therefore training-opaque; Naive Bayes models are not. Decision tree algorithms, for example, C4.5, are mostly non-incremental. Some incremental algorithms do exist, however, and some of these, such as ID4 and ID5, lead to training-opacity. However, the family of decision tree induction algorithms also reminds us of the subtle difference between incrementality and training-opacity: *Some algorithms are incremental, but guaranteed to produce the same decision tree regardless of how the training instances are*

[11]The training-opacity of *k*-means follows from how the final model state is, in general, sensitive to the random initialization. We need to know how the centroids are initialized to predict the final model state—in the absence of strong convexity. Incremental *k*-means learns from small batches of data at a time and may be sensitive to the order in which the training data are seen.

[12]In machine learning, a separable problem refers to a type of problem in which the data points from different classes or categories can be perfectly separated by a decision boundary or hyperplane. In other words, it is a problem where you can draw a clear and distinct line, curve, or boundary that completely separates one class of data from another without any overlap.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

ordered. ID5R is an example of such an algorithm (Utgoff, 1989). Here, incrementality does not lead to relevant epistemic opacity with respect to *how* the model was induced as a result of the training data.

In sum, incrementality is sufficient for training-opacity (in the absence of convergence guarantees), but not for inference-opacity.

7. Interactions

A general objection to the discussion above would be that I have treated each potential opacity inducer in isolation. This is arguably far too generous. The problem with DNNs, some might say, is not that they are large, continuous, nonlinear, and so on, but that they are all of these things at once. My reply to this objection is twofold:

First of all, it is simply not true that all DNNs satisfy the conjunction of the five properties above. Many DNNs omit nonlinear activation functions; others rely on binary-valued parameters (Hubara et al., 2016). It is true that all (relevant) DNNs are large, instrumental, and incremental, but the same holds true for many other learning algorithms, some of which are usually considered inference-transparent, for example, online random forests (Lakshminarayanan et al., 2014). Here is a sample list of learning algorithms satisfying all three of the above properties, size, instrumentality (Inst), and incrementality (Incr):

Algorithm	Reference	Size	Inst	Incr
Online random forests	Lakshminarayanan et al. (2014)	+	+	+
FOGD	Lu et al. (2016)	+	+	+
NOGD	Lu et al. (2016)	+	+	+
Online multiple kernels	Sahoo et al. (2016)	+	+	+
PIKLM	Manica et al. (2018)	+	+	+

These algorithms not only present a challenge to philosophers who claim that DNNs present us with unique kinds of opacity, for example, Boge (2022), but they also show that the conjunction of instrumentality and incrementality does *not* cause inference-opacity. *Small* online random forests are clearly inference-transparent, and random forests are often said to be *inherently* inference-transparent.

Generally—and this is my second reply to the objection above—I have not found *any* examples in the literature of small and inference-opaque models, regardless of how instrumental they were, or how incrementally they were trained. A bold empirical hypothesis seems to emerge: *No small models are inference-opaque.* The discussion above, in my view, presents strong evidence that model size, in general, is sufficient for opacity. I have also presented arguments against seeing continuity, nonlinearity, instrumentality, or incrementality as sufficient for opacity. But I have *no* proof for size being the only sufficient source of opacity, of course. Such a proof would rely on an iteration over *all possible* sources of opacity, showing that none of them, except for size, are sufficient.

8. Mitigation

My main observation is that the number of parameters in a model is a sufficient source of its inference-opacity, whereas incrementality—training on small batches in ways that are sensitive to

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

the order in which the batches are presented—is a sufficient source of training-opacity (in the absence of convergence guarantees). Other candidates, such as continuity, nonlinearity, and instrumentality, are insufficient to establish opacity.

This observation comes with the promise of making research on DNN transparency more focused, highlighting the need for knowledge distillation or summarization in terms of models with fewer parameters. It also indicates that the opacity of DNNs is related to the opacity of many other learning algorithms. Nothing about the DNN model opacity is unique to DNNs. I will use this section to show that it also follows from this observation that inference-opacity mitigation amounts to computing something akin to an abstractiveness–faithfulness trade-off, known from traditional summarization problems, and that training-opacity amounts to finding a similar trade-off—but on a different scale.

8.a Abstractiveness–faithfulness

Many strategies for mitigating inference-opacity focus on linearizing or discretizing DNNs. Linearization and discretization may contribute to our understanding of DNNs, but as I have shown, these methods will, on their own, be insufficient for inference-transparency. Many methods exist for approximating DNNs with smaller DNNs. The so-called lottery ticket extraction methods, for example, have shown to lead to only small performance drops, even after pruning away 19/20 parameters. However, for modern DNNs with millions or billions of parameters, this is clearly insufficient for inference-transparency, since 1/20 of a million parameters is still 50,000 parameters. Saliency maps approximate DNNs with very small linear models, with just a single feature per input unit. These are much easier to wrap your head around, but research suggests that the approximations produced by existing methods are unreliable (Arun et al., 2021; Kindermans et al., 2019).

There is, in other words, a trade-off between how faithful these interpretations are, and how much inference-opacity they remove. This trade-off is familiar from the summarization literature and is often referred to as the abstractiveness–faithfulness trade-off:[13] "the number of unfaithful sentences […] increases as the summary becomes more abstractive […]" (Durmus et al., 2020). A sentence in a summary is unfaithful if it has no support in the original document, and an abstractive sentence is one that abstracts away from the details of the original document to deliver the message more succinctly. This trade-off is, thus, a trade-off between precision and how easy it is to quickly understand the document by reading the summary. If, for example, you write a one-page summary of *Lord of the Rings*, you have to toss out information and abstract away from details. If you only have *half* a page for your summary, you need to toss out some more.

An approximation of a DNN in the same way needs to balance the same two objectives: faithfulness to the original DNN and abstractiveness, that is, generalization over its dynamics to present a more succinct and easy-to-understand alternative. Our objective thus becomes a weighted minimization problem, minimizing a weighted sum of the discrepancy between our approximation and the original model, and the size of our approximation. Both terms are analogous to terms of learning theory, since the former is a loss, and the latter is a regularization term.[14]

[13]It is of course also related to similar trade-offs in learning theory, including the bias–variance trade-off.

[14]This amounts to the following kind of objective:

$$\arg\min_{\theta^a} \sum_{i \leq n} \ell\left(y_i^a, y_i^o\right) + \lambda \|\theta^a\|^0,$$

where λ is a parameter that controls how important model size is to us, balancing the two objectives, faithfulness and abstractiveness. Over n many data points, we minimize the discrepancy between the output of our original model y_i^o and the output of our approximation of it y_i^a. $\ell(\cdot,\cdot)$ is a loss function quantifying this distance. The weighted regularization term controls abstractiveness by penalizing the size of the approximation, its L_0-norm.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

This observation sets the direction for future work on making DNNs transparent because we can now apply standard learning theory to the inference-opacity problem. Scientists have recently argued how work on making DNNs transparent "both overstates the benefits and undercounts the drawbacks of requiring black-box algorithms to be explainable" (Babic et al., 2021) but with the application of learning theory, we can now quantify these benefits and drawbacks, as well as provide provable guarantees for transparency. We know from scaling laws for various tasks, for example, Kaplan et al. (2020), that high performance often requires large models, but uptraining—that is, transferring the knowledge of a complex DNN with desired performance to a smaller, more compact model (Ba & Caruana, 2014; Bucila et al., 2006)—provides proof-of-concept that even the behavior of very large models can often be summarized by much smaller ones.

8.b Incrementality

Can we apply learning theory to the problem of training-opacity in the same way? While we can mitigate inference-opacity by balancing abstractiveness and faithfulness, training-opacity seems to be more of an uphill battle. Asking how an incremental learning algorithm arrived at a model after training on random batches of data is like asking how someone came to be the person they are now. This would require a complete recording of past events in the person's life. Similarly, mitigating the training-opacity of a DNN requires its complete training history. The number of parameters you need to summarize in order to mitigate training-opacity is also much larger than the number of parameters you need to summarize to make a model w-transparent. If a DNN has n parameters and saw m batches of data, you need to summarize at least $m \times n^2$ interactions to summarize the network's training history, that is, the number of updates to n parameters. One obvious strategy for doing so is learning to predict the model from a small set of d dataset characteristics, reducing the set of parameters from $m \times n^2$ to $d \times n$, and predicting the final state rather than tracking all the intermediary state transitions. Since, in practice, m and n tend to be six to eight digits and d to be two digits, this reduces the complexity a lot. This glosses over the stepwise training history and may be insufficient for transparency, since the model itself (n) may be too large. If, instead, you try to predict properties of the output from properties of the input, for example, performance from dataset characteristics, or class distribution from dataset characteristics, the number of parameters can be reduced even further, but possibly at a severe information loss.

9. Human Baselines

DNNs are said to be opaque machines, and I have discussed the sources of their alleged opacity, and how to mitigate it. Here, in the final section, I ask: Opaque compared to what? Opacity is only of practical importance, if there is a transparent alternative. Where DNNs replace manual processes, the alternative is often a human being. Clearly, humans are not exactly *see-through.*

9.a Bushy eyebrows bias

Consider the following situation: Mary is a doctor. Throughout her career, she has seen many patients with sleep apnea. In an unlikely series of events, all the patients she saw with this diagnosis, however, also had bushy eyebrows. Today, she is seeing a patient called Eric, who is complaining about shallow and infrequent breathing during sleep, and Mary has to decide whether she thinks Eric suffers from sleep apnea. Eric also has bushy eyebrows. Mary diagnoses Eric with sleep apnea, but Eric has a hunch that it is not really about his sleep symptoms. Why was Mary looking at his eyebrows all the time? Eric decides to complain to the National Board of Health, but how will they decide whether Mary was biased by Eric's eyebrows when arriving at her diagnosis?

When it comes to the question of what moves us to do one thing or the other, we cannot trust ourselves or others. While philosophers from Descartes to Chalmers have often granted

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

introspection a privileged epistemic status, there is ample evidence that introspective reports can be unreliable (Pronin, 2009; Schwitzgebel, 2006). Mary, in other words, may not be aware of her bias and thus unable to confirm or deny it. And while you can learn a lot about a person by observing their behavior, when a doctor weighs evidence for and against a diagnosis and decides whether to operate or not, there is typically little behavioral evidence for why the doctor chose this or that. In our case, Mary's gaze led Eric to speculate something was wrong. But clearly such accidental giveaways cannot be relied on. It seems we would need a mind reader in order to truly open up human black boxes.

This is in sharp contrast with DNNs. During inference (a forward pass in the DNN), input is multiplied with DNN model parameters, transformed, and squashed through nonlinearities. The process may be over-whelming, typically involving thousands or millions of computations, but all is explicit and could, in theory, be printed out. The only reason what happens inside DNNs at inference time is not directly interpretable is complexity. The number of computations is much more than we can grasp. The same is true, of course, of computations inside humans.

I said we would need a mind reader in order to truly open up human black boxes. The idea of mind readers has fascinated us for millennia. Philip Breslaw, the first magician to perform "mind reading" on stage, played in 1781 at the Haymarket Theatre in London; later, Harry Houdini would do the same. Ancient mythology is filled with mind reading, magical creatures. In Japanese folklore, *satori* are mind reading monkey-like monsters said to live in the mountains of the Gifu prefecture. In the Book of Job, it is said that only God has private thoughts. God, in other words, is a mind reader. Mind reading and telepathy also figure prominently in modern myths, such as those of the superheroes in the DC universe. In Harry Potter, Voldemort reads minds, and memories can be extracted from our brains and stored in a so-called perceive, a form of memory basin. In the words of Albus Dumbledore, "one simply siphons the excess thoughts from one's mind, pours them into the basin, and examines them at one's leisure."

While brain image decoding is still in its infancy, it is clear that some of our thoughts can be observed by bystanders, if equipped with appropriate technology to measure the oxygenation or electricity in our brains. SSVEP-based spelling systems enable spelling from EEG scans with better than 90% accuracy, surpassing our accuracy when typing text messages on our smartphones. Neurotechnologies can, to a large extent, decode our mental states by analyzing neural activity patterns, and have led to the development of a wide range of educational, entertainment-related, and even military applications.

However, none of these mythical or technical mind readers tell us the exact computations happening inside human brains. Why not? Presumably, Albus Dumbledore or God would know. The exact computations are not *practically* important. Explanations in terms of billions of neurons are explanations but not useful explanations. Humans are inference-opaque and training-opaque, and if the best brain scanners opened up the lids of our black boxes, it would provide us with valuable information, but *not* make us completely transparent (to resource-bound humans).

Forgetfulness also contributes to the opacity of doctors. Consider the bushy eyebrow bias example. Doctor Mary may or may not be aware of how she has learned to associate bushy eyebrows with sleep apnea. She may not even be able to recall the faces of past patients any longer. Human doctors may also be sensitive to their initialization and the order in which they had their experiences, including possibly their education, but these data are, likely, forever lost.

Mary is principally opaque in the moment, but we also do not have access to the events that shaped her convictions over time. And here it seems our doctor will never become fully transparent, even if we had almost unlimited resources. For this would require not only perfect brain image decoding, but lifelong, continuous brain image decoding and, possibly, also monitoring of the surroundings. As well as maybe working out the dynamics between these two systems. That seems like an uphill battle.

The picture that emerges is: What happens inside DNNs at inference time is not inaccessible or black-boxed, but simply "too complex." Or, in the words of Beisbart, something that a human being

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

cannot wrap their heads around "in a reasonable amount of time" Beisbart (2021). This is what their inference-opacity amounts to. What happens inside humans at inference time *is* inaccessible—in the absence of on-the-fly brain image decoding technologies, that is—*and* too complex. What led the DNNs to induce the functions they represent from their training data is inaccessible in the absence of the training history, but not in its presence. While DNNs and humans form a continuum here, providing the training history of humans is considerably more difficult than providing the training history of DNNs. So again, DNNs seem less opaque than humans in practice. Humans may be preferable over DNNs for other reasons: a) they may exhibit superior and more robust performance, or b) they may satisfy auxiliary criteria, for example, have a different legal status than computer software, or be considered more dignifying in, say, some patient–doctor settings. But when it comes to opacity or transparency, DNNs have obvious advantages over humans.

10. Concluding Remarks

I have presented five common properties of DNNs and two different kinds of (internal, algorithmic) opacity. I have explored how these five properties and two kinds of opacity interact, to decide which of these properties are sufficient for what type of opacity. Both kinds of opacity stem from the individual properties of learning algorithms: Inference-opacity stems from size; training-opacity stems from incrementality. This has serious consequences for transparency techniques, such as discretization and linearization, which will be insufficient to obtain transparency. I discussed how the two kinds of opacity can be mitigated, and showed how optimal transparency is a trade-off between abstractiveness and faithfulness. Identifying opacity sources and presenting a framework for formalizing transparency, push the horizon for the research program of explainable artificial intelligence.

Acknowledgments. The author would like to thank Akos Kadar, Thor Grünbaum, and the anonymous reviewers, as well as audiences at University of California Los Angeles and University of Tartu. This work was supported by the Carlsberg Research Foundation's Grant CF22-1432: Center for Philosophy of Artificial Intelligence.

Competing interest. The author declares none.

Anders Søgaard is a full professor at the University of Copenhagen. Previously he worked at the University of Potsdam, as well as for Google Research. He holds an ERC Starting Grant, a Google Focused Research Award, and a Carlsberg Semper Ardens Advance.

References

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., & Kalpathy-Cramer, J. (2021). Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3. https://doi.org/10.1148/ryai.2021200267

Askari, A., d'Aspremont, A., & Ghaousi, L. E. (2020). Naive feature selection: Sparsity in Naive Bayes. In *23rd International Conference on Artificial Intelligence and Statistics*.

Aytekin, C. (2022). *Neural networks are decision trees.* Preprint,. arXiv:abs/2210.05189.

Ba, J., & Caruana, R. (2014). Do deep nets really need to be deep? In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc.

Babic, B., Gerke, S., Evgeniou, T., & Cohen, I. G. (2021). Beware explanations from ai in health care. *Science*, 373(6552), 284–286.

Beisbart, C. (2021). Opacity thought through: On the intransparency of computer simulations. *Synthese*, 199(3–4), 11643–11666.

Boge, F. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, 32, 43–75.

Bucila, C., Caruana, R., & Niculescu-Mizil, A. (2006). Model compression. In T. Eliassi-Rad, L. H. Ungar, M. Craven, & D. Gunopulos (Eds.), *Knowledge Discovery and Data Mining* (pp. 535–541). ACM.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society*, 3(1), 1–12.

Cartuyvels, R., Spinks, G., & Moens, M.-F. (2021). Discrete and continuous representations and processing in deep learning: Looking forward. *AI Open*, 2, 143–159.

https://doi.org/10.1017/can.2024.1 Published online by Cambridge University Press

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4), 568–589.

Cuthill, J. F. H., Guttenberg, N., Ledger, S., Crowther, R., & Huertas, B. (2019). Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. *Science Advances*, 5(8). https://doi.org/10.1126/sciadv.aaw4967

Durmus, E., He, H., & Diab, M. (2020). FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5055–5070). Association for Computational Linguistics.

Dziugaite, G. K., Ben-David, S., & Roy, D. M. (2020). *Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability.* Preprint, arXiv:abs/2010.13764.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread*. September 12, 2022.

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1), 62–70.

Goetze, T. S. (2022). Mind the gap: Autonomous systems, the responsibility gap, and moral entanglement. In *ACM Conference on Fairness, Accountability, and Transparency* (pp. 390–400). ACM.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI – Explainable artificial intelligence. *Science Robotics*, 4(37).

Halkidi, M., Spiliopoulou, M., & Pavlou, A. (2012). A semi-supervised incremental clustering algorithm for streaming data. In *16th Pacific-Asian Conference on Knowledge Discovery*. Springer.

Hubara, I., Courbariaux, M., Soudry, D., El-Yaniv, R., & Bengio, Y. (2016). Binarized neural networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese*, 169(3), 615–626.

Jo, Y., Park, S., Jung, J., Yoon, J., Joo, H., Kim, M., Kang, S.-J., Choi, M. C., Lee, S. Y., & Park, Y. (2017). Holographic deep learning for rapid optical screening of anthrax spores. *Science Advances*, 3(8). https://doi.org/10.1126/sciadv.1700606

Johansson, U., Sönströd, C., Norinder, U., & Boström, H. (2011). Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Medicinal Chemistry*, 3, 647–663.

Joneidy, S., & Ayadurai, C. (2021). Artificial intelligence and bank soundness: Between the devil and the deep blue sea – Part 2. In A. Petrillo, F. D. Felice, G. Lambert-Torres, & E. L. Bonaldi (Eds.), *Operations Management – Emerging Trend in the Digital Era, Chapters*. IntechOpen.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling laws for neural language models.* CoRR. abs/2001.08361.

Karthikeyan, K., & Søgaard, A. (2021). *Revisiting methods for finding influential examples.* CoRR. abs/2111.04683.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., & Kim, B. (2019). The (un)reliability of saliency methods. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI*. Springer Verlag.

Lakshminarayanan, B., Roy, D. M., & Teh, Y. W. (2014). Mondrian forests: Efficient online random forests. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (Vol. 27). Curran Associates, Inc.

Landgrebe, J. and Smith, B. (2021). Making AI meaningful again. *Synthese*, 198, 2061–2081.

Lu, J., Hoi, S. C. H., Wang, J., Zhao, P., & Liu, Z.-Y. (2016). Large scale online kernel learning. *Journal of Machine Learning Research*, 17(1), 1613–1655.

Manica, M., Cadow, J., Mathis, R., & Martínez, M. R. (2018). Pimkl: Pathway-induced multiple kernel learning. *NPJ Systems Biology and Applications*, 5. https://doi.org/10.1038/s41540-019-0086-3

Marques-Silva, J., Gerspacher, T., Cooper, M., Ignatiev, A., & Narodytska, N. (2020). Explaining naive bayes and other linear classifiers with polynomial time and delay. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 20590–20600). Curran Associates, Inc.

Mittelstadt, B. D., Russell, C., & Wachter, S. (2018). Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY: Association for Computing Machinery.

Pedapati, T., Balakrishnan, A., Shanmugam, K., & Dhurandhar, A. (2020). Learning global transparent models consistent with local contrastive explanations. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 3592–3602). Curran Associates, Inc.

Price, W. N. (2018). Big data and black-box medical algorithms. *Science Translational Medicine*, 10(471).

Pronin, E. (2009). The introspection illusion. *Advances in Experimental Social Psychology*, 41, 1–67.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Rust, P., & Søgaard, A. (2023). Differential privacy, linguistic fairness, and training data influence: Impossibility and possibility theorems for multilingual language models. In *40th International Conference on Machine Learning (ICML)*.

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable ai decision-support-system to automate loan underwriting. *Expert Systems with Applications*, 144.

Sahoo, D., Hoi, S., & Zhao, P. (2016). Cost sensitive online multiple kernel classification. In R. J. Durrant, & K.-E. Kim (Eds.), *Proceedings of the 8th Asian Conference on Machine Learning, Volume 63 of Proceedings of Machine Learning Research* (pp. 65–80). University of Waikato.

Schwitzgebel, E. (2006). The unreliability of naive introspection. *Philosophical Review*, 117, 245–273.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Shrestha, Y. R., Krishna, V., & von Krogh, G. (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, 123, 588–603.

Sullivan, E. (2022a). Inductive risk, understanding, and opaque machine learning models. *Philosophy of Science*, 89(5), 1065–1074.

Sullivan, E. (2022b). Understanding from machine learning models. *British Journal for the Philosophy of Science*, 73(1), 109–133.

Tang, W., Hua, G., & Wang, L. (2017). How to train a compact binary neural network with high accuracy? *AAAI Conference on Artificial Intelligence*, 31(1).

Utgoff, P. E. (1989). Incremental induction of decision trees. *Machine Learning*, 4, 161–186.

Vaassen, B. (2022). Ai, opacity, and personal autonomy. *Philosophy and Technology*, 35(4), 1–20.

Witowski, J., Heacock, L., Reig, B., Kang, S. K., Lewin, A., Pysarenko, K., Patel, S., Samreen, N., Rudnicki, W., Łuczyńska, E., Popiela, T., Moy, L., & Geras, K. J. (2022). Improving breast cancer diagnostics with deep learning for MRI. *Science Translational Medicine*, 14(664).

Yang, L., Liu, X., Zhu, W., Zhao, L., & Beroza, G. C. (2022). Toward improved urban earthquake monitoring through deep-learning-based noise suppression. *Science Advances*, 8(15).

Zerilli, J. (2022). Explaining machine learning decisions. *Philosophy of Science*, 89(1), 1–19.

Zhong, X., Gallagher, B., Liu, S., Kailkhura, B., Hiszpanski, A., & Han, T. Y.-J. (2022). Explainable machine learning in materials science. *NPJ Computational Materials*, 8(1). https://doi.org/10.1038/s41524-022-00884-7

Cite this article: Søgaard, A. 2023. On the Opacity of Deep Neural Networks. *Canadian Journal of Philosophy* 53: 224–239, doi:10.1017/can.2024.1