Understanding from Machine Learning Models

Emily Sullivan

forthcoming in *British Journal for the Philosophy of Science*

penultimate version

Abstract

Simple idealized models seem to provide more understanding than opaque, complex, and hyper-realistic models. However, an increasing number of scientists are going in the opposite direction by utilizing opaque machine learning models to make predictions and draw inferences, suggesting that scientists are opting for models that have less potential for understanding. Are scientists trading understanding for some other epistemic or pragmatic good when they choose a machine learning model? Or are the assumptions behind why minimal models provide understanding misguided? In this paper, using the case of deep neural networks, I argue that it is not the complexity or black box nature of a model that limits how much understanding the model provides. Instead, it is a lack of scientific and empirical evidence supporting the link that connects a model to the target phenomenon that primarily prohibits understanding.

1 Understanding from Minimal and Complex Models

A common view in philosophy of science is that simple idealized models provide more understanding than complex or hyper-realistic models (Bokulich [2008]; Kuorikoski and Ylikoski [2015]; Strevens [2008]). Simpler models are easier to understand. Simpler models are more tractable. Simpler models seem to answer more what-if or w-questions, and do a better job at highlighting salient difference-makers. Moreover, understanding how a model works seems necessary to understand the phenomenon that model seeks to capture.

However, as philosophers are gaining better insight into minimal models, an increasing number of scientists are going in the opposite direction by utilizing deep neural net (DNN) machine learning algorithms using large data corpuses to create classifications, make predictions, and draw inferences. One example is the *deep patient* model (Miotto et al. [2016]). This model takes as inputs large amounts of patient medical data and gives as an output a generalizable patient representation that can be used to predict future medical problems. The model provides surprising results. It is able to predict a wide array of medical problems, such as schizophrenia, attention-deficit disorder, and severe diabetes, with a higher degree of accuracy than competing predictive models. However, with this increased accuracy comes increased opacity.

If we begin from the lessons learned from philosophical work on understanding and minimal models, it appears that scientists are curiously opting for models that have less potential to increase understanding. DNNs are opaque to modelers, they are increasingly complex and have less modeler control, and the amount of w-questions they address are seemingly limited. Are scientists trading understanding for some other epistemic or pragmatic good when they choose an opaque and complex machine learning model? Or are the assumptions behind why minimal models provide understanding misguided? In this paper, I argue that model simplicity and transparency are not needed for understanding phenomena. It is not simply the complexity or opaqueness of DNN models that limits how much understanding they provide. Instead, it is the level of *link uncertainty* present—that is, the extent to which the model fails to be empirically supported and adequately linked to the target phenomena—that prohibits understanding.

To make my argument, I first consider a simple model to illustrate how models can explain and provide understanding (§2) before looking to cases of DNN models (§4, §5). I

clarify the roles that algorithms play in explaining phenomena and the importance of explanatory questions for understanding (§2). I argue that the principle way that algorithms are black boxed is by obscuring implementation at various levels (§3). I then argue that it is not presence of implementation black boxes that prohibits understanding. Instead, it is the level of link uncertainty present that prohibits understanding (§5). In the end, understanding phenomena is not directly dependent on model simplicity and transparency. It is high levels of link uncertainty that undermines understanding phenomena from opaque models.

2 Algorithms, Explanatory Questions, and Understanding

The use of algorithms in scientific inquiry is not new. An algorithm is simply a series of steps or set of rules that carries out an action or solves a problem. Any model that utilizes a simulation employs an algorithm. Algorithms on their own are not explanations. It is only when algorithmic models are used to answer a question about some event or phenomenon that they explain. Some examples: How is it possible that the eye evolved in so many diverse systems? Why is segregation so prevalent? Or what effect does carbon dioxide have on current and future weather patterns? Other distinguishing features of explanation (causal, counter-factual, law-covering, et cetera.) are still widely discussed. As a starting point, I will adopt the increasingly common view that explanation aims at understanding (De Regt [2017]; Grimm [2010]; Khalifa [2017]; Potochnik [2011], [2015]; Strevens [2008]). In a slogan: explaining why helps us to understand why.

The exact relationship between understanding and explanation is still the subject of widespread disagreement. Some question whether explanation is necessary or sufficient for understanding. It may be that one can gain understanding without explanation (Lipton [2009]), or it may be that in addition to explanation, agents need to meet other epistemic conditions to understand (Grimm [2014]; Hills [2016]; Lawler [2018]; Sullivan [2018]). I set these issues aside. The central issue in this paper is whether explanations that utilize complex and opaque models are unable to provide understanding of phenomena in virtue

of the fact that the model itself is not well understood or black boxed.[1] My arguments do not so much trade on any positive notion of what understanding or explanation is, but on whether, given a lack of information, it is still possible to gain insight about a phenomenon. I am concerned with what Humphreys ([2004], [2009]) calls *epistemic opacity*: the extent to which the process of the model and its derived output are inaccessible to scientists and modelers.[2]

Before considering the complex case of DNNs, it is worthwhile to first consider a simple case to illustrate the way that models explain and provide understanding of phenomenon.

Thomas Schelling was interested in understanding why so many human populations are segregated. He created a model that aims to simulate a neighbourhood where individuals act on simple preferences in order to see the conditions under which segregation occurs. Schelling's checkerboard model has been discussed extensively in philosophy (Rohwer and Rice [2013]; Mäki [2009]; Grüne-Yanoff [2009]) and in the social sciences (Clark [1991], [1992]; Bobo & Zubrinsky [1996]). The model is simple. It is a simulation that consists of a grid with two types of actors, A and B, where both types act on one simple preference—that at least 30% of their neighbours are the same type. The simulation follows a simple algorithm: if more than 70% of the actors adjacent to a particular actor are of a different kind, move that actor to the closest unoccupied space. Repeat until no actors move. The equilibrium result, after several iterations, is a segregated board.

There are several possible explanatory questions that one could ask of this model, and depending on the question, a different explanation is called for. First, one could ask: how does the simulation work? To answer this question, one needs to know the details of the algorithm including expected input and output. The basic algorithm is so simple that Schelling originally designed this simulation not using a computer, but with two different types of coins on a checkerboard. Since then, the model has been implemented on computational systems in many different ways and at varying levels of complexity

[1] The use of explanation in machine learning often corresponds to justification, such as explaining how a model operates in order to justify a decision. In this paper, I am using explanation in a broader sense, in terms of explaining a target phenomenon. Models are part of explanations that are answers to questions that enable understanding of the target phenomenon.

[2] The type of opacity I am concerned with is close to Burrell's ([2016]) use of opacity in regards to the way algorithms "operate at the scale of application".

(Muldoon et al. [2012]). Questions about how the model or algorithm itself works takes our focus away from the phenomenon it bears on. There is a distinction between understanding and explaining how the model works and using that model to understand a phenomenon of interest. If one is chiefly concerned about explaining or understanding how a given model is implemented, it is not necessary to know how the model maps on to some real-world phenomenon. The question of this paper, on the other hand, is to what extent understanding the model is necessary for gaining understanding of the phenomenon that the model explains.

Schelling himself was interested in explaining phenomena with his model. He asked whether it is possible that segregation could occur based on individual preferences alone without institutional racism, thus going beyond the algorithm and toward understanding possibilities surrounding a real-world phenomenon. In order to explain how it is possible segregation could occur, the explanation must include how the algorithmic model simulates a possible population that could be affected by segregation by identifying the key mechanism behind the algorithm and how it maps onto a possible population. In this case, the coins represent people of different races. The empty spaces represent move-in ready houses. The catalyst for moving is a preference of nearest-neighbours being the same race. This mapping allows us to interpret the results of the simulation as identifying a possible causal mechanism of segregation. Explaining segregation in this fashion is an example of what philosophers of science have recently called a how-possibly explanation (Rohwer and Rice [2013]; Reutlinger et al. [2017]). The model is used to explain how it is possible that a neighbourhood could become segregated through a possible causal mechanism. From this explanation, we are able to gain understanding of the possible mechanisms that could operate in real-world populations because it isolates key possible causal mechanisms.

However, how-possibly explanations stop short of answering how-actually questions or why-questions about actual real-world populations (Sullivan and Khalifa [2019]; van Riel [2015]). If we want to explain and understand why so many real-world populations are segregated, or why a particular population is segregated, we need to go beyond mere possibilities. The explanation employing Schelling's model must include details about how the algorithm simulates a *real* population, that is, how the key features

of the algorithm map on to real-world populations. Furthermore, the explanation needs to include empirical justification of the claim that individual racial preferences is a salient mechanism that drives real world population moving patterns (Mäki [2009]; Sullivan and Khalifa [2019]).

Without empirical evidence validating that the possible causes identified by Shelling's model are actual causes, there is no link connecting the model to the phenomenon. There is a high level of *link uncertainty*, that is, a lack of scientific and empirical evidence supporting the link that connects the model to the target phenomenon. At the time Schelling's model was introduced, it failed to explain or enable understanding due to the fact that there was no empirical evidence connecting personal preferences to the causes of actual segregation in real-world populations. It wasn't until many years after Schelling's model was introduced that it was tested empirically, providing some limited evidence that individual preferences are an actual cause of segregation, not merely a possible cause (Clark [1991], [1992]). Indeed, if we suppose instead that the empirical evidence suggested that all segregation is the result of institutional racism or individual racial *prejudice*, so that Schelling's mechanisms for segregation were never realized in any real-world system, then we would have no reason to think that Schelling's model uncovers any actual causes of segregation, and thus would not be able to explain segregation or enable any real understanding. Thus, Schelling's model only provides understanding in so far as there is there is a link connecting the model to the phenomenon.

The way of establishing the necessary link is with additional scientific evidence that supports the connection between the causes or dependencies that the model uncovers to those causes or dependencies operating in the target phenomenon. What constitutes the amount and kind of scientific evidence needed to reduce link uncertainty will differ depending on the phenomenon and the model. In the case of Schelling's model, link uncertainty is inversely related to the amount and quality of empirical evidence connecting individual preferences to causes of segregation and a lack of evidence that suggests a different overriding causal factor. Importantly, establishing the necessary link connecting the phenomenon to the model does not thereby replace the need for or the epistemic value of the model. The model still explains even once the link between the model and the phenomenon is no longer uncertain. Schelling's model could still provide insight into why

a population that has the preferences required for the affect to take place is likely to become segregated. In fact, it is precisely when the link uncertainty is removed that the model is able to explain and provide understanding.

I'll return to the notion of link uncertainty later. For now, the important takeaway is that when we consider DNN models and how their opacity may prevent understanding of phenomena, we cannot consider the model in isolation. The focus should not be unduly placed on how the model works, but instead consider the explanatory question we ask of the model, the role that the algorithm or model plays in the explanation, and the amount, quality, and kind of scientific evidence needed in order to connect the model to the target phenomenon. We now turn to considering model opacity and the impact black boxes have on understanding.

3 Black Boxes

Black box explanations are also commonplace in scientific inquiry. Many explanations in various domains obscure low level details to explain higher level causal mechanisms or non-causal dependencies. Examples include explanations of universality in physics and explanations of convergent evolution, among others (Batterman and Rice [2014]). One can gain understanding of these phenomena without knowing all the details. A special problem seems to emerge with DNN models because the opacity goes beyond simply black boxing low level or irrelevant details. Engineers or modelers that design DNN models themselves do not fully know how the model determines the output. If the modeler cannot explain how the DNN algorithm works, then how can it be used to explain or understand some phenomenon? In this section, I clarify that the black box at issue here is one of implementation. I discuss varying levels of implementation black boxes and the potential impact they have on explaining and understanding phenomena.

3.1 Implementation black boxes

One way algorithmic models can be black boxed is that some level of detail regarding how the model is implemented is obscured. A modeler may know the broad outline of the

algorithm's structure without knowing how each step is exactly implemented. In such a case, there is a black box around implementation; the implementation is either unknown or illegible to the modeler, the explainer, or the understander.

Consider a simple example of computing factorials. There are different ways one could implement factorials in a computer system. Two basic methods are an iterative process and a recursive process. Using the language Scheme, each process can be expressed as shown in figure 1 (Abelson et al. 1996, 32-34).

Recursive Process

```
(define (factorial n)
   (if  (= n 1)
        1
        (* n (factorial (- n 1)))))
```

Iterative Process

```
(define (factorial n)
   (fact-iter 1 1 n))
(define (fact-iter product counter max-count)
   (if (> counter max-count)
       product
       (fact-iter (* counter product)
                  (+ counter 1)
                  max-count)))
```

Figure 1

Code examples of a recursive and iterative process for computing factorials

Not only is each method syntactically different, but semantically and operationally different under the hood as well (see figure 2, and Abelson et al. [1996], pp. 32-4). Thus, the exact implementation makes a difference to how the simulation and the computer system operate. However, in many cases it is not important to know exactly how a simulation or computer system is implemented, as long as the inputs and outputs remain the same. If a climate model involves computing factorials, it is unnecessary for the modeler to know, or make explicit in explaining climate patterns, how exactly the factorials were implemented. Details regarding the implementation are unnecessary for explaining and understanding why a particular climate pattern emerged.

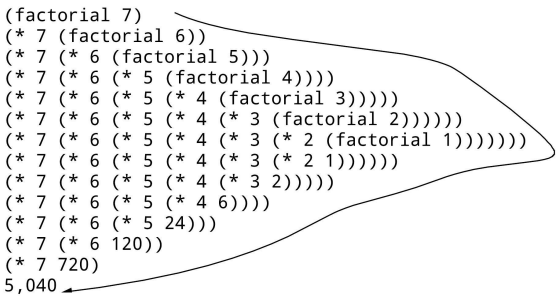

Figure 2a:
Recursive process for computing 7!

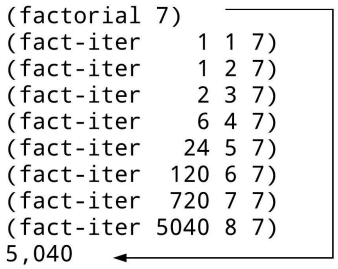

Figure 2b:
Iterative process for computer 7!

The irrelevance of implementation goes beyond simple computations. Implementation black boxes are present and similarly irrelevant to understanding the causes of segregation using Schelling's checkerboard model. There are countless implementations of Schelling's model that follow the same basic higher-level algorithm regarding satisfying neighbourhood preferences. In order to gain understanding of possible mechanisms of segregation, or actual mechanisms of segregation, one does not need to know whether Schelling's model was implemented using a functional, object-oriented, or actor-based language, even though these implementation differences make a difference to how the computer system operates and executes the algorithm. More drastically, in Schelling's case one does not even need to know whether the model was implemented on a computer system at all or whether it was implemented on a checkerboard, chessboard or a Go board in order to explain or understand segregation using the model. Thus, implementation black boxing in itself does not undermine our ability to explain or understand phenomena.

Of course, there are cases where the implementation matters. As argued above, the question and phenomenon of interest determine the scope of the explanation and level of detail necessary for success. If the low-level implementation details made a difference to the high level results of Schelling's model, then the implementation would matter for explaining and understanding segregation. However, as it so happens, these details do not make a difference. On the other hand, if the explanatory question concerns implementation or why building a model in a particular way is preferable, then knowledge of the

implementation is needed. In the factorial case, one may choose the iterative process if it is necessary to be able to track the state of the system at any point in the process, something that is not possible with a recursive implementation. So while it is possible for implementation black boxes to impede understanding, they are not in principle problematic for explaining or understanding phenomena.

3.2 Levels of implementation black boxes

Implementation black boxes can occur at varying levels, especially in an algorithmic model. Each step in an algorithm can often be broken down into further sub-steps. We can talk about the algorithm for the whole task, say the algorithm for computing factorials, or the algorithm for its each of its sub-steps, the sub-steps of its sub-steps, and so on. In the iterative process for computer factorials (figure 1), the first sub-step is `fact-iter`. `Fact-iter` has further sub-steps, such as computing multiplication and addition. In the iterative code example in figure 1, the `fact-iter` algorithm is not black boxed, but the multiplication and addition algorithms are. One of the goals in designing computer systems is to build modular systems. That way, a division of labour among engineers working on varying levels of the system can take place without each person needing to know how each of the sub-steps (lower level algorithms) are implemented.

While some levels of implementation black boxes do not get in the way of understanding, there do seem to be other instances of implementation black boxes that might pose a problem. For some algorithms, not all sub-steps are fixed or comprehensible to the modeler. It can be that in the course of running the simulation, some aspect of the algorithm itself changes or updates. In this case, there is less modeler control since the system is in an evolving state. Being in the dark about how the algorithm changes or is impacted through the execution of the simulation is just another instance of an implementation black box: there is something about the implementation of a higher level algorithm that is obscured.

Applying this to Schelling's model, we see that it utilizes these more mid-level implementation black boxes. The coins occupying the squares are moving around the board and continuously fed back into the program to determine which coin will move next. There

is an updating algorithm that is black boxed. Each time the simulation runs with different initial conditions, the ordering of movements, and which coins move, differs. The resulting segregated pattern is not known to the modeler beforehand because of the underlying complexity of the model. Despite this, the high-level algorithm stays intact: an iteration of coin movements until at least 30% of neighbours are of the same type. The result of the algorithm—a segregated board—also stays the same. In order to gain understanding of segregation using Schelling's model, you do not need to peer inside and see the implementation here. We have the same level of understanding of the mechanisms of segregation whether or not the implementation black box is removed. The explanation does not rely on the specific movements of, say, coin-267, but on the macrolevel emergence of a segregated pattern. This is true even if the algorithm is indeterminate, such that it is not even possible in theory to predict the resulting segregated pattern because the algorithm involves random choice of which actors to move next. Instead, as argued above, it is the link connecting Schelling's model to the phenomenon of segregation that would provide understanding. It is not implementation black boxing that gets in the way of understanding, it is link uncertainty.

That said, what if there is *nothing* about the algorithm that is known? What if we have the highest-level of implementation black boxing? This occurs when all the details of the algorithm are obscured—only the inputs and outputs are known. In cases where there is only a mapping of inputs to outputs, there is a strong intuition that the possibility for explanation and understanding is quite limited. Consider hypothetically that if all we knew of Schelling's model is that it takes in a board with two different coloured dots dispersed randomly as an input and gives back dots clustered into segregated patterns as output; it is hard to see how this can be used to explain segregation based on individual preferences, or anything for that matter. Similarly, if a black box computes factorials and we know nothing about what factorials are, then our understanding is quite limited. However, if we already know what factorials are, then this highest-level black box, for factorials, turns into a simple implementation black box that is compatible with understanding. This suggests that the level of the black box, which is coupled with our background knowledge of the model and the phenomenon the model bears on, matters for understanding, in addition to the type of explanatory question asked and the amount of link uncertainty present. Thus, when we

look to cases of DNN models, we need to ask whether we have the highest level of implementation black boxing, which is the question we turn to next.

4 The Black Boxes of Deep Neural Networks

Consider the deep patient DNN model designed by researchers at Mount Sinai (Miotto et al. [2016]). The model builds a generalizable abstract patient representation that can inform clinical decision making. The model takes as input electronic medical records represented numerically and gives as output the deep patient representation. In theory, the model can be part of a process used for several different tasks including identifying patients for clinical trials, detecting similarity between patients, and predicting disease. The researchers' first focus has been on disease prediction, aiming to show that DNN models can be applicable to a wide range of disease detection problems instead of being optimized for one specific disease, such as melanoma (discussed below). Down the line, researchers hope to build on this model to aid doctors in devising treatment recommendations. The model was trained on over 700,000 patients, with a subset of 76,214 patients analysed for disease prediction. Each of the patients in the subset had at least ten records between 1980-2013 and at least one new medical diagnosis in 2014. The model tested each patient against seventy-eight different medical problems ranging from severe diabetes and particular cancers to schizophrenia. Interestingly, the model was able to increase predictive accuracy (the proportion of true results, both true positives and true negatives, among the total number of examined cases) by 15%.

The deep patient model is quite impressive, especially given its high accuracy for predicting medical problems like schizophrenia, which can be difficult for physicians to predict. However, the model is largely opaque. One of the modelers of deep patient, Joel Dudley, expressed the point as follows: "we can build these models, but we don't know how they work" (Knight [2017]). Is it possible to gain understanding from the model despite its opacity? As I have been arguing, in order to consider how opacity limits understanding, we need to isolate the explanatory question of interest, the level of link uncertainty present, and determine the depth of the black boxes in the deep patient model,

and in DNN models more generally. In what follows, I first consider the latter and then return to explanatory questions and link uncertainty in section 5.

4.1 DNN structure

LeCun et al. ([2015], pp. 438) define deep-learning architectures broadly as "a multilayer stack of simple modules, all (or most) of which are subject to learning, and many of which compute non-linear input–output mappings". The basic multilayer structure of deep neural networks, including the deep patient model, is shown in figure 3. There are many different types of DNN models, such as convolutional neural networks (CNN), recurrent neural networks (RNN), and multilayer perceptrons (MLP). The exact structure and implementation of each type of DNN varies.[3] However, the nuances of these different techniques do not impact the larger argument concerning opacity and understanding; thus, the following discussion is a simplified overview of DNNs, which, in particular, highlights fully-connected layers.[4] DNN models are inspired by their namesake of neural networks in the brain; however, machine learning researchers are not aiming to model or simulate how a human brain works when building the types of models discussed in this paper.[5]

[3] For an accessible overview of different DNN techniques and methods see Guo et al. ([2016]) and LeCun et al. ([2015]). Generally speaking, recurrent neural networks (RNN) are ideal when data is sequentially ordered, as is the case with natural language or in time series. Convolutional neural networks (CNN), on the other hand, are often used when data has a clear spatial structure, as is the case with image classifiers (Shickel et al. [2018]).

[4] Different DNN architectures can include different types of layers that play different roles. For example, CNNs consists of the type of fully-connected layers described in the main text, while also first made up of convolution layers that create feature maps using local connectivity followed by pooling layers that reduce the dimensions of the feature maps (Guo et al. [2016], Shickel et al. [2018]).

[5] See Goodfellow et al. ([2016], Chapter 1) for a discussion on how DNNs used in computer science are inspired by the brain and the limits of this analogy. For a philosophical treatment of these limits, see Bailer-Jones and Bailer-Jones ([2002]). Alternatively, the field of computational neuroscience *does* seek to model the brain and draw inferences using DNNs (Glorot et al. [2011]). See Buckner ([2018]) for a philosophical argument that DNNs (CNNs in particular) capture processes of the brain. Since this paper is not about brain processing explanations in computational neuroscience, I leave these latter considerations aside.

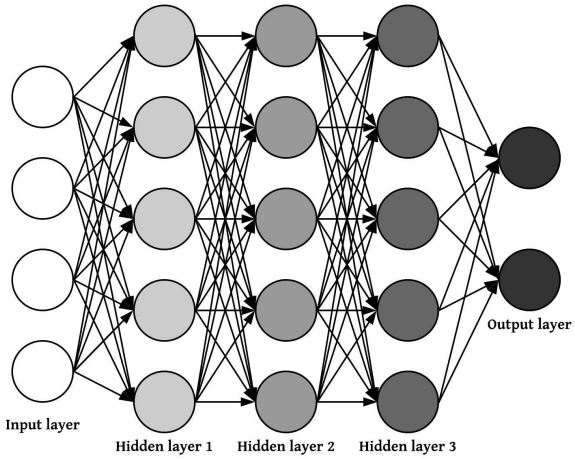

Figure 3:

Standard Deep Neural Network Structure

DNN models consist of an input layer, one or even hundreds of *hidden layers* (making the neural net *deep*), and an output layer. The edges that point from the nodes of the input layer to the nodes in the first hidden layer represent weights assigned to each piece of input data. The nodes of the hidden layer represent an *activation function* that is a non-linear function that takes as input the value of each of the previous nodes with its associated edge weight (and sometimes a bias value). The activation function that is often used is sigmoid, but modelers may choose other non-linear functions such as the hyperbolic tangent function or rectified linear units. The output of the activation function then serves as the new node input value for the next hidden layer (figure 4). The process repeats for each node in each layer, and for each edge connecting each node in each layer, until it reaches the output layer and delivers the final output. A DNN with hundreds of hidden layers could have hundreds of millions of connections and adjustable weights. The activation functions, the number of layers, the input data (and how it is represented), and the number of nodes in each layer are all determined by the modeler.

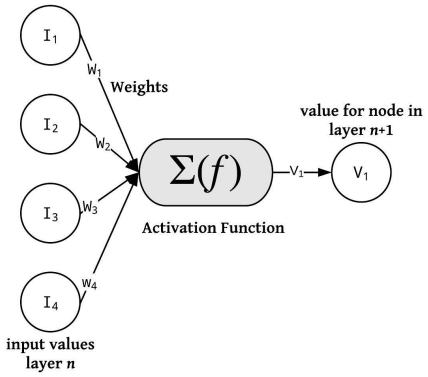

Figure 4:

Computing the value of a single node in the subsequent layer

DNNs are designed to learn which weights should be assigned to each feature in order to maximize predictive power and identify patterns in data that are not easily detectible by humans. While this is quite abstract, it has an intuitive basis. When a physician diagnoses someone with a particular disease, there are several data points that the physician considers and weighs differently. For example, the age of the patient is no doubt relevant, but presumably less relevant than a known blood marker for the disease. So, just as the physician weighs these pieces of information differently when making a determination, DNNs tease out the relevant features from the irrelevant for the task at hand. The exact representational relationship between the layers, and the exact functioning of the DNN, will differ depending on the phenomena modelled, type of DNN, and the choice of activation and learning functions. For example, in a standard image classifier, the first layer analyses the picture in the form of individual pixels, with numerical values representing the hue, saturation, and more. The next layer isolates a collection of pixels that start to pick out higher level arrangements, such as lines or edges. Each resulting layer gradually picks out higher and higher level abstractions until it reaches a classification of the image. In the deep patient model, the input data consists of electronic medical records of patients, each of which was pre-processed to identify clinically relevant phenotypes, and normalized to

lie between zero and one. The output of the model provides a more abstract patient representation where only the most important derived dimensions remain.[6]

4.2 DNN modelling process

The process of creating a DNN model starts with a training and learning phase, followed by validation and testing. Training, testing and validation are part of the modelling process that results in a model that can be applied to novel cases with success. A fundamental statistical assumption behind this procedure is the inductive learning hypothesis: that if a model[7] fits well over a sufficiently large set of training examples, it will also fit well over other unobserved or new examples. It is in the training phase that the machine *learns* and makes adjustments to the weights of each connection throughout the network.[8] The way in which the model "learns" which weights are optimal and corrects for error is through the backpropagation process, also determined by the modeler. This involves correcting for errors, often using stochastic gradient descent, to get the desired results. For example, in the case of a supervised learning image classifier, training data is labelled with the correct value, and the machine works through the backpropagation process to reduce error and settle on a set of weights that best captures the phenomenon.

After the model is trained, validation and testing are required to ensure the model is generalizable. The modeler determines what regularization methods to employ in order to avoid overfitting the data. For example, one regularization technique called dropout randomly omits a certain number of weights or activations (edges or nodes). The thought is that such a method can control against the model learning any odd particularities of the

[6] Some have compared DNN procedures to other coarse-graining abstraction procedures in physics, such as renormalization group methods (Mehta and Schwab [unpublished]). See also the information bottleneck method and its application to DNNs (Tishby et al. [1999]; Tishby and Zaslavsky [2015]). Despite the differences in the way that neural nets analyze data, they all seek to tease out what is relevant from the irrelevant for the task at hand.

[7] In machine learning, the word "hypothesis" is often used interchangeably with "model". A hypothesis is a function described by the set of weights that is presumed to capture the phenomenon. The *target function* is the function that will truly capture the phenomenon. Often the target function is an ideal that is not reachable; the goal is to settle on a hypothesis that is as close to the ideal target function as possible (see Mitchell [1997]).

[8] In more complex DNN models a *bias* parameter is also included in the weighted sum and can be modified through the learning process.

training data that would prevent generalizability (Baldi et al. [2013]; Wan et al. [2013]). The modeler then tests the accuracy of the model with data different from the training phase. If all goes well, then the modeler has a model that is able to generalize over the intended use cases. Importantly, modelers are not working completely in the dark. They rely on fundamental statistical assumptions and theories to ensure that the resulting model is generalizable.[9] Depending on the size and type of dataset, different statistical guidelines are utilized. For example, testing and training data should have a similar distribution to ensure statistical validity.[10]

The result of the modelling process produces a DNN model that follows its own algorithm that it learned through the modelling process. The modelling process is what determines the set of steps or rules that the resulting model will follow with any new input data it receives.

4.3 Levels of DNN black boxes

As argued in section 3, the level of the black box plays a central role in determining how much understanding is possible. Is the level of black boxing of DNNs problematic for understanding? It should be clear from the above discussion that DNNs are not black boxed at the highest level either during the modelling process, or in the resulting model.

First, the modelling process of DNNs involve basic implementation black boxing. The modeler is often working with higher level function calls where the exact implementation is unknown. For example, the modeler does not know how the sigmoid or gradient descent function is implemented, and would simply make a function call, like "sig(…)" when using the functionality to implement the activation functions for the nodes in each hidden layer. As argued above, this type of implementation black boxing does not hinder understanding phenomena. However, the implementation black boxes are deeper still. As the computer learns different weights for each data point, the output from each layer changes and serves as a new input for the next layer throughout the execution of the program. The modeler cannot predict which data points will be most salient, nor can the

[9] See Cristianini ([2010]) for a discussion on the history and use of statistical methods and machine learning.
[10] For a discussion of statistical guidelines, see Mitchell ([1997]) especially chapter 5.

modeler interpret the ways in which the machine settled on certain weights for certain pieces of data given the complexity and high number of data points. The implementation black boxing is increased when regularization methods like dropout are applied. The modeler does not even know which weights or activations will be deployed in a given iteration. Thus, the modeler does not have direct control over assigning weights to specific data points. In fact, usually the process starts with randomly assigned weights, with the system making changes through various iterations. However, the modeler relies on a wealth of knowledge and research about what methods to follow to build a generalizable model for the task at hand. So, while the modeler does not have direct control over the modelling process, a contrast case with Schelling's model, still the process is not black boxed at the highest level, such that it would prevent understanding of the phenomenon the resulting model aims to capture.

Once the model is trained, the modeler still has a general idea of how the finalized model works in virtue of having knowledge about how the model was trained and validated. However, it is only through indirect means that the modeler can investigate whether the model is picking up on what seems to be the most relevant features for the task at hand. For example, in an image classifier the modeler can deploy saliency maps where the model highlights certain areas of the image that was found to be the most relevant for its calculation. This goes a long way in determining the suitability of the model. Consider that there can be cases where the model focuses on aspects of the image that are clear proxies and not real difference makers, such as the time of day for a tank classifier, or the snowy background for a dog versus wolf classifier (Ribeiro et al. [2016]). Saliency maps can help us identify cases like this and rule the model unsuitable. Indirect methods like saliency maps still consist only of general approximations, not detailed assessments, with several low-level and mid-level dependencies remaining obscured. However, different types of saliency testing are enough to satisfy our need to know the high level details of how the model works to open the door to understanding the phenomenon the model bears on.

Putting all this together, when Dudley says "we can build these models, but we don't know how they work", he is saying that, on the one hand, we know enough about the modelling process (the higher level algorithms that train and create the deep patient model) such that one can build a model—and make intelligent changes to improve output and

prediction. On the other hand, the fact that there still remains low-level 'reasoning' and implementation black boxes in the resulting deep patient model—even after indirect saliency testing—undermines understanding. Thus, there is a lingering concern that even though black boxes of DNN models are not at the highest level, still the understanding we gain from these models is limited because of the lower level opacity in the DNN model itself. In the next section, I argue that when we focus on the types of explanatory questions we can ask of the models, the lingering problems for understanding that remain are not foremost due to the implementation black boxes, but because there is a certain level of link uncertainty.

5 Understanding, Explanation, and Link Uncertainty

Recall that the way that models explain depends in part on the explanatory question asked. In section 2, three different types of questions were introduced: questions directed at how the model works, how-possibly questions about a target phenomenon, and why or how-actually questions concerning a real-world phenomenon or target system. We saw with Schelling's model that implementation black boxes are compatible with understanding possibilities surrounding target systems and compatible with how-actually and why questions surrounding target systems. On the other hand, not surprisingly, implementation black boxes can inhibit understanding of how the model works.

Similarly, there are several questions that we could ask of DNN models, each with varying degrees of answerability. For example, in the deep patient case, we could ask questions about the structure of the model and how it works. We could ask simple classificatory questions, such as: Which disease is patient *x* likely to develop? We could also ask explanatory questions of the model, like the one the researchers were after: how is it possible to use a single model to predict a variety of diseases instead of relying on several models each designed for one disease? We could also ask more pointed questions: why is patient *x* predicted to develop disease *y*? Why are certain medical indicators associated with high risk for developing disease *y*?

The main challenge is whether the various implementation black boxes that prevent us from explaining and understanding certain aspects about how the model works also

prohibit explaining and understanding the *phenomenon* the model bears on. As we saw with Schelling's model, explanation and understanding of real-world segregation was only possible when the link connecting the model to the phenomenon was established, in other words by eliminating link uncertainty. On the contrary, knowing more about the model implementation did not increase understanding. It is my contention that the same is true with DNNs.

5.1 DNNs and how-possibility explanations

The deep patient model provides us with how-possibly explanations. Recall that a how-possibly explanation simply highlights a possibility concerning the causes or dependencies of some phenomenon; it falls short of explaining how the target phenomena *actually* is caused or the actual dependences concerning the phenomenon. The deep patient model can explain the researchers' main motivation—answering how it is possible to predict disease development for a range of diseases—simply by appealing to the input data and the higher level workings of the model discussed in the previous section. The model can also be used to explain how it is possible to predict schizophrenia (or any of the other seventy-seventy medical problems) through past medical records alone. Simply having a highly predictive model, and knowing the high-level emerging properties of the model, uncovers that it is possible to use a machine learning representation for disease prediction. Importantly, it is not necessary to look inside the implementation black box to answer these types of how-possibly questions. All that is needed is the higher-level understanding of how the system is able to identify high-level patterns within data.

With Schelling's model, peering inside the black box and examining how each coin moves around the board does not improve our understanding of how it is *possible* segregation can occur based on individual preferences alone. Similarly, with the deep patient model, learning more about the exact fine-grained weights of different data points, and the exact way the machine learning algorithm settles on and applies certain weights,

does not improve our understanding of how it is *possible* schizophrenia can be predicted and correlated with features found in medical records.[11]

However, things do seem to change when we move from a how-possibly explanation to a how-actually explanation or a more pointed why question. It does seem as though we cannot give a satisfying explanation for why a particular patient developed a particular medical problem using the deep patient model. After all, the model does not speak to *why* a certain marker is linked to a disease. Moreover, it also seems that we are unable to get a satisfying explanation for why it is actually the case that certain indicators reliably go hand in hand with a given disease using the deep patient model. Pointing to the gradient descent algorithm does not give us the right sort of insight here. We want some indication that the model is picking out the real difference makers for identifying a given disease and not proxies, general rules of thumb, or artefacts within a particular dataset. While it is tempting to attribute this gap in understanding to the implementation black box of the deep patient model, we should not take the bait.

5.2 DNNs and link uncertainty

Recall that link uncertainty constitutes a lack of scientific and empirical evidence supporting the link connecting the model to the target phenomenon. In the case of Schelling's model, link uncertainty was reduced after empirical evidence suggested that in real world populations individual preferences were a considerable causal factor that governed moving choices in segregated cities.

In the deep patient case, the model is greatly informed by existing empirical evidence concerning diseases. The modelers made particular determinations of which medical problems to include in their predictions and which ones to exclude. For example, they did not seek predictions of HIV because of the large behavioural aspect to the disease

[11] A worry that someone might have here is that since the knowledge of how a given DNN is trained is generally applicable to all (or most) DNN models, that this knowledge is epistemically independent from understanding a target phenomenon. However, generality *qua* generality does not make something epistemically independent. In this case, it seems that it is precisely because of the generality of the knowledge that it can help to uncover how-possibly explanations. The core of how Schelling's model works is also generally applicable to any type clustering behaviour. It is because of this general applicability that it serves as a possible explanation for segregation.

(Miotto et al. [2016], pp. 5). Having prior knowledge about which records are salient for medical diagnosis helped lead to the success of the model. However, gaps in medical knowledge still exist. Part of the reason for building the deep patient model is because the level of uncertainty about why certain patients develop certain medical problems remains high. Link uncertainty is prevalent. This is highlighted by the ways the model did not meet expectations. For example, the modelers found that the model had trouble predicting certain diseases that otherwise should have been predicted with ease, such as diabetes mellitus without complications. Their hypothesis was that since the screening process of diabetes often occurs during routine check-ups, the frequency of those tests was not a valid discriminating factor. This suggests that the model in part tracks proxies of disease development, such as previous physicians' decisions to carry out a diagnostic test. Given this, there is still link uncertainty that prevents understanding of real-world instances of disease development.

The deep patient model is still able to provide how-possibly explanations and understanding of possibilities. More than this it points to possible correlations that are worthy of future scientific and empirical research. It is the patterns that the deep patient model indicates that gives researches hypotheses to test and gain additional evidence for the strength of these patterns in real-world cases. Exploring these hypotheses further would reduce the link uncertainty and increase the level of understanding the deep patient model could provide on disease development. The type of scientific evidence needed to reduce the link uncertainty, in this case, could consist of building additional statistical models that makes the deep patient results more robust, conducting clinical trials, or conducting various longitudinal studies. The scientific standards of evidence for the domain in question determines how to establish an acceptable link connecting the model to the phenomenon. The stronger the link, the greater possible understanding the model can provide.

As we saw with Schelling's model, once the link uncertainty is resolved, the additional empirical evidence does not replace the usefulness of the model to explain and enable understanding. So too in the deep patient case, once the link uncertainty is resolved, the deep patient model is able to explain and enable understanding of disease development. Indeed, it is precisely because of reducing link uncertainty that this understanding is possible. For example, physicians can use the model, along with the link connecting the

model to the phenomenon, to explain and enable understanding for patients about their risk factors.

5.3 Differences in understanding; differences in link uncertainty

In order to strengthen the case that it is the level of link uncertainty present that prohibits explanation and understanding of phenomenon, and not model opacity, in what follows I will consider two additional DNN models that have the same level of model opacity and black boxing, but differ in their level of link uncertainty. One model has a lower level of link uncertainty compared to the deep patient model, the other has a higher level of link uncertainty compared the deep patient model.

Consider a DNN model that identifies cases of melanoma (Esteva et al. [2017]). The model works similarly to the deep patient model except that instead of medical records, images of melanoma and healthy moles serve as inputs to the system. The model is trained with semi-supervision, where the training set includes accurate labels of the images, and is then applied to a novel set of images to classify. The results of the model are significant, with the researchers claiming it outperforms dermatologists at classification.

The important point for our purposes is that the melanoma DNN model has the same level of implementation black boxing as the deep patient model. And like the deep patient model, there are several explanatory questions that might be interesting to ask of the model, each with varying levels of answerability. Some of these questions are how-possibly questions and some of these questions are more pointed why- and how-actually questions. However, compared to the deep patient model, the link uncertainty is greatly reduced. The level of scientific justification and background knowledge linking the appearance of moles to instances of melanoma is extensive. Visual appearance serves as the leading deciding factor for initial medical intervention, and for explaining why these interventions were made, such as explaining why a mole is more likely to be cancerous, or why a particular mole should be biopsied. The DNN model uses the visual appearance of the mole and immediate surrounding skin to identify problematic skin lesions. In the process, correlations between appearances of moles and the likelihood of the mole being healthy is found. The model can help physicians gain understanding about why certain

medical interventions are relevant, and using the model can help explain medical interventions to patients. Moreover, the model can discover new visual patterns that are highly correlated with health or disease. This can further understanding, especially once these newly discovered patterns undergo further investigation.

Implementation black boxes do not get in the way of understanding phenomena in the melanoma case because the model is operating within a background of existing scientific understanding. So, although we do not understand all the low level details about how the model works, and even though the model is complex in its data-points, we gain understanding of skin classification nonetheless. There are limits to the model. The model was trained primarily on white skin, for example, and thus is unreliable on other skin tones. This raises important medical ethical questions; however, the limited scope does not take away from the understanding that we gain using the model. Understanding is narrowed to a population subset, which is common in medical sciences. Just like many other scientific models, the usefulness of the model depends on the target system and the explanandum. If certain parameters change, the given model ceases to be the right model for explaining.

Lastly, consider the other extreme: a DNN model that has even greater link uncertainty compared to the deep patient case. Researchers developed a facial recognition model that seeks to identify the sexual orientation of individuals (Wang and Kosinski [2018]). This model uses roughly the same method as the melanoma model. The input data consisted of images of heterosexual men and women along with images of openly self-identifying gay men and lesbians. The images were of exclusively white American men and women taken from dating websites where users documented their orientation. The model is able to give striking accuracy in identifying sexual orientation when the model had five images of the same person. In the scenario where the model was presented with two faces, one of which was an image of someone who self-identified as gay and the other an image of someone who self-identified as straight, the model had a 91% labelling accuracy rate for men and 83% for women.[12]

The same level of implementation black boxing is present in this DNN model as with the deep patient and the melanoma model. What differs is the level of link uncertainty. The researchers built the model for two scientific purposes. First, to see whether it was

[12] The accuracy metric used here is AUC accuracy or the receiver operating characteristic curve.

possible to identify an individual's sexual orientation based on facial features alone. Second, to add evidential support for the parental hormone theory (PHT), an origin theory for sexual orientation. According to the theory, same-gender sexual orientation stems from the underexposure of male foetuses and over-exposure of female foetuses to prenatal androgens. Such a theory predicts gay and lesbian individuals would display gender atypical features (LeVay [2010]).

Just like with the deep patient case and the melanoma case, there are how-possibly explanations that the model can answer.[13] One such question is how it is possible to determine, just by facial features alone, someone's openly self-identified sexual orientation. However, the researchers take this *how-possibly* evidence and argue further that the model serves as supporting evidence for existing scientific theories, such as the PHT theory, and theories connecting facial morphology to psychological traits and processes. Wang and Kosinski say that:

> [I]dentifying links between facial features and psychological traits by employing methodology similar to the one used here could boost our understanding of the origins and nature of a broad range of psychological traits, preferences, and psychological processes ([2018], pp. 254).

However, we should be very careful here. The way to gain understanding of the actual relationship between facial features and psychological traits, and the origins of sexual-orientation, involves answering a how-actually or a pointed why question that this model, and models like it, cannot answer without resolving the requisite empirical questions and link uncertainties. In this case, the link uncertainty is vast. As the researchers themselves note, many of the features that the model tracks are cultural features, such as certain grooming patterns, and dating-profile picture conventions. Both these features have no relationship to androgens and facial morphology, thus severing the connection between the model and the phenomenon concerning the actual causes of sexual orientation**.**

[13] Several critiques by academic researchers emerged online following the release of Wang and Kosinski's paper. (e.g. Bergstrom and West [unpublished] and Mattson [unpublished]). The critiques raise several scientific drawbacks of the study that suggests that the model even falls short of answering how-possibly questions. Such models have also inspired artistic critiques concerning privacy and bias (Blas [2013]).

Furthermore, the idealized assumptions underlying the model—that sexual orientation is binary and static, that those who are openly gay on social media are representative of the whole gay population, and ignoring gender and racial variance—distort important difference makers in real-world populations (Miller [2018]). To make matters worse, existing scientific and social-scientific evidence either speaks against PHT theory, and against a dependency relation between facial features and sexual orientation or other personality traits (LeVay [1996], [2010]; Magnet [2011]; Mustanski et al. [2002]), or speaks against gender atypical traits being the driving factor (Valentova et al. [2014]). Given the problems with linking this model to actual real-world phenomena, the model is not able to provide understanding of how innate facial features reflect sexual orientation or other personality traits, let alone provide understanding of why differences in sexual orientation develop. The model, however, could be used to explain and enable understanding, even with its level of model opacity, if the surrounding scientific evidence did actually suggest that there was a link between facial features and origins of sexual orientation.

I want to stress here that the lack of understanding is not due to implementation or model illegibility. The level of implementation black boxing is the same in the deep patient case, the melanoma case, and sexual orientation case. If what I have been arguing here is right—that there is clear difference between the satisfying level of explanations and understanding we can get from each of these three models—then there is something other than implementation black boxing that governs the level of understanding the models provide. I have argued that the difference in each of these models is the level of link uncertainty (the amount, kind, and quality of scientific and empirical evidence supporting the link connecting the model to the target-phenomenon) that is present.

6 Conclusion

Are scientists trading understanding for some other epistemic or pragmatic good when they choose an opaque DNN? Not quite. DNN models are able to provide how-possibly explanations of various phenomena that, just like many minimal models, are the first steps to determining which causal mechanisms or dependency relations should be explored

further. Moreover, I have argued that so long as we do not have complete black boxing at the highest level, understanding is possible from opaque models, so long as there is an adequate link connecting the model to the phenomenon of interest. Since DNNs are not black boxed at the highest level, the central question that remains—whether a particular DNN can explain or enable understanding of the phenomenon it bears on—comes down to a question of link uncertainty.

This general claim about the importance of removing link uncertainty in order to gain understanding stretches beyond the cases of minimal and complex models. For example, Strevens ([2017]) argues that black box explanations in convergent evolution fail to provide adequate explanations or understanding. When making his argument, he specifically appeals to not knowing whether certain aspects unique to each species' evolutionary history made a difference to the evolved behaviour (and he suspects that it does). What I have been arguing here suggests that the problem is not with the black box surrounding the implementation of how a particular species evolved, but that the empirical link between the convergent evolution model and individual populations is in some sense uncertain. If the link uncertainty is resolved, our explanations of phenomena and the understanding we gain from these explanations can tolerate implementation black boxing.

Much more can be said on what precisely it takes for there to be an adequate link connecting a model to the phenomenon of interest. The cases discussed in this paper indicate that that the stronger the connection we have between the model and the phenomenon, the greater the understanding. Moreover, the cases discussed here require an empirical connection that involves going outside of the model in question. On the other hand, model-based explanations that aim to explain mathematical or structural dependences may require something other than empirical evidence to connect the model to the phenomenon of interest. I leave this question for future work.

Before ending, I want to call attention to three possible worries. First, what if there are no explanatory questions that a given DNN model can answer? Indeed, many DNN models address simple classification tasks, like identifying a number from a handwritten note. One could reasonably argue that there are no *explanatory* questions one could ask of

such a model; only mere prediction is possible.[14] Maybe so. What I have been arguing for in this paper is that the complexity and black box nature of DNN models does not prevent understanding of phenomena. There may be other reasons that many DNN models cannot provide understanding of phenomena. For instance, it might be that there are no explanatory questions the model can answer or that some models are mere predictive tools, but these are different considerations than the opacity and complexity issue one taken up here.

Second, one might worry whether the scientific evidence that connects the model to the phenomenon of interest is what constitutes our understanding such that the model itself no longer plays any epistemic role in our understanding. I have taken some steps to address this worry by highlighting the way in which the model is still necessary to explain even once we resolve link uncertainty. However, there is a deeper question about whether models, in general, are temporary tools to be discarded once we gain more empirical insight about the phenomenon. But even if models are mere temporary tools, the point made here—that getting better insight into how the model works is not necessary to gain better understanding of the phenomena it bears on—still stands.

Lastly, all the cases I have discussed here, including Schelling's model, have a level of inductive risk.[15] Diagnosing medical conditions involves high stakes. Schelling's model and facial recognition models can perpetuate harmful stereotypes and lead to greater marginalization of oppressed groups. Thus, there is a deeper question about how these risks and social values in general impact the level of explanatory understanding we can gain from these models.[16] While I focus in this paper simply on black boxes, the way that inductive and ethical risk impacts explanatory understanding deserves greater attention.

[14] That said, even in the simple handwritten number classifier, computational neuroscientists seek answers to contrastive explanatory questions between the way machines learn versus humans (Lake et al. [2015]). This type of question however is different from the types of questions discussed in this paper.

[15] For discussions on inductive risk and values in science see Douglas ([2000]), Elliott ([2013]), and Winsberg ([2012]).

[16] See Potochnik ([2015]) for a discussion of how social values in science may impact when we have explanatory understanding.

Acknowledgements

For many helpful comments and conversations, I would like to thank the anonymous referees, Mark Alfano, Henk de Regt, Insa Lawler, Michael Strevens, with special thanks to John Mumm. I presented this work at various stages of development at the University of Copenhagen, Ghent University, the 11th annual MuST conference in Turin, and the Free University of Amsterdam. I thank the attendees for the very fruitful discussions and suggestions that followed. I wrote part of this paper while being supported by a subaward agreement from the University of Connecticut with funds provided by Grant No. 58942 from John Templeton Foundation. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of UConn or John Templeton Foundation.

Emily Sullivan

Web Information Systems, EWI

Delft University of Technology

Delft, Netherlands

e.e.sullivan-mumm@tudelft.nl

References

Abelson, H., Sussman, G. J., & Sussman, J. [1996]: *Structure and Interpretation of Computer Programs.* MIT Press.

Bailer-Jones, D. M., and Bailer-Jones, C. A. [2002]: 'Modeling Data: Analogies in Neural Networks, Simulated Annealing and Genetic Algorithms', in L. Magnani and N.J. Nersessian (*eds*), *Model-Based Reasoning*, New York: Kluwer Academic/Plenum Publishers, pp. 147–65.

Baldi, P., and Sadowski, P. J. [2013]: 'Understanding Dropout', in C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (*eds*) *Proceedings of the 26th International Conference on Neural Information Processing Systems*, **Volume 2** (NIPS'13), USA: Curran Associates Inc., pp. 2814–22.

Batterman, R. W., and Rice, C. C. [2014]: 'Minimal Model Explanations, *Philosophy of Science*, **81**(3), pp. 349–76.

Bergstrom, C. and West, J., [unpublished]. 'Machine Learning about Sexual

Orientation?', available at, *Calling Bullshit: Data Reasoning in a Digital World* <https://callingbullshit.org/case_studies/case_study_ml_sexual_orientation.html>.

Blas, Z. [2013]: 'Escaping the Face: Biometric Facial Recognition and the Facial Weaponization Suite', *Media-N, Journal of the New Media Caucus*, available at <http://median.newmediacaucus.org/caa-conference-edition-2013/escaping-the-face-biometric-facial-recognition-and-the-facial-weaponization-suite/>.

Bobo, L., & Zubrinsky, C. L. [1996]: 'Attitudes on Residential Integration: Perceived Status Differences, mere in-Group Preference, or Racial Prejudice?', *Social Forces*, **74**(3), pp. 883–909.

Bokulich, A. [2008]: *Reexamining the Quantum-Classical Relation*, Cambridge: Cambridge University Press.

Buckner, C. [2018]: 'Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks', *Synthese*, **195**(12), pp. 5339–72.

Burrell, J. [2016]: 'How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms', *Big Data & Society*, **3**(1), pp. 1–12.

Clark, W. A. V. [1991]: 'Residential Preferences and Neighborhood Racial Segregation: A Test of the Schelling Segregation Model', *Demography*, **28**(1), pp. 1–19.

_____. [1992]: 'Residential Preferences and Residential Choices in a Multiethnic Context', *Demography*, **29**(3), pp. 451–66.

Cristianini, N. [2010]: 'Are we there yet?', *Neural Networks*, **23**(4), pp. 466–70.

De Regt, H. W. [2017]: *Understanding Scientific Understanding*, Oxford University Press.

Douglas, H. [2000]: 'Inductive Risk and Values in Science', *Philosophy of Science*, **67**(4), pp. 559–79.

Elliott, K. C. [2013]: 'Douglas on Values: From Indirect Roles to Multiple Goals', *Studies in History and Philosophy of Science Part A*, **44**(3), pp. 375–83.

Esteva, A., Kuprel, B., Nova R., Ko, J., Swetter, S., Blau, H., and Thrun S. [2017]: 'Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks', *Nature* **542**, pp. 115–8.

Glorot, X., Bordes, A., & Bengio, Y. [2011]: 'Deep Sparse Rectifier Neural Networks',

In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 315–323.

Goodfellow, I., Bengio, Y., & Courville, A. [2016]: *Deep Learning*, MIT Press.

Grimm, S. [2010]: 'The Goal of Explanation', *Studies in History and Philosophy of Science A*, **41**(4), pp. 337–44.

_____. [2014]: 'Understanding as Knowledge of Causes.', In A. Fairweather (*ed*), *Virtue Epistemology Naturalized: Bridges Between Virtue Epistemology and Philosophy of Science*, Dordrecht: Springer, pp. 329–45.

Grüne-Yanoff, T. [2009]: 'Learning from Minimal Economic Models', *Erkenntnis*, **70**(1), 81–99.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. [2016]: 'Deep Learning for Visual Understanding: A Review', *Neurocomputing*, **187**, pp. 27–48.

Hills, A. [2016]: 'Understanding Why', *Noûs*, **50**(4), pp. 661–88.

Humphreys, P. [2004]: *Extending ourselves: Computational Science, Empiricism, and Scientific Method*, Oxford University Press.

_____. [2009]: 'The Philosophical Novelty of Computer Simulation Methods', *Synthese*, **169**(3), pp. 615–26.

Khalifa, K. [2017]: *Understanding, Explanation, and Scientific knowledge*, Cambridge University Press.

Knight, W. [2017]: 'The Dark Secret at the Heart of AI', *MIT Technology Review*, available at <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/>.

Kuorikoski, J. and Ylikoski P. [2015]: 'External Representations and Scientific Understanding', *Synthese* **192**(12), pp. 3817–37.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. [2015]: 'Human-Level Concept Learning through Probabilistic Program Induction', *Science*, **350**(6266), 1332–8.

Lawler, I. [2018]: 'Understanding Why, Knowing Why, and Cognitive Achievements', *Synthese*, pp. 1–21.

LeCun, Y., Bengio, Y., & Hinton, G. [2015]: 'Deep Learning', *Nature*, **521**(7553), pp. 436.

LeVay, S. [1996]: *Queer science: The Use and Abuse of Research into Homosexuality*, MIT Press.

Lipton, P. [2009]: ‘Understanding Without Explanation’, in H. W. de Regt, S. Leonelli, & K. Eigner (*eds*), *Scientific Understanding: Philosophical Perspectives*, Pittsburgh: University of Pittsburgh Press, pp. 43–63.

Mäki, U. [2009]: ‘MISSing the World. Models as Isolations and Credible Surrogate Systems’, *Erkenntnis*, **70**(1), 29–43.

Mattson, G. [unpublished]: ‘Artificial intelligence discovers gay face. Sigh.’, *Scatterplot* available at, <https://scatter.wordpress.com/2017/09/10/guest-post-artificial-intelligence-discovers-gayface-sigh>

Miller, A. E. [2018]: ‘Searching for Gaydar: Blind Spots in the Study of Sexual Orientation Perception’, *Psychology & Sexuality*, **9**(3), pp. 188-203.

Mehta, P., & Schwab, D. J. [unpublished]: ‘An Exact Mapping Between the Variational Renormalization Group and Deep Learning.’, available at, <arxiv.org/abs/1410.3831>

Mitchell, T. M. [1997]: *Machine learning*. Burr Ridge, IL: McGraw Hill.

Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. [2016]: ‘Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records’, *Scientific Reports*, **6**(26094), pp. 1-10.

Muldoon, R., Smith, T., & Weisberg, M. [2012]: ‘Segregation that No One Seeks’, *Philosophy of Science*, **79**(1), pp. 38–62.

Mustanski, B. S., Chivers, M. L., & Bailey, J. M. [2002]: ‘A Critical Review of Recent Biological Research on Human Sexual Orientation’, *Annual Review of Sex Research*, **13**(1), pp. 89–140.

Potochnik, A. [2011]: ‘Explanation and Understanding’, *European Journal for Philosophy of Science*, **1**(1), pp. 29–38.

_____. [2015]: ‘The Diverse Aims of Science’, *Studies in History and Philosophy of Science Part A*, **53**, pp. 71–80.

Reutlinger, A., Hangleiter, D., & Hartmann, S. [2017]: ‘Understanding (with) Toy Models’, *The British Journal for the Philosophy of Science*, **69**(4), pp. 1069-99.

Ribeiro, M. T., Singh, S., & Guestrin, C. [2016]: ‘Why should I Trust You?

Explaining the Predictions of Any Classifier', In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135-44.

Rohwer, Y., & Rice, C. [2013]: 'Hypothetical Pattern Idealization and Explanatory Models', *Philosophy of Science*, **80**(3), pp. 334–55.

Schelling, T. C. [1971]: 'Dynamic Models of Segregation', *The Journal of Mathematical Sociology*, **1**(2), pp. 143–86.

Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. [2018]: 'Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record Analysis', *IEEE journal of biomedical and health informatics*, **22**(5), pp. 1589–604.

Strevens, M. [2008]: *Depth: An Account of Scientific Explanation*, Cambridge: Harvard University Press.

_____. [2017]: 'The Whole Story', in D. Kaplan (*ed*) *Explanation and Integration in Mind and Brain Science*, Oxford University Press.

Sullivan, E. [2018]: 'Understanding: Not Know-How', *Philosophical Studies*, **175**(1), 221–40.

Sullivan, E. and Khalifa, K. [2019]: 'Idealization and Understanding: Much ado about nothing?', *Australasian Journal of Philosophy*, pp. 1–17.

N. Tishby, F. C. Pereira, and W. Bialek, [1999]: 'The information bottleneck method', in *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–77.

Tishby, N., & Zaslavsky, N. [2015]: 'Deep Learning and the Information Bottleneck Principle', in *IEEE Information Theory Workshop* (ITW), pp. 1-5.

Valentova, J. V., Kleisner, K., Havlíček, J., & Neustupa, J. [2014]: 'Shape Differences between the Faces of Homosexual and Heterosexual Men', *Archives of Sexual Behavior*, **43**(2), pp. 353–61.

van Riel, R. [2015]: 'The Content of Model-Based Information', *Synthese*, **192**(12), 3839–58.

Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., & Fergus, R. [2013]: Regularization of Neural

Networks Using DropConnect', in *Proceedings of the 30th International Conference on Machine Learning* PMLR **28**(3), pp. 1058–66.

Wang, Y. and Kosinski, M. [2018]: 'Deep neural networks are more accurate than humans in detecting sexual orientation from facial images', *Journal of Personality and Social Psychology* **114**(2), pp. 246–57.

Winsberg, E. [2012]: 'Values and uncertainties in the predictions of global climate models.,' *Kennedy Institute of Ethics Journal*, **22**(2), pp. 111–37.