Understanding from Deep Learning Models in Context

by Michael Tamir and Elay Shech

Forthcoming in Lawler, I., Khalifa, K. and E. Shech (Eds.) 2023. *Scientific Understanding and Representation Modeling in the Physical Sciences*. Routledge.

This paper places into context how the term *model* in machine learning (ML) contrasts with traditional usages of scientific models for understanding and we show how direct analysis of an estimator's learned transformations (specifically, the hidden layers of a deep learning model) can improve understanding of the target phenomenon and reveal how the model organizes relevant information. Specifically, three modes of understanding will be identified, the difference between implementation irrelevance and functionally approximate irrelevance will be disambiguated, and how this distinction impacts potential understanding with these models will be explored. Additionally, by distinguishing between empirical link failures from representational ones, an ambiguity in the concept of link uncertainty will be addressed thus clarifying the role played by scientific background knowledge in enabling understanding with ML.

1. Introduction

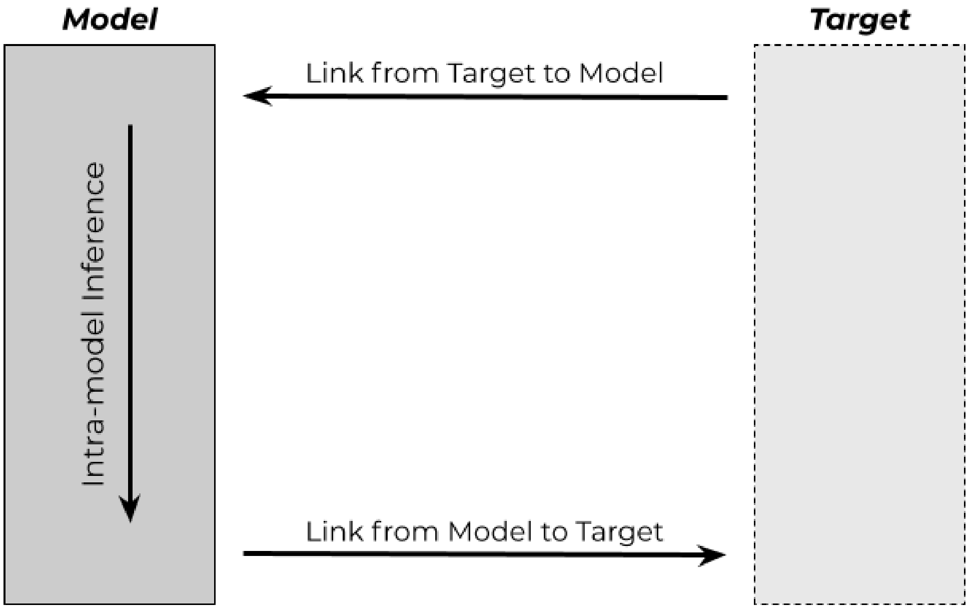
Advances in machine learning (ML) techniques, deep learning (DL) especially, are drawing increased philosophical consideration. ML-trained algorithms are often called models, encouraging questions about how such automated effective estimation techniques fit in with existing accounts of scientific modeling and representation for understanding. In contrast to how simple idealized models arguably enable understanding by reducing complexity (Bokulich 2008; Khalifa 2017; Potochnik 2017; Strevens 2008), Sullivan (2022) rejects the possible claim that ML models enable understanding by reducing complexity. In particular, Sullivan (2022, 110) claims that “model simplicity and transparency are not needed for understanding phenomena,” arguing that DL models can provide understanding despite the ostensible opaqueness or “blackbox” nature of how particular estimations are generated. Instead, she suggests that understanding with a DL model depends on what she calls *link uncertainty*, or “the extent to which the model fails to be empirically supported and adequately linked to the target phenomena.”

In this paper, we place into context how the term *model* in ML contrasts with traditional usages of scientific models for understanding, resolving core ambiguities involving representational links to the target phenomenon. We explore standard techniques involving direct analysis of an estimator’s learned transformations (viz., the hidden layers of a DL model). Next, we show that such direct analysis can improve understanding of the target phenomenon and reveal how the model organizes relevant information. In Section 2, we lay the groundwork for contrasting non-ML scientific models with ML models, and set the stage for our interaction with Sullivan’s take on understanding with ML models. Section 3 provides a brief overview of ML and DL models and considers a candidate for framing the proper target of ML understanding leveraged in later sections. Section 4 identifies three modes of understanding given the proposed target of ML models. We then disambiguate what we describe as the difference between implementation irrelevance and functionally approximate irrelevance, and explore how this distinction impacts potential understanding with these models. Section 5 addresses an ambiguity in the concept of link uncertainty, arguing that distinguishing empirical link failures from representational ones clarifies the role played by scientific background knowledge in enabling understanding with ML. In Section 6, we conclude with a brief summary.

2. C-schema Models and Sullivan on Understanding from ML Models

In one of her paradigm examples of non-ML models, Sullivan (2022) considers Thomas Schelling’s model for explaining and understanding why human populations tend to be segregated. Schelling wanted to investigate “some of the individual incentives and individual perceptions of difference that can lead collectively to segregation” (Schelling 1971, 138). This means that segregation behavior among populations is the *target phenomenon* of interest, that is, the object of study that we want to understand. He constructed a simple model: A checkerboard represents spatial locations with two types of individual households, represented by dimes and nickels. Each household, or "actor," is stipulated to prefer that at least 30% of its neighbors be of the same type (similarity preference parameter). If this condition is met, the actor remains in place; if not, one moves the actor to the closest unoccupied space. It turns out that for most initial configurations, the equilibrium state of the board results in segregation. Such results may suggest that racial segregation can occur without an organized institutionally racist influence. But how do models like Schelling's afford understanding? Sullivan (2022, 3) adopts the view (found in, e.g., Khalifa 2017; Strevens 2008) that explanation aims at understanding whereby "explaining why helps us understand why." Schelling's model succeeds in affording understanding of actual segregation behavior found in some populations if it links faithfully with causal factors that explain such segregation. Similarly, Schelling's model fails at providing understanding and explanations of actual populations when, for instance, the similarity preference parameter is inaccurate.

It is helpful to view a paradigmatic non-ML model like Schelling's model, along with the explanations and understanding that it may provide, within a framework that we will call the "C"-shaped modeling schema, or *C-schema*.



There are many accounts of scientific modeling, representation, and inference that fit the C-schema. For example, similarity accounts (e.g., Weisberg 2013), mapping isomorphism or structuralist accounts (e.g., da Costa and French 2003), inferentialist accounts (e.g., Suárez 2004), and more recent views (e.g., Frigg and Nguyen 2016) can all be viewed within a C-schema. Such models M are used or interpreted to explain, understand, or investigate some target T, which is a phenomenon in the world, with specified audiences and goals in mind (e.g., epistemic or pragmatic). The horizontal top of the C-schema concerns the modeling or representational relationship between M and T. For example, some accounts note that a scientist must stipulate that M denotes T (e.g., Frigg and Nguyen's (2016) DEKI account). On the vertical side of the C-schema, one interacts with M directly to draw inferences, for example, by running iterations of Schelling's model and identifying specific relationships between similarity preference parameters and state equilibria, or by performing mathematical calculations given certain conditions. Last, on the horizontal bottom side, M is used to relevantly draw inferences and answer questions about, or impute properties to, T.

Models that fit into the C-schema can range in kinds such as concrete models like the San Francisco Bay-Delta model, mathematical models such as the Lotka-Volterra predator-prey model, or simulation/computational models like Schelling's model (Weisberg 2013, Ch. 2). Such models tend to facilitate inferences about systems and phenomena of interest, limited to the scope of representation, and afford understanding in a number of ways. For instance, in a typical map (or map-like) model, we can not only use M to draw inferences about T but we can also see how such inferences are drawn, for example, because of relevant similarities or isomorphic relations associated with M and T (say, a map and a city). There are many varied accounts as to how the C-schema works and facilitates successful inference, such as those accounts mentioned above. In one prominent family of C-schema accounts, to which we return, understanding is powered by some relevant structural similarity or isomorphism between M and an abstraction of features of T.

In what follows, we will focus on two of Sullivan's main claims. The first is the claim that opacity and complexity of ML models that can occur at various levels is not an in-principle impediment to understanding target phenomena. Models do not have to be transparent for them to be useful for understanding and explanation. For example, Sullivan notes how one can compute factorials using an iterative process or a recursive process, but that such implementation differences are irrelevant to understanding with, for example, a climate model that uses factorials. Analogously, "one does not need to know whether Schelling's model was implemented using a functional, object-oriented, or actor-based language" and so "implementation blackboxing in itself does not undermine our ability to explain or understand phenomena" (Sullivan 2022, 114–115). We agree with Sullivan on this point—implementation opacity does not matter anymore than drawing a map with red or blue ink matters—but we think that there is an important overlooked distinction between this type of implementation irrelevance and (what we call) functionally approximate irrelevance.

Second, Sullivan holds that a fundamental impediment to understanding in the context of both ML and non-ML models concerns the link between model and target (horizontal sides of the C-schema), what she calls "link uncertainty," which "constitutes a lack of scientific and empirical evidence supporting the link connecting the model to the target phenomenon" (2022,124). For instance, in discussing Schelling's model, she notes that without "empirical evidence validating that the possible causes identified by Schelling's model are actual causes, there is no link connecting the model to the phenomenon" (2022,125). Sullivan (2022, 128) then continues to argue for the same point in the context of ML models, holding that "lack of understanding is not due to implementation or model illegibility." Instead, it is due to high "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." Generally, we are in agreement but we will highlight an important contrast between empirical link uncertainty (e.g., found in Schelling's model due to an inaccurate similarity preference parameter, and other potential empirical questions concerning actual populations) and what we describe as data-misrepresentation link uncertainty.

3. The Target of ML Models

ML classification[1] is an algorithmic process for generating an estimator $f : x_i \mapsto p_i$ that, for given "input elements" $\{x_i\}_{i \in N} \subset \Omega_X$ in a data set, estimates a scoring function p_i defined over the set of "output elements" or *y-targets*[2] $\hat{y}_i \in \Omega_Y$. Typically, f is parameterized by some set of parameters $\theta \in \Theta$, establishing a family of estimators $\{f_\theta\}_{\theta \in \Theta}$ for the given estimation procedure. Training an ML estimator f can be distinguished from non-ML rule-based estimation algorithms in that the parameter values θ determining the trained f_θ are optimized so as to best "fit the data" according to a prescribed loss function. Namely, the parameters defining the estimator f_θ are optimized by taking a set of sample pairs $\{(x_i, y_i)\}_{i=0}^{N} \in \Omega_X \times \Omega_Y$,[3] called the *training set*, and, for a given loss function $L : (p_i, y_i) \to l_i \in \mathbb{R}$, finding the optimal parameterization θ^* such that:

$$\theta^* = argmin_{\theta \in \Theta} \sum_{i=0}^{N} L(f_\theta(x_i), y_i)$$

The process of finding the optimal f_{θ^*} for a given training set is called *training the model.*

[1] In the context of classification, scoring functions p_i are commonly probability distributions on the state space $\mathbf{\Omega}_Y$. In the special case where a single output is desired, p_i may also be an indicator function returning 1 for a single element $\hat{y}_i \in \mathbf{\Omega}_Y$ and 0 for other elements.

[2] The scored "output elements" are commonly referred to as the *target* or *target feature* of an ML model, which in our present contexts risks a misleading conflation of y-targets with the target phenomena of understanding of models as in the C-schema. To disambiguate, we shall use 'y-' to prefix 'y-targets' whenever the former is intended.

[3] In the context of sampling, $\mathbf{\Omega}_X$ and $\mathbf{\Omega}_Y$ can here be thought of as state spaces for (marginalized) random variables *X* and *Y* respectively, which are described by some joint distribution corresponding to the data sampling process.

DL is a family of ML techniques in which estimators are constructed through the composition of multiple linear and nonlinear transformations, called a *neural network*. Neural networks can consist of multiple (hidden) layers of learned parameterized transformations of the data, where each *hidden layer* (parameterized transformation) is learned through ML optimization. We argue below that while DL models do not necessarily provide understanding in the same manner as (say) map representations, learned representation layers may be leveraged to improve understanding by providing insight into how a DL model learns to organize raw input data to optimally estimate y-targets.

As an example, consider an ML model trained to solve "complete the sentence" tasks on text sampled from a given corpus. An x_i might be 'In the morning, I enjoy _____' where the answer y_i for a particular data point is 'coffee.' An ML model f_θ is trained to best estimate how to fill in this blank given the x_i input. Other (similar) data points may have the same input for $x_{i'}$ but a different $y_{i'}$. Underlying how English tends to be written (for our samples) is some distribution over the entire vocabulary; 'coffee' may be likely, but terms like 'tea,' 'eggs,' 'sunshine,' etc., also get non-negligible probability mass, whereas other arbitrary terms like 'economic,' 'about,' or 'transcendental' do not. We can imagine there is some "true" distribution $p(Y = y_i | X = x_i)$ underlying the conditional probability for the (x_i, y_i) data generated by a given sampling process. An ML algorithm learns from actually sampled examples in the training set to best estimate this distribution. That is, the result of an ML training process is an estimator f_θ for random variable Y given X, such that the model can estimate conditional probabilities $p_\theta(Y = y | X = x_i)$ induced by the scoring function f_θ for given x_i values.[4] If, for appropriate metrics, the induced $p_\theta(Y|X)$ matches the actually sampled y_i values well on data not used for training, called *test data*, we can start making judgments about how reliably it can relate such data more generally, but, strictly speaking, we are still just talking about the model and its (potential) data.

[4] In practice, the standard construction for DL classifiers includes a final "softmax" (multinomial logistic) layer estimating *p(Y|X)*, by reducing cross-entropy between the estimated distribution of scores over all *y* values and the actually sampled y-target in the data ensuring that *p(Y|X)* satisfies the formal constraints of probability distributions and that (with enough data) learns to assign probability mass to non-peak values. The "hard" max of this learned distribution is returned when a single $\hat{y}$ value must be returned.

We can view this within a generalized C-schema framework: Inferences about how well the model can estimate y-targets given x-data, exploration of what particular features of x-data tend to play important roles in (correct) estimations of y-targets, and the study of how network parameters and hidden layer transformations organize and restructure x-data to effectively estimate y-targets are all examples of "vertical" inferences in a C-schema. As Sullivan (2022) observes, we need to "link" to the external phenomena intended to be understood. What is missing are the *horizontal* links of a C-schema connecting facts about the model and the data considered in isolation of a target phenomenon.

In order to horizontally link an ML model to a target of understanding (and hence evaluate the links' level of uncertainty which Sulivan correctly argues is vital), we must be specific about the target. Our hypothesis is that the appropriate target of understanding with ML models closely relates to how a learned distribution $p_\theta(Y|X)$ estimates the "actual" distribution $p(Y|X)$ describing a phenomenon's studied features. Specifically we propose the following:

> **(TML) Target of ML Hypothesis:** The target phenomenon of understanding with ML models is the *relationship(s) of features represented by the data*.

The *target phenomenon* is not a particular object or sampling instance, but relationships of the properties or features found potentially in individual or multiple objects or object types.[5] The target is the relationship patterns of these represented features. These relationship(s) are typically those described (indirectly or directly) by some underlying actual distribution $p(Y|X)$ estimated by the model. While the targeted relationships are described by such a $p(Y|X)$, we emphasize that it is the real world relationships between features implied by the description $p(Y|X)$ that are the target phenomenon not $p(Y|X)$ itself. The concept of *represented features* intuitively can be thought of as measurable properties associated with the phenomenon,[6] but as we elaborate in Section 5 the specific features represented by data are intimately tied to measurement

[5] We use the phenomena/data distinction in a similar manner to Bogen and Woodward (1988).

[6] Note that in practice data scientists and ML researchers refer to each of the dimensions represented by x-data as a *feature*, and the y-target as the *label*, or the (y-)*target feature*, in the context of supervised learning. To be explicit, for our usage here both *x* and *y* data represent features of a target phenomenon.

methodology and sampling practice. We argue that focusing on precisely what features sampled data do *and do not* represent is paramount to evaluating ML model link uncertainty.

4. Implementation vs. Functionally Approximate Irrelevance

Relationships of features in the TML hypothesis play an important role for interpreting the target of ML understanding. Assume that with proper methodology one rules out the sort of link uncertainty (discussed in Section 5 below) associated with how data represents features of the target phenomenon (horizontal top side in the C-schema). Assume further that an ML model is well trained with appropriate metrics, data, and research practices for testing generalizability to support the claim that $p_\theta(Y|X)$ reliably estimates similarly sampled data (vertical side in the C-schema). In order to study trained ML models, especially complex DL models, for insight into external targets, we must first clarify how a link from ML models to TML targets (horizontal bottom side of a C-schema) may work.

To frame this challenge, let us contrast understanding from ML models and their targets with examples of understanding from subway maps and their targets. In the latter case, relevant facts about the target are abstracted and then represented typically as some graph-like visualization, a map. The map can then be used to understand how to navigate the represented subway system so long as these navigation insights are circumscribed by what is faithfully supported through the captured facts. For instance, one can infer which paths are available from point A to point B, but not (always) details about the decor or physical distance associated with taking the paths. The map represents an abstraction of the right details, enabling judgments based directly on the map's topology to be linked back to things like understanding how to navigate the represented subway system.

There are multiple details about the map that are also irrelevant to understanding navigation. For instance, color choices, whether it is physical or digital, and so on, are irrelevant details that fall under what Sullivan (2022) describes as "implementation" details. However, topological facts depicted by a map are relevant and cannot be ruled out as mere "implementation" details of the visualization, ostensibly because one must use those details directly to generate judgments like "this path from A leads to B." What about complex DL trained neural networks? After all, an ML model's estimations of $p_\theta(Y|X)$ depend directly on the parameter values θ and network architecture. Sullivan (2020, 7-8) argues that when the "[d]etails regarding the implementation are unnecessary for explaining and understanding," opacity of these details ("implementation blackboxing") are "not in principle problematic for explaining or understanding phenomena." Sullivan's argument refutes that DL models cannot be used for understanding merely due to *some* opacity, but the reliance on the *implementation irrelevance* of certain details, particularly their irrelevance to the target of understanding, is essential for this defense.

We have a dilemma: To understand feature relationships with an ML model despite detail opacity with respect to learned parameter instantiations, said details ostensibly cannot be used for insight into the relationship, but if estimations of $p_\theta(Y|X)$ directly depend on the instantiation of these parameter values how can they be used to understand the relationship between X and Y without such details? Parameter instantiation details are not purely irrelevant implementation details like coding language, color choices, material constitution, or even provably equivalent algorithmic techniques. The dilemma resolves by disambiguating implementation irrelevance, where the variations in question make no difference to the target of understanding, from a second sort of *functionally approximate irrelevance*, active in the case of DL understanding. What distinguishes functionally approximate irrelevance from implementation irrelevance is that in the former *varied details matter to the studied target*, but they are varied only in ways that approximately preserve the relevant aspects of the phenomenon to be understood.[7]

When the target is well specified by the TML hypothesis, we argue that by also accounting for the role and degree of functionally approximate irrelevance, DL may help us understand the following aspects and relationships of features represented by the data:[8]

IR (Informative Relationship): If and to what extent mutual information between the features of a target phenomenon exists.

[7] Sullivan's view may be that what we distinguish as functionally approximate irrelevance and implementation irrelevance here both fall under some more general concept of implementation irrelevance. Our account may be interpreted as a refinement of Sullivan, making explicit that accounting for role and degree of approximation matters to understanding.

[8] We make no claims that this list is comprehensive.

FI (Feature Importance): Which features associated with the input x-data either individually or in combination are more/less important to such an informative relationship.

LR (Learned Representation): How transforming the input x-data to better enable estimation of the y-target reveals informative ways of organizing the x-data and the features it represents for this or other estimation purposes.

To illustrate how functionally approximate irrelevance plays a role, consider the sentence completion task discussed in Section 3. Using the same neural network architectures and the same (or similar sampling of) training data we could train two different models. Random differences in how the parameters are initialized before training, order differences in how the models see the training data, or differences in hyper-parameters used to define how the models are trained can result in two substantively different learned parameterizations θ and θ', even though the probability estimates $p_\theta(Y|X)$ and $p_{\theta'}(Y|X)$ generated by the respective DL models approximate the feature relationships described by $p(Y|X)$ equivalently. Sullivan defines "highest level" ML opacity (blackboxing) as cases where one merely has access to model inputs and outputs but not execution details. IR is compatible with Sullivan's highest level opacity. If both models are successfully trained, performing sufficiently and equivalently across different strata of the data, we can infer that since $p_\theta(Y|X)$ and $p_{\theta'}(Y|X)$ both approximate $p(Y|X)$, which in turn describes features of the target phenomenon, significant parameter detail differences (i.e. $||\theta - \theta'|| \geq \epsilon$) are irrelevant to understanding that there is an informative relationship. With the ML model's target clearly defined as the relationships of features represented by X and Y, the fact that $||\theta - \theta'|| \geq \epsilon$ has approximate irrelevance becomes clear: $p_\theta(Y|X)$ approximates $p(Y|X)$ well enough, so while the details matter to *how the estimates are made*, they can still be (approximately) irrelevant for the target of understanding. With the functionally approximate irrelevance of particular parameter details ensured, IR understanding even with "highest level" opacity is possible. At minimum, the reliable generalizability of an ML model to similarly sampled data increases our understanding that there is *some* signal in the x-data useful for y-target estimation. The existence of even an opaque but sufficiently reliable estimator $p_\theta(Y|X)$ can entail that there is mutual information between the features represented by X and Y.

Turning to FI, assigning feature importance in DL is an active research area. Certain "permutation" style techniques treat the DL model with highest-level opacity where x-features are manipulated (e.g., occluding part of an image as in (Seiler and Ferguson 2014)) and the impact on estimation performance is then analyzed. Other methods attribute importance from internal network properties, tracing individual contributions (typically involving gradients) from neuron to neuron (Shrikumar et al. 2017; Simonyan et al. 2013). Arguments for FI-based understanding for either permutation methods, or methods that meticulously trace contributions through the network are similar to the arguments for IR above. If two similarly trained models irreconcilably diverge in their feature importance implications, approximate irrelevance comes into question and the researcher should doubt whether such FI attribution yields very much genuine understanding of which features are predictive of the target. In contrast, if two similarly trained models tend to agree on which features are important, or better yet, multiple feature attribution techniques agree on which features matter to model predictions for certain sorts of input, then the functionally approximate irrelevance of the specific layer-by-layer calculations is evident, supporting FI understanding of the phenomenon. Again, the details matter in the sense that changing them has a direct impact; however, by establishing that certain detail variations approximately preserve resulting insights into the target phenomenon's feature relationships (viz. FI relationships), understanding is possible.

In their influential discussion of representation learning, Bengio et al. (2013) describe how DL models must "learn to identify and disentangle the underlying explanatory factors hidden in the observed milieu of low-level sensory data." Research into how DL models "disentangle" and organize the "low-level" input data focuses not only on studying the informative relationship between input and output substantiated by the possibility of training a DL model to detect these relationships successfully (IR), but also on how the data are represented via transformations from one hidden layer to the next hidden layer of the network (LR), revealing what kind of information is preserved and how it is represented in the associated vector (tensor) spaces of these layers.[9] Sullivan (2022, 119) echoes Bengio et al., describing how DL models learn to “tease out the relevant features from the irrelevant,” as in image classification where each hidden “layer gradually picks out higher and higher-level abstractions until it reaches a classification of the image.” How hidden layer representations interact and combine for the ultimate classification goal is studied not just for a deeper understanding of the neural networks themselves, but also to understand how they organize and incorporate relevant information content in the “low-level sensory data” as they mathematically transform it into new representations for optimal estimation (Olah et al. 2020).

A simple example of such informative hidden layer representations is traditional word vectorization.[10] The early efforts of Mikolov et al. (2013) used shallow neural networks to map individual terms to vector representations, which were optimized for “fill in the blank” style tasks similar to the example in Section 3. Such vectorized *word-embeddings* are not merely useful for the original task. The representations could be reused as *pre-trained* representations for novel text-based tasks.[11] Embeddings were widely used to study and leverage ostensible semantic relationships (e.g., analogies, synonymy clusters) manifested by their usage patterns for practical applications.

In contrast to IR and FI, LR must engage *directly* with the learned representations of data (like word-embeddings) associated with hidden layers or neurons. As above, the individual parameterizations of the learned representations are opaque, but in the case of LR, understanding is achieved through direct engagement with these representations.[12] The ostensible target phenomenon of the original Mikolov et al. (2013) task is the relationship between the terms

[9] See, for example, (Kingma and Welling 2013; Chen et al. 2018; Achille and Soatto 2018; Olah et al. 2020; and Tamir and Shech 2022), for a discussion of the various DL and information theoretic techniques for studying such representations.

[10] More complex contemporary Transformer techniques use much deeper pre-training of text embedding methods bearing some similarities to shallower early word vector pre-training representations but have been adapted to embed both chunks of text and embed individual terms in the context of the surrounding text in which they are written (Vaswani et al 2017; Devlin et al. 2018).

[11] Because information about frequency in contexts learned by these models can be useful in more general text-based tasks, these representations and more advanced techniques are commonly also used as pre-trained representatives for new models. Such pre-training is standard for both novel text tasks and image classification tasks. As we see in Section 5, both Esteva et al. (2017) and Haenssle et al. (2018) use versions of pre-trained Inception models (v3 and v4 respectively) for image classification.

[12] There are numerous methods of visualizing and analyzing learned representations directly, including projecting to a lower-dimensional space, inspecting particular weights indicating strength of importance to the task, developing generative models, sweeps in latent feature spaces (Chen et al. 2018; Kingma and Welling 2013), and more.

"filled in" (y-targets) and their surrounding context terms (x-data). This relationship is described by some distribution capturing (some of) this frequency in context information. Study of these vectorizations, such as the relative position of vector differences can capture analogical ($\vec{man} - \vec{king}$ vs $\vec{woman} - \vec{queen}$) or morphological ($\vec{smart} - \vec{smarter}$ vs $\vec{hard} - \vec{harder}$) term usage in (sampled) text.[13] Similarly, projecting word-embeddings onto lower dimensional visualizations, or studying clustering patterns of embeddings can inform understanding of term usage as a surrogate for synonymy.[14] As with the discussion of IR, different embeddings likely have non-identical parameterizations due to differences in the training data, the way that the models were initialized, neural architecture, training process, etc. However, if these representations can be used for LR understanding of the represented features, properties such as the relative positions of word-embeddings used to complete analogies should be evident in the respective representations despite these differences. Dev et al. (2019) explain that "rotation or scaling of the entire dataset will not affect synonyms (nearest neighbors), linear substructures (dot products), analogies, or linear classifiers" because "there is nothing extrinsic about any of these properties." For example, in studying the impact of basis rotations to align GloVe (Pennington et al. 2014) and Word2Vec (Mikolov et al. 2013) embeddings, they confirm that using a vector from Word2Vec to complete an analogy using GloVe embeddings "is very poor, close to 0; that is, extrinsically there is very little information carried over" by the (basis dependent) parameter values themselves. However, when the learned rotations were used to align the embeddings first, near equivalent performance was recovered.

Studies like these illustrate how individual parameterization details can differ but still have functionally approximate irrelevance to the specific method of studying the properties of the DL model. In the simple case of word-embeddings, we see that properties such as relative angles or positions that are invariant to certain changes of coordinate values allow for a direct engagement with hidden layer representations to gain LR understanding.[15] This suggests a path

[13] When learned representations are particularly effective as pre-trained representations, it is tempting for additional applications to infer that they capture relationships about language (or vision, etc.) in general. However, there are epistemic risks associated with adopting patterns in the data on which a pre-trained model was trained, leading, for example, to gender bias in word-vectorizations (Bolukbasi et al. 2016).

[14] Esteva et al. (2020) discussed below does a similar low dimensional projection of learned representation vectors of dermatology photos to study how clustering patterns of these images relates to their respective diagnosis labels.

[15] For a more complex example, see Olah et al.'s (2020) universality hypothesis.

to reconciling Sullivan’s contrast of DL models with idealized models. Namely, although DL models have an overwhelming “blackbox level” number of parameter details, focusing on how scientists and researchers leverage these models for LR (FI and MI) understanding reveals the approximate irrelevance of these particular details. By attending to which details are (approximately) irrelevant to these more illuminating relationships and properties, we can see how such opacity (at an approximately irrelevant individual parameter detail level) need not prohibit improved understanding.

5. Disambiguating Link Uncertainty

Our above discussion explored the TML hypothesis that understanding with ML models targets the relationships of features represented by the data and described by some underlying distribution directly or indirectly estimated by the model. Further, if varying certain “blackbox” details impact the target relationships of an ML model only approximately, then DL understanding is still possible at least for MI, FI, and even LR when we “open up the black box” to acquire this understanding. In this section, we consider examples discussed by Sullivan to disambiguate two kinds of link uncertainty that we submit are conflated in Sullivan (2022) so that the understanding gained from FI and LR can potentially be leveraged to prevent such uncertainty.

Sullivan (2022, 126) presents Esteva et al.’s (2017) melanoma classifier as an example of low link uncertainty. Esteva et al. (2017) use the Inception-v3 (Szegedy et al. 2016) model and then fine-tune (further train) it for dermatology images. The authors tested their model using data points with biopsy verified labels, inferring that it could outperform most expert dermatologists evaluated on “the same data.” What links justify the conclusion that their model was able to outperform expert dermatologists tested on “the same data?” First, although the model’s x-data are processed from images that the dermatologists looked at, even these data are not identical: Inception-v3 takes as input 2-dimensional arrays of RGB values, whereas human dermatologists view images. Issues of proper lighting at test time, etc. might affect human performance but not the model. Further, issues of data leakage (Kuehlkamp et al. 2017; McCoy et al. 2019; Torralba and Efros 2011), where unintended signal correlated to the y-target is inadvertently left in the x-data, pose a risk to the validity of this horizontal link. For instance, Haenssle et al. (2018) reported that their DL trained algorithm with the similar Inception-v4 convolutional neural network (CNN) architecture performed competitively with expert dermatologists. However, in a follow-up study, Winkler et al. (2019) note that "[i]n clinical routine, suspicious lesions are frequently marked before being excised or photographed" and that benign images were "frequently labeled as being malignant by the CNN when ink markers were visible at the periphery of the dermoscopic image." In subsequent tests of the CNN on new data taken before and after marking, they found that adding the marking significantly increased false positivity, suggesting that human markings made before the original study leaked unintended human expert information about the y-target into the x-data.[16] Since ML models primarily "learn" from the provided data, ruling out data leakage and confounding bias in sampling methodology, data preparation, and y-target labeling are fundamental to preventing link uncertainty with ML models. In data leakage cases, the actual features represented by the data were not just the intended features (specifically, how skin looks (x-data) and dermatological state (y-target)). The data also included influential features, namely, experts selectively added markings in a way that correlated with suspicions of dermatological states. Conclusions drawn from a successfully trained DL model are about these features (also), because that is what the actually sampled data represent, rendering the link with intended features, namely, unmarked skin and their dermatological state uncertain. Let us call this kind of link uncertainty, resulting from relevant and unintended misrepresentation of target features by the data, *misrepresentation uncertainty*.

Misrepresentation uncertainty can clearly corrupt the horizontal links between the model and target (in both directions), confounding the intended features of the target phenomenon to be understood. This is different from the kind of empirical link uncertainty introduced in Section 2 that occurred with Schelling's model. In that case, the uncertainty did not arise from an *unintended* mismatch between what elements of the simulation model (coins as residents, squares in the grid as houses) are supposed to represent. Rather, the alleged uncertainty

[16] Esteva et al. (2017) have a similar potential source of data bias where dermatologists selectively used a ruler when capturing images of lesions, one of the authors is quoted in popular media as recognizing that "[i]n our data set, dermatologists tended to do this only for lesions that were a cause for concern" (Patel 2017).

concerned a mismatch between empirical facts about homeowner reactions and intended model parameters (namely, the 30% similarity preference threshold), and it is not clear that these model assumptions veridically correspond to the target. In misrepresentation uncertainty such as leakage cases, however, the data still veridically correspond to *some feature* actually measured and encoded in the data, but substantively *not the intended features* to be understood.

Sullivan (2022, 126) argues that "[i]mplementation 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." She highlights that "[t]he level of scientific justification and background knowledge linking the appearance of moles to instances of melanoma is extensive," noting that it is a "leading deciding factor for medical intervention" and biopsies. This supports claims that there may be meaningful relationships between features of the phenomenon, but it does not provide clear horizontal links for understanding. Certainly, background scientific knowledge can inform the kinds of features to target with an ML model, but in order to establish a link between the model and the target, more is needed. Disambiguating misrepresentation link uncertainty from empirical link uncertainty helps clarify what is needed. The data used by an ML model must *represent* the targeted features of the phenomenon as intended. If not, a background of scientific evidence does not prevent misrepresentation uncertainty. Haenssle et al. (2018) rely on the same "background knowledge" for a similar use case and even DL model but demonstrably fail to link the model to the target. Complementary to the important role of background scientific knowledge for informing how ML models link with their target phenomenon, misrepresentation uncertainty also highlights that appropriate sampling methodology and data preparation are necessary to establish a representation link between the ML model's data and the intended features of the target phenomenon.

Moreover, FI and LR, discussed in Section 4, can be instrumental not just as examples of understanding, but also in helping to rule out misrepresentation uncertainty. Consider Sullivan's (2022, 127) discussion of Wang and Kosinski's (2018) DL model. She explains that "researchers built the model . . . to see whether it was possible to identify an individual's sexual orientation based on facial features alone." Their model is trained on profile pictures taken from dating websites (x-data) and self-reported orientation for said websites (y-target). Although the reported estimator quality metrics are high enough to suggest mutual information between *these features* (profile pictures and self-reports), Sullivan (2022, 127) observes that 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." According to our account, this is a case of misrepresentation uncertainty. Instead of using data that represents "facial features alone" and "sexual orientation" they allow for data leakage in the form of "grooming patterns and profile conventions." The extent to which this data leakage was influential could be achieved by alternative data sampling methods. It might also be better understood through an analysis of FI using permutation importance techniques such as visual occlusion of grooming features, etc., which are believed to have leaked confounding information,[17] or more gradient-based techniques used to identify which parts of an image have a greater impact on model estimation. Similarly, studying how the DL model transforms and represents higher-level features may also inform greater LR understanding of the actually represented features.

Although the three modes of ML understanding explored in Section 4 are not intended to be comprehensive, we emphasize that a predictable relationship between the represented features is not necessarily causal. Background scientific theory is vital for inferring causal claims from information-theoretic claims about features. Sullivan argues effectively against the existence of any background scientific knowledge supporting parental hormone theory (PHT), an origin theory for sexual orientation. Even if the model were reconciled of its misrepresentation uncertainty using improved data sampling methods, etc., further inference from such an IR (between the features veridically representing facial structure and appropriately defined orientation features) cannot be made based on an ML model alone. Though PHT causal hypotheses may be loosely related to the features targeted by such a methodologically rectified ML model, far more scientific work (in the context of an appropriate background theory) is required before it could be said to even partially provide supporting evidence. It is a question of what kind of evidentiary patterns do (and do not) support causal claims within a scientific

[17] FI might similarly be used to detect the above identified potential dermatology data leakage risks, by understanding the importance of markings and ruler usage included in the x-data.

domain. Origin theories about causal links such as PHT are not merely suspect; they fall outside the scope of target features to be understood by ML models according to the TML hypothesis.

6. Conclusion

Our account can be taken in part as a development of Sullivan (2022), adding further distinctions. Specifically, (1) we have explored the TML hypothesis as an account of the appropriate kind of target for ML models, and (2) we have identified MI, FI, and LR as (at least) three modes of understanding such targets with ML models. (3) We have argued that functionally approximate irrelevance be distinguished from implementation irrelevance, and we have suggested that this distinction helps illuminate why parameter detail proliferation does not necessarily render the level of questions answered by MI, FI, and LR opaque. Last, (4) we have argued that the difference between empirical and misrepresentation-based link uncertainty brings more clarity to the role of background scientific knowledge in supporting ML model based understanding.

References

Achille, A., & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19(1), 1947-1980. https://doi.org/10.1109/ITA.2018.8503149

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

Bogen, J, and Woodward, J., (1988). “Saving the Phenomena,” *Philosophical Review*, XCVII (3): 303–352. https://doi.org/10.2307/2185445

Bokulich, A. (2008). *Reexamining the Quantum–Classical Relation*, Cambridge: CUP. https://doi.org/10.1017/CBO9780511751813

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv:1607.06520. https://doi.org/10.48550/arXiv.1607.06520

Chen, R. T., Li, X., Grosse, R. B., & Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. In NeuroIPS, 2610-2620. https://doi.org/10.48550/arXiv.1802.04942

Da Costa, N. C. A. and s. French, 2003, *Science and Partial Truth: A Unitary Approach to Models and Scientific Reasoning*, Oxford: OUP. https://doi.org/10.1093/019515651X.001.0001

Dev, S., Hassan, S., & Phillips., J. M., (2019) Closed form word embedding alignment. In IEEE ICDM, 130–139. https://doi.org/10.1109/ICDM.2019.00023

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. nature, 542(7639), 115-118. https://doi.org/10.1038/nature21056

Frigg, R. and J. Nguyen, (2016). “The Fiction View of Models Reloaded”, *The Monist*, 99(3): 225–42. https://doi.org/10.1093/monist/onw002

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A. and Uhlmann, L. (2018). Man against machine. Annals of Oncology, 29(8), 1836-1842. https://doi.org/10.1093/annonc/mdy166

Khalifa, K. 2017. *Understanding, Explanation, and Scientific Knowledge.* Cambridge: CUP. https://doi.org/10.1017/9781108164276

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. In ICLR. https://doi.org/10.48550/arXiv.1312.6114

Kuehlkamp, A., Becker, B., & Bowyer, K. (2017). Gender-from-iris or gender-from-mascara?. In 2017 IEEE Winter conference on applications of computer vision 1151-1159. https://doi.org/10.1109/WACV.2017.133

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. arXiv:1902.01007. https://doi.org/10.18653/v1/P19-1334

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26. https://dl.acm.org/doi/10.5555/2999792.2999959

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., Carter, S.(2020) Zoom In: An Introduction to Circuits, Distill, 2020. https://doi.org/10.23915/distill.00024.001

Patel, N. 2017. “Why Doctors Aren’t Afraid of Better More Efficient AI Diagnosing Cancer.” *The Daily Beast*. https://www.thedailybeast.com/why-doctors-arent-afraid-of-better-more-efficient-ai-diagnosing-cancer

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In EMNLP 1532-1543. https://doi.org/10.3115/v1/D14-1162

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press. https://doi.org/10.7208/chicago/9780226507194.001.0001

Schelling, T. C. (1971) Dynamic Models of Segregation, The Journal of Mathematical Sociology, 1, 143–186. https://doi.org/10.1080/0022250X.1971.9989794

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In International conference on machine learning (pp. 3145-3153). PMLR. https://doi.org/10.48550/arXiv.1704.02685

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks. arXiv:1312.6034. https://doi.org/10.48550/arXiv.1312.6034

Strevens, M. (2008). *Depth: An Account of Scientific Explanation,* Cambridge, MA: Harvard University Press.

Sullivan, E. 2022. “Understanding from Machine Learning Models.” The British Journal for the Philosophy of Science, 73(1): 109–133. https://doi.org/10.1093/bjps/axz035

Suárez, Mauricio, (2004). “An Inferential Conception of Scientific Representation”, *Philosophy of Science*, 71(5): 767–779. https://doi.org/10.1086/421415

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition 2818-2826. https://doi.org/10.1109/CVPR.2016.308

Tamir, M. and E. Shech. 2022. Manuscript. “Understanding and Information Flow in Deep Learning Representations.”

Torralba, A., & Efros, A. A. (2011, June). Unbiased look at dataset bias. In IEEE CVPR 2011 1521-1528. https://doi.org/10.1109/CVPR.2011.5995347

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762. https://doi.org/10.48550/arXiv.1706.03762

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. Journal of personality and social psychology, 114(2), 246. https://doi.org/10.1037/pspa0000098

Weisberg, D. S., & Gopnik, A. (2013). Pretense, counterfactuals, and Bayesian causal models: Why what is not real really matters. Cognitive science, 37(7), 1368-1381. https://doi.org/10.1111/cogs.12069

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W. & Haenssle, H. A. (2019). Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA dermatology, 155(10), 1135-1141. https://doi.org/10.1001/jamadermatol.2019.1735

Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham https://doi.org/10.1007/978-3-319-10590-1_53