Philosophy of Science (), 1–
doi:

PSA PHILOSOPHY OF SCIENCE ASSOCIATION PHILOSOPHY OF SCIENCE

Deep Neural Networks, Architectural Constraints, and Scientific Representation

Phillip H. Kieval[1]

[1]Department of Philosophy, University of Florida
*Corresponding author: Phillip H. Kieval; pkieval@ufl.edu

Abstract

I argue that the practice of applying generic deep neural network (DNN) architectures with minimal theoretical constraints falls short of the conditions required for scientific representation on both substantive or deflationary accounts. Substantive views fail because the characteristic interpretive activities that establish representation relations are absent from generic DNN practice. Deflationary views fare better but risk trivializing the concept of surrogative inference when applied to generic DNNs. I then propose that theoretically-motivated architectural constraints function as a form of indirect characterization that grounds deflationary scientific representation.

1. Introduction

Whether and how machine learning models based on deep neural networks (DNNs) represent their targets has become the subject of recent debate among philosophers of science (Boge, 2022; Sullivan, 2023; Tamir and Shech, 2022; Kieval, 2025). These questions follow naturally from the central role that representation has played in the philosophy of model-based science. Representationalism has been the dominant approach to understanding how inferences from models to their targets (surrogative inferences) can be justified (see Frigg and Nguyen (2017) for a review). This view holds that models must represent their targets for us to learn from them. Hereafter, I call a representation that licenses surrogative inferences a *scientific representation.* Whether DNNs can function as scientific representations thus has important consequences for how we should conceptualize the epistemology of machine learning in science.

Doubts about whether DNNs function as scientific representations stem from the observation that their internal structure lacks the conceptual richness required to license surrogative inferences from elements of a model to aspects of a real-world target (Boge, 2022). At best, the parameterized structure of a DNN is very difficult to interpret. Artificial neurons tend to be *polysemantic;* they do not respond to an interpretable property of the input, or they appear to respond to confounding and chimerical combinations of features (Elhage et al., 2022). At worst, some parts of a DNN's structure such as the weight values between neurons may be semantically meaningless. Hence, their structure seems unable to support surrogative inferences and thus cannot function as a scientific representation of target phenomena.

In reply, Sullivan (2023) has recently argued that DNNs represent their targets in a manner akin to highly idealized toy models. Sullivan argues that skeptical doubts about DNNs' capacity to represent assume an inflexible conception of representation in terms of substantive relations like similarity or isomorphism. She rightly observes that scientific representation depends on the context-sensitive, interpretive activities that license valid surrogative inferences. On such an interpretive view of representation, highly idealized toy models can lack high degrees of similarity and nevertheless represent their target. Sullivan thus argues that, since DNNs function in a similar way to idealized toy models, they represent their targets by similar lights.

While I think that Sullivan's (2023) emphasis on idealization and interpretation is on the right track, her view fails to fully appreciate how the modeling activities involving DNNs comprise a significant departure from the ways that philosophers of science have tended to characterize surrogative inference. In particular, her account still trades on a substantive conception of scientific representation in terms of mapping activities that do not feature in the practices of constructing and applying DNNs as scientific models.

In this paper, I argue that the practice of applying generic DNN architectures with minimal theoretical constraints falls short of the conditions required for scientific representation. I then argue that attention to machine learning practice reveals how theoretically-motivated architectural choices function as a form of indirect characterization that can ground scientific representation. This claim draws on deflationary views of scientific representation that tie representation to the social-epistemic practices that justify surrogative inferences rather than to any objective relation between model and target. The key move lies in recognizing that some deflationary views leave open the question of what kinds of practices can do this justificatory work (see Suárez, 2024). I argue that theoretically-motivated architectural constraints can play this role by encoding domain-specific background knowledge as inductive biases on learning. The fittingness of these choices provides modelers with defeasible warrant that a trained model's internal representations track generalizable structure in the target system. This warrant distinguishes a DNN that functions as a scientific representation from one that merely produces accurate predictions. Hence, my critique also targets a specific and widespread practice of applying generic, off-the-shelf DNN architectures to scientific problems with little regard for domain-specific theoretical knowledge.

In Section 2, I distinguish substantive from deflationary views of scientific representation. Section 3 argues that substantive views must rule out DNNs as scientific representations, since the characteristic interpretive activities that establish the representation relation are absent from DNN modeling practice. In Section 4, I suggest that deflationary views may fare better, but that accommodating generic DNNs forces a choice between restricting surrogative inference to practices DNNs do not exhibit, or expanding it so broadly that the concept risks triviality. I conclude by arguing that attention to architectural choices in machine learning practice can resolve this tension. I argue that theoretically-motivated architectural constraints function as a form of indirect characterization of the target system. By this I mean a way of giving empirical meaning to a model's structure in light of background knowledge of its target, which can ground deflationary scientific representation without requiring the interpretive activities that DNNs lack. My account of indirect characterization suggests an important epistemological distinction between practices that apply generic DNN architectures to any

data and more sophisticated modeling activities that involve theoretically constrained architectural choices.

2. Substantive and Deflationary Views of Scientific Representation

Recent work in the philosophy of science tends to characterize accounts of scientific representation into substantive and deflationary varieties (Suárez, 2015). Substantive views characterize representation in terms of some objective property or relation that explains how models enable surrogative inferences. Substantive views have largely converged around a position I will call *pragmatic representationalism*. Pragmatic representationalism gives a prominent role to the users of a representation in fixing the conditions under which a model M represents a target T. These views stress the importance of human judgment in the successful use of models to represent the world. Scientific representation must be a representation *for us*; to be of any epistemic value, a representation must satisfy the cognitive needs of limited knowers such as we. Whether they understand the representation relation in terms of similarity (Weisberg, 2013; Giere, 2010), morphisms between abstract structures (van Fraassen, 2008), or denotation (Hughes, 1997; Contessa, 2007; Frigg and Nguyen, 2016), the overwhelming majority of substantive views commit themselves to a version of pragmatic representationalism.

Such views acknowledge that context-sensitive norms of assessment partially determine whether the representation relation holds between M and T. Even so, substantive views maintain that a substantive relation between M and T explains how surrogative inferences can be justified. Take, for example, Weisberg's weighted feature mapping account of similarity. For Weisberg (2013), "models consist of an *interpreted structure* that can be used to represent a real or imagined phenomenon" (Weisberg, 2013, 15, emphasis original). Such interpretations mean agents adopt specific *construals* of a model that fix relations of denotation between elements of M and T (2013, 39). Part of a construal is what Weisberg calls the *assignment*, which determines the relationship between parts of M's structure and features of T. Assignments are thus "explicit specifications of how parts of real or imagined target systems are to be mapped onto the parts of the model" (Weisberg, 2013, 39). Assignments tend to be implicit in the interpretive norms and practices adopted by specific epistemic communities. Although the details vary, Hughes's (1997) denotation-demonstration-interpretation (DDI) account and Frigg and Nguyen's (2016) DEKI account offer similar analyses of how modelers interpret the structure of M in light of background knowledge of T.

Be it denotation, isomorphism, or similarity, substantive views invoke an objective representation relation to explain how models enable surrogative inferences, and marry this to pragmatic representationalism. These activities involve context-sensitive norms that govern how modelers establish mappings between elements of M and features of T to license valid surrogative inferences. Yet the representation relation itself still does the heavy lifting when it comes to explaining the epistemic success of a model.

Deflationary views take pragmatic representationalism to its furthest conclusion by reducing scientific representation to the social-epistemic practices that constitute or maintain the norms governing valid surrogative inferences (Suárez, 2004, 2024; Khalifa et al., 2022). Deflationary accounts deny that any substantive, factual relationship obtains between models and their targets, instead tying representation to surrogative

inference. They make no attempt to define representation in terms of a substantive relation between M and T; instead, they explain how M relates to T by reference to an account of surrogative inference.

3. DNNs and Substantive Views of Representation

Working with DNNs involves using domain-general techniques and procedures to construct predictive models that are adequate-for-purpose. One recent lesson data scientists have learned from DNNs is that flexible methods of learning from large quantities of data tend to produce the best predictions when dealing with very complex phenomena. Such flexibility means modelers make minimal assumptions about their target phenomenon. Rather than relying on domain-specific theoretical models to tightly constrain the space of possible functions a model could learn, machine learning researchers engineer domain-general neural architectures that learn predictive features from the data on their own. This difference tracks the distinction between what Knüsel and Baumberger (2020) call process-based models and data-driven models, or what Breiman (2001) calls data modeling versus algorithmic modeling (cf. Shmueli, 2010).

Training makes flexible learning effective by teaching networks to transform representations of data into forms that simplify processing at later layers. Machine learning relies on a geometric framework to explain this process. We can view each data sample as a point in a multidimensional 'feature space,' where the axes represent the dataset's qualitative features. The network learns transformations of this feature space that reduce task complexity. In classification tasks, for example, the network may learn to transform inputs into a representation that becomes linearly separable at the penultimate layer, allowing the network to draw a decision boundary that divides inputs into distinct categories. Feature space exists in such high dimensions that we cannot visualize or even conceive the structural properties of learned representations. The network itself thus serves as the user of these representations, not human inquirers.

To treat a trained DNN as a scientific representation, we might try to approach it like simpler parametric statistical models, examining how the network's learned parameters represent the complex relationship between predictor features and target variables. After all, there is one sense in which a DNN's internal structure is fully transparent. Even the most complex DNN constitutes "a closed system of effectively computable operations where rules and transformations are mechanically applied to inputs to determine outputs" (Leslie, 2019, 41). We can, in principle, manually inspect a DNN's complete internal structure.

Yet inspecting the total internal workings of a DNN seems unlikely to enable surrogative inferences. There are several types of transparency in machine learning models (cf. Zerilli, 2022; Creel and Hellman, 2022; Boge, 2022). Lipton (2018) defines *fathomability* as the degree to which someone can immediately grasp how the model's features relate to its predictions. A fathomable model allows "a human to take the input data together with the parameters of the model and in reasonable time step through every calculation required to produce a prediction" (Lipton, 2018, 38-39). Yet even very small models by today's standards typically contain tens of millions of trainable parameters, and this sheer number prevents us from contemplating the model's structure holistically. Deep learning models remain unfathomable due to their size and complexity alone.

A lack of fathomability marks a practical—not in-principle—obstacle to interpretability.[1] Linear statistical models, for instance, can also become too complex to fathom, yet remain interpretable as scientific representations because each parameter plays a semantically interpretable role. Linear models typically map the target system's features to parameters in a linear probability model. Linearity means every term constitutes either a constant or a feature variable weighted by a tunable parameter, and the model is additive, meaning there are no interactions between predictor variables. Linearity as such implies conditional monotonicity: conditional on all other explanatory variables, a change in one feature variable leads to a predictable change in the dependent variable. Without at least conditional monotonicity, our intuitions about the relationship between feature variables and the target phenomenon tend to fail. Linearity also permits validation techniques like sensitivity analysis, which allows modelers to determine how uncertainty in the model's parameters accounts for uncertainty in its outputs—providing information about patterns of counterfactual dependency crucial for explanatory power (see Woodward, 2003).

Though linearity is by no means the sole measure of interpretability, it marks one way of ensuring that statistical models remain interpretable even when very complex.[2] My point here is that the distinctive combination of non-linearity, hierarchical organization of processing, polysemantic distributed representations (see e.g. Smolensky, 1988; Elhage et al., 2022), and massive overparameterization results in an inability to interpret individual parts of a DNN in terms of cognitively meaningful features of the target system (see also Boge, 2022).

I take this analysis to reveal that DNNs lack interpretability in a sense that goes beyond their unfathomable number of parameters. DNNs contain extreme nonlinearities due to their many stacking layers of functions embedded within functions, making it practically impossible to disentangle how a model relates features of the target to produce its predictions. There is no way to meaningfully interpret individual parameter values in terms of qualitative features of the target system, because there is no presumed connection between model structure and target. In scientific practice, the number of neural network parameters often far outstrips the number of quality dimensions in the data. These parameters correspond to learned weights governing the transformation of signal between layers, but those transformations need not have any intrinsic connection to the target phenomenon for the model to prove useful.

Insofar as DNNs license surrogative inferences from M to T, they appear to do so through brute calculation rather than by enabling modelers to reason about features of M to draw out inferences about T. Sullivan (2023) is correct to point out that choices about network architectures and data processing involve idealizing assumptions about the target distribution. But the presence of idealizing assumptions alone does not establish that using DNNs involves the distinctive interpretive activities and patterns of reasoning involved in surrogative inferences from toy models. These activities paradigmatically involve partitioning the structure of M into various parts and interpreting these parts as denoting or standing in for features of T. But the actual practice of constructing and

[1] On the distinction between essential and non-essential epistemic opacity, see Humphreys (2009).

[2] Decision trees, for instance, are non-linear but tend to be interpretable in virtue of how their representation partitions a feature space into simple geometric regions. Both Lipton (2018) and Räz (2024) show convincingly that interpretability cannot be reduced to a monolithic concept, and different modeling techniques exhibit trade-offs between different aspects of interpretability.

applying a DNN tends to be totally divorced from such activities. Once trained, the model supplies conclusions about the global behavior of T directly through calculations without any need for an interpretive map between structured parts of M and features of T.

While we may attempt to establish such a map using post hoc interpretability methods or mechanistic interventions like linear probes or activation patching, these methods are in an important sense directed at the wrong target. They aim to establish that a DNN uses information about semantically interpretable features of T and that this information plays a causal role in producing accurate predictions. But the mere fact that a network uses such information need not suffice to establish a *usable* scientific representation that human inquirers can successfully latch on to and which licenses valid surrogative inferences about T. Such techniques therefore aim primarily at explaining the activity of a model itself without necessarily providing firm ground for a model-induced explanation of T vis-a-vis a usable scientific representation (cf. Lawler and Sullivan, 2021).

Inferences from DNNs to their targets thus exhibit what we can call the *bare form* of surrogative inference:

M says that P
Therefore, C.

where C is a claim about the target system T inferred from a premise P derived from the model M. When it comes to DNNs, 'derived' is quite literal, since P follows from calculations performed by the trained model. But when dealing with more traditional mathematical models, such derivations may turn on researchers' abilities to inspect or manipulate elements of the model to draw out some conclusion P (Khalifa et al., 2022, 268).

Substantive views hold that characteristic activities establish mapping relations between the interpreted structure of M and features of T. Modeling communities thus justify the bare form of surrogative inference in terms of a representation relation established through interpretive activities. But modeling practices with DNNs lack these characteristic activities. We lack agreed-upon norms for interpreting the structure of M to reason about T, and making effective use of a DNN does not typically depend on reasoning about the interpreted structure of M or establishing mappings between claims about M and T.[3] It involves the bare form of surrogative inference alone. Modelers infer conclusions about T from premises derived from calculations performed with M without any well established norms for interpreting elements of the structure of M or mapping claims about those elements to features of T. Without such interpretive norms, there can be no shared conventions for establishing the representation relation. Hence, substantive views of scientific representation must rule out that DNNs represent their targets.

[3] I am bracketing here the use of DNNs as computational models of the processes underlying various cognitive capacities. There is substantial work, for instance, demonstrating significant correspondence between the hierarchical stages of processing in the layers of convolutional neural networks and the primate ventral stream (see e.g. Yamins and DiCarlo, 2016; Kieval, 2022; Cichy et al., 2016). In general, I take the use of DNNs as models in connectionist approaches to cognitive science to be a special case with considerable methodological differences from applications of DNNs as predictive models in science more broadly.

4. DNNs and Deflationary Views of Representation

Deflationary views hold that what counts as a scientific representation depends on the practices of justifying surrogative inferences (Suárez, 2024; Khalifa et al., 2022). In their recent defense of inferentialist-expressivism, Khalifa et al. (2022) provide the following necessary and sufficient conditions on scientific representation:

> M is a scientific representation of T if and only if M has scientifically justified surrogative consequences that are answers to questions about T, where a surrogative consequence is justified if and only if the entitlements of its inferential pedigree have been secured (Khalifa et al., 2022, 274).

According to their view, justifying surrogative inferences depends on securing five different kinds of entitlements. These entitlements form an interdependent web of epistemic statuses achieved by a given scientific community which jointly justify surrogative inferences (Khalifa et al., 2022, 272). Once secured, this justification establishes a model's 'inferential pedigree.'

These five entitlements include DERIVATION, CHARACTERIZATION, RELEVANCE, MEASUREMENT, and NO DEFEATERS (see Figure 1). DERIVATION largely corresponds to what I have called the bare form of surrogative inference—inferring conclusions about T from premises derived from M. Khalifa et al. (2022) argue that DERIVATION depends on further entitlements that fix how modelers give empirical meaning to elements of the model. CHARACTERIZATION involves appropriately interpreting elements of M in light of background knowledge about T. RELEVANCE involves evaluating whether a given CHARACTERIZATION of M can produce answers to questions about T, and is thus sensitive to scientists' context-relative tolerance for idealization. MEASUREMENT addresses choosing and justifying measurement methodologies and establishing values for model parameters; it is interdependent with CHARACTERIZATION, since one cannot justify measurements without some interpretation of variables, and measurement possibilities constrain available interpretations. Finally, NO DEFEATERS requires the absence of facts that would refute the surrogative conclusion or undercut the premise's support of it.

I take Sullivan's (2023) discussion of idealization in machine learning to show that modelers working with DNNs make choices about data processing and model architecture that secure RELEVANCE, and that data sampling and input variable choices secure MEASUREMENT. DNNs also involve reading off predictions derived by calculations performed by the model itself—a special case of DERIVATION where $P = C$ (cf. Khalifa et al., 2022, 268). But, as I have already argued, CHARACTERIZATION does not generally feature in the practice of building, validating, and applying DNNs. Using DNNs involves only the bare form of surrogative inference without appealing to established norms of characterizing a model in terms of a partitioned structure whose elements have determinate empirical meanings. If these entitlements are each necessary and jointly sufficient, then DNNs plainly cannot support warranted surrogative inferences.

There remains a minimal sense in which DNNs enable surrogative inference: they produce predictions with determinate representational content that practitioners use to make inferences about a target. In this sense, DNNs *produce* scientific representations without themselves being subject to the activities that ordinarily characterize inferential

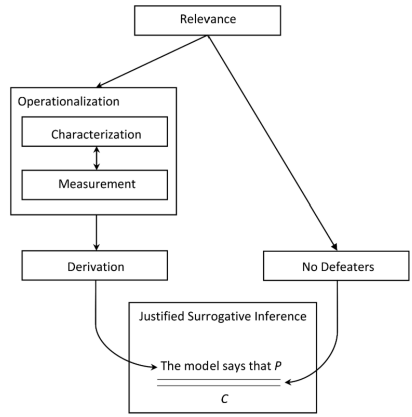

Figure 1. The structure of "inferential pedigree." Figure adapted from Khalifa et al. (2022).

pedigree. This distinction between the products of a model and the model itself matters for how we characterize surrogative inference. It is questionable whether inferences made on the basis of predictions *merely produced* by an otherwise opaque model count as instances of surrogative inference. Even for deflationists, inference *qua surrogative inference* seems to essentially involve reasoning from part of a partitioned description of M to features of T under a prescribed description (cf. Khalifa et al., 2022; Suárez, 2024, Ch. 7). But CHARACTERIZATION — the set of activities that makes this possible — is absent from the practice of deep learning.

Suárez's deflationary inferentialism (2024) contains potential resources to overcome this objection. Suárez could maintain—in a thoroughgoing deflationary spirit—that *any practices* of using one object to draw out inferences about another object might count as an instance of surrogative inference.[4] This reply holds that cases of surrogative inference need not involve the particular activities described in the inferential pedigree. Suárez himself does not consider cases of surrogative inference that do not involve reasoning from partitioned features in a description of M to their correlates in T. Yet there is some textual support for this suggestion: Suárez holds that it is best to think of modeling activities "in terms of their effective upshot, namely, the licensing of some surrogative inferences" (Suárez, 2024, 156). One way to read Suárez (2024) is that there is nothing special about the particular activities in the inferential pedigree aside from the contingent historical fact that model-based science has relied on them to manage valid rules of inference. Different practices could emerge which eschew these activities but nevertheless count as surrogative inference. In its most minimal form, Suárez's inferential conception can accommodate DNNs as licensing a peculiar species of surrogative inference that involves a very different set of activities and norms.

This conclusion leaves us at a fork in the road. First, one could maintain that surrogative inference must involve the kinds of characteristic practices that establish epistemic entitlements to use M as a representation of T. Perhaps, these practices are constitutive

[4] I have come to appreciate the potential viability of this strategy thanks to Mauricio Suárez (personal correspondence, August 12, 2025).

of surrogative inference. In this case, we may have to concede that DNNs do not function as scientific representations. On the other hand, one may accept that any practices of using M to infer conclusions about T will suffice. On this view, even the bare form of surrogative inference will be sufficient to establish warranted surrogative inferences. I suspect that some may view this move as a perversion of the meaning of surrogative inference and a substantive revision of philosophers' conception of the epistemic capacities of scientific models.

Neither option is particularly satisfying. The first requires us to deny that a wide class of successful scientific tools count as representations at all. The second deflates representation so thoroughly that it threatens to become trivial. The shared meanings of our talk about surrogative inference and representation are historical products shaped by the gradual establishment of norms governing a particular approach to doing model-based science. Disrupting this style of modeling and its concomitant normative practices seems to break down the loosely unifying elements that fix what we mean by surrogative inference and representation. Our philosophical vocabulary was developed to capture one style of model-based science, and DNNs seem to genuinely disrupt these practices. Our conceptions of 'inferential pedigree' and 'surrogative inference' grew out of specific historical contexts where they proved useful for thinking about the epistemology of the actual practice of model-based science. One may thus worry that deflating surrogative inference even further to *any* inference from one thing to another would involve a significant shift in the meaning of the term, even as Suárez deploys it. This rejoinder is not necessarily a decisive blow against the move, but it does suggest a need for further specification and defense of the new, even further deflated sense of surrogative inference that applies to DNNs.

5. Conclusion: Model Architectures as Indirect Characterization

I have argued that if DNNs are to function as scientific representations, it is on the most minimal, deflationary conception available. Scientists can use DNNs without necessarily appealing to a characterization of the model in terms of a partitioned structure whose parts can be mapped to determinate empirical meanings. In my view, however, attention to machine learning practice can begin to shed light on emerging epistemic strategies for justifying choices about *architectures*, or the conceptual structure and logical organization of the different components and processes that make up a machine learning system.

These modeling strategies reflect the fact that flexible, domain-general learning is a double-edged sword. On the one hand, neural networks are universal function approximators: a large enough network can represent any continuous, differentiable function to an arbitrary degree of accuracy (Hornik et al., 1989). The success of these methods on exceptionally complex problems like protein folding (Jumper et al., 2021) and global weather forecasting (Lam et al., 2023) reflects an abundance of learnable structure in natural data, often more complex than can be represented using simple law-like mathematical models. But the expressivity and flexibility of these models also means modelers must worry about overfitting to spurious correlations (Geirhos et al., 2020) and memorization (Zhang et al., 2021), which undermine generalization. To mitigate these issues, modelers make architectural choices that impose interpretable, top-down constraints on learning. These architectural constraints encode inductive biases, or preferences over

the space of functions a model can learn (see Goyal and Bengio, 2022; Wilson, 2025; Liu et al., 2026; Lampinen et al., 2024). These constraints aim to balance selectivity, or sensitivity to complex features, against tolerance for irrelevant variation. The most successful models tend to involve boutique architectures specifically tailored to the task (see e.g. Jumper et al., 2021; Lam et al., 2023). AlphaFold2, for example, implements modified self-attention functions that impose theoretically informed constraints such as requiring estimated distances between amino acids to conform to the triangle inequality (Jumper et al., 2021, 586).

I suggest that the fittingness of neural architectures with background knowledge of a theoretical problem provides evidence that a trained model represents relevant, generalizable structure in data. To assess whether a DNN architecture in fact imposes such fitting constraints, modelers need to understand how different architectural choices encode specific biases on statistical learning. This understanding depends on a large repertoire of applied mathematical techniques built up through experimental investigations of neural networks. Modelers then generalize these techniques to new situations where data have similar structural properties. As such, the practices that justify the use of DNNs in specific scientific contexts are always inductive, local, and provisional. This picture broadly aligns with Norton's (2021) material theory of induction, on which inductive inferences are always warranted by context-specific material facts rather than universal rules or formal schemas. In this way, the practices surrounding architectural choice begin to fill out the gaps left by the absence of explicit CHARACTERIZATION. Rather than interpreting a model's internal structure in terms of features of T, modelers instead impose theoretically-grounded constraints on what the model is permitted to learn, providing defeasible warrant that its learned representations track genuine structure in the target.

I claim that we can view such architectural choices as a kind of *indirect characterization* whereby modelers use their domain-specific theoretical knowledge of a target system to impose fitting top-down constraints on statistical learning. These choices give modelers some confidence that their DNN model will learn internal representations that track generalizable features, even when the precise content of those representations remains inaccessible. Indirect characterization does not recover full CHARACTERIZATION in Khalifa et al.'s (2022) sense, because the internal representations of the model remain largely inaccessible. But such indirect characterization provides enough epistemic purchase to ground the bare form of surrogative inference on a deflationary view. On this account, the theoretically-motivated reliability that indirect characterization underwrites is what distinguishes a DNN functioning as a scientific representation from one that merely produces accurate predictions.

This understanding of network architectures suggests that DNNs only function as scientific representations when constructed to encode theoretically-informed inductive biases. A DNN that encodes theoretically-motivated inductive biases provides stronger warrant for the bare form of surrogative inference, because indirect characterization makes it more likely that the model's learned representations track generalizable structure in the target. Conversely, a generic architecture applied without regard for domain-specific background knowledge offers weaker warrant (see also Rathkopf, 2026). This speaks against the tendency in machine learning practice to apply generic, off-the-shelf architectures to scientific problems with little regard for domain-specific background knowledge.

References

Boge, F. J. (2022) Two Dimensions of Opacity and the Deep Learning Predicament. *Minds and Machines*. *32*(1), 43–75. 10.1007/s11023-021-09569-4.

Breiman, L. (2001) Statistical modeling: The two cultures. *Statistical Science*. *16*(3), 199–215. 10.1214/ss/1009213726.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. and Oliva, A. (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*. *6*(January), 1–13. 10.1038/srep27755.

Contessa, G. (2007) Scientific representation, interpretation, and surrogative reasoning. *Philosophy of Science*. *74*(1), 48–68. 10.1086/519478.

Creel, K. and Hellman, D. (2022) The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*. *52*(1), 26–43. 10.1017/can.2022.3.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M. and Olah, C. (2022) *Toy Models of Superposition*.

Frigg, R. and Nguyen, J. (2016) The Fiction View of Models Reloaded. *The Monist*. *99*(3), 225–242. 10.1093/monist/onw002.

Frigg, R. and Nguyen, J. (2017) Models and Representation In *Springer Handbook of Model-Based Science*, Magnani, L. and Bertolotti, T. (eds). Springer. pp. 49–102. Available at: `http://link.springer.com/10.1007/978-3-319-30526-4_3`. 10.1007/978-3-319-30526-4_3.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. and Wichmann, F. A. (2020) Shortcut learning in deep neural networks. *Nature Machine Intelligence*. *2*(11), 665–673. 10.1038/s42256-020-00257-z.

Giere, R. N. (2010) An agent-based conception of models and scientific representation. *Synthese*. *172*(2), 269–281. 10.1007/s11229-009-9506-z.

Goyal, A. and Bengio, Y. (2022) *Inductive Biases for Deep Learning of Higher-Level Cognition*.

Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*. *2*(5), 359–366. 10.1016/0893-6080(89)90020-8.

Hughes, R. I. G. (1997) Models and Representation. *Philosophy of Science*. *64*(S4), S325–S336. 10.1086/392611.

Humphreys, P. (2009) The philosophical novelty of computer simulation methods. *Synthese*. *169*(3), 615–626. 10.1007/s11229-008-9435-2.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. and Hassabis, D. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*. *596*(7873), 583–589. 10.1038/s41586-021-03819-2.

Khalifa, K., Millson, J. and Risjord, M. (2022) Scientific Representation: An Inferentialist-Expressivist Manifesto. *Philosophical Topics*. *50*(1), 263–292. 10.2307/48681556.

Kieval, P. H. (2022) Mapping representational mechanisms with deep neural networks. *Synthese*. *200*(3). 10.1007/s11229-022-03694-y.

Kieval, P. H. (2025) Representation learning without representationalism: A non-representational account of deep learning models in scientific practice In *Philosophy of Science for Machine Learning: Core Issues and New Perspectives*, Durán, J. M. and Pozzi, G. (eds). Synthese Library.

Knüsel, B. and Baumberger, C. (2020) Understanding climate phenomena with data-driven models. *Studies in History and Philosophy of Science Part A*. *84*, 46–56. 10.1016/j.shpsa.2020.08.003.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., Merose, A., Hoyer, S., Holland, G., Vinyals, O., Stott, J., Pritzel, A., Mohamed, S. and Battaglia, P. (2023) Learning skillful medium-range global weather forecasting. *Science*. *382*(6677), 1416–1421. 10.1126/science.adi2336.

Lampinen, A. K., Chan, S. C. Y. and Hermann, K. (2024) *Learned feature representations are biased by complexity, learning order, position, and more*.

Lawler, I. and Sullivan, E. (2021) Model Explanation Versus Model-Induced Explanation. *Foundations of*

Science. *26*(4), 1049–1074. 10.1007/s10699-020-09649-1.
Leslie, D. (2019) Understanding artificial intelligence ethics and safety.
Lipton, Z. C. (2018) The mythos of model interpretability. *Communications of the ACM*. *61*(10), 35–43. 10.1145/3233231.
Liu, Z., Sanborn, S., Ganguli, S. and Tolias, A. (2026) From Kepler to Newton: Inductive Biases Guide Learned World Models in Transformers.
Norton, J. D. (2021) *The Material Theory of Induction*. University of Calgary Press. 10.2307/j.ctv25wxcb5.
Rathkopf, C. (2026) From Hallucination to Reliability: Generative Modeling and the Structure of Scientific Inference.
Räz, T. (2024) ML interpretability: Simple isn't easy. *Studies in History and Philosophy of Science*. *103*(November 2023), 159–167. 10.1016/j.shpsa.2023.12.007.
Shmueli, G. (2010) To explain or to predict? *Statistical Science*. *25*(3), 289–310. 10.1214/10-STS330.
Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences*. *11*(1), 1–74. 10.1017/S0140525X00052432.
Suárez, M. (2004) An inferential conception of scientific representation. *Philosophy of Science*. *71*(5), 767–779. 10.1086/421415.
Suárez, M. (2015) Deflationary representation, inference, and practice. *Studies in History and Philosophy of Science Part A*. *49*, 36–47. 10.1016/j.shpsa.2014.11.001.
Suárez, M. (2024) *Inference and Representation: A Study in Modeling Science*. University of Chicago Press. Chicago. doi:10.7208/chicago/9780226830032.
Sullivan, E. (2023) Do Machine Learning Models Represent Their Targets? *Philosophy of Science*. pp. 1–11. 10.1017/psa.2023.151.
Tamir, M. and Shech, E. (2022) Understanding from Deep Learning Models in Context In *Scientific Understanding and Representation: Modeling in the Physical Sciences*, Lawler, I., Khalifa, K., and Shech, E. (eds). Routledge. New York. pp. 323–340.
van Fraassen, B. C. (2008) *Scientific Representation: Paradoxes of Perspective*. Oxford University Press. 10.1093/ACPROF:OSO/9780199278220.001.0001.
Weisberg, M. (2013) *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press. New York. 10.1093/acprof:oso/9780199933662.001.0001.
Wilson, A. G. (2025) *Deep Learning is Not So Mysterious or Different*.
Woodward, J. (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.
Yamins, D. L. and DiCarlo, J. J. (2016) Using goal-driven deep learning models to understand sensory cortex. 10.1038/nn.4244.
Zerilli, J. (2022) Explaining Machine Learning Decisions. *Philosophy of Science*. *89*(1), 1–19. 10.1017/psa.2021.13.
Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2021) Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*. *64*(3), 107–115. 10.1145/3446776.