Searching for Features with Artificial Neural Networks in Science: The Problem of Non-Uniqueness

Siyu Yao and Amit Hagar

Department of History and Philosophy of Science and Medicine, Indiana University, Bloomington, USA

Correspondence Author: Siyu Yao

Address: 1020 East Kirkwood Avenue, Ballantine Hall 916, Bloomington, IN, 47405, USA

Email: siyuyao@iu.edu

Disclosure Statement: T*he authors report there are no competing interests to declare*.

Acknowledgment: We thank Claus Beisbart, Dan Li, Shannon Abelson, Evan Arnet, Sihao Cheng, and Jordi Cat for their constructive feedback on an earlier draft of this paper. This paper stems from a graduate seminar, 'The Robot Scientist', in the Department of History and Philosophy of Science and Medicine, Indiana University Bloomington.

Searching for Features with Artificial Neural Networks in Science: The Problem of Non-Uniqueness

Artificial neural networks and supervised learning have become an essential part of science. Beyond using them for accurate input-output mapping, there is growing attention to a new feature-oriented approach. Under the assumption that networks optimized for a task may have learned to represent and utilize important features of the target system for that task, scientists examine how those networks manipulate inputs and employ the features networks capture for scientific discovery. We analyse this approach, show its hidden caveats, and suggest its legitimate use. We distinguish three things that scientists call a "feature": parametric, diagnostic, and real-world features. The feature-oriented approach aims for real-world features by interpreting the former two, which also partially rely on the network. We argue that this approach faces a problem of non-uniqueness: there are numerous discordant parametric and diagnostic features and ways to interpret them. When the approach aims at novel discovery, scientists often need to choose between those options, but they lack the background knowledge to justify their choices. Consequentially, features thus identified are not promised to be real. We argue that they should not be used as evidence but only used instrumentally. We also suggest transparency in feature selection and the plurality of choices.

Keywords: machine learning, non-uniqueness, artificial neural network, evidence, transparency

Word count: 7,991

1. Introduction

The rise of data-driven science in the last few decades has evoked questions concerning its epistemic and methodological characteristics. Data-driven science is often believed to consist of the utilization of large amounts of data and the application of data-analysis methods like regression, decision trees, and machine learning. Leonelli (2012, 1) points out two features key to data-driven science: 'the intuition that induction from existing data is being vindicated as a crucial form of scientific inference', and the central role of automated reasoning with machines in extracting meaningful patterns from data. These features pinpoint a new context to rethink traditional topics in the philosophy of science like induction, inference, exploration, and the role of theories within research practices (Leonelli 2012; Pietsch 2015).

Machine learning with artificial neural networks (ANN) has been increasingly applied in the natural sciences. Its power to perform accurate input-output mapping has proven it efficient in tasks like data classification and prediction. However, fitting a statistical model to data and making accurate predictions are not all that science is about. The novelty of science often consists in building models that are also causal, explanatory, analyzable, or applicable to a broader scope. Given these diverse scientific goals, how ANNs can further contribute to science is under heated discussion. By analyzing the underlying logic of the various applications of ANNs in science, philosophers can explicate the roles ANNs play and draw normative lessons.

One common approach that scientists take toward ANNs is to utilise their accurate output given certain inputs for tasks like classification and prediction. We call this the *output-oriented approach*, as here scientists are only interested in the outputs without scrutinizing how the network produces them.

Transcending the focus on outputs, many scientists have also started to take a different approach that stresses the use of ANNs' internal parameters after training. Because network parameters are optimized for performing a scientific task, they are anticipated to reflect some *features* in the target system for that task. ANNs might thus contribute to discoveries of real patterns or causal relations in the target system, and thus help to refine our scientific representation (Samek and Müller 2019; Zednik and Boelsen 2022). Because this approach is interested in how elements of ANNs correspond to features in the target system, we call it the *feature-oriented approach*.

The output-oriented approach has been implemented by scientists for more than a decade, but the feature-oriented approach has only recently received attention. Its major technical difficulties, analyzing and understanding how a trained ANN works, have only recently obtained solutions in the explainable artificial intelligence (XAI) project (Adadi and Berrada 2018). Due to the immature status of this approach, its epistemological and methodological aspects have not received sufficient scrutiny. The aim of our paper is thus to examine how the feature-oriented approach can contribute to science, identify potential epistemic caveats, and offer philosophically informed prescriptions.

In section 2, we introduce the feature-oriented approach, including its general conceptualization, methods, aims, and present application. In section 3, we show that the central step of this approach is to identify features belonging to an ANN trained for a scientific task and relate them to features of the target system, which scientists are really interested in. We call the latter *real-world features* and further specify the former into two types, *parametric* and *diagnostic features*. By this, we provide a taxonomy for the different things that scientists call a "feature". We further argue that this approach faces an epistemic issue of non-uniqueness, the existence of multiple equally good mathematical characterizations of a system (Baker 2013; Hagar and Hemmo 2013; Li 2023). When the feature-oriented approach is used for generating novel scientific knowledge, scientists can identify numerous discordant parametric and diagnostic features, but they lack the background knowledge that justifies their selection and interpretation. We argue that non-uniqueness has tricky practical consequences. In section 4, we use a case in cosmology to demonstrate how the approach is performed and present the caveats of non-uniqueness. As a prescription, in section 5, we emphasise the instrumental use of features and the importance of plurality and transparency in ANN-guided scientific discoveries[1].

2. The Feature-Oriented Approach to ANNs in Science

ANNs are computational systems inspired by biological neurons (Fig. 1). They consist of interconnected units. Each unit receives an input and produces an output for other units. The connections among units are assigned a relative weight. In a deep network, the units lie on multiple layers. A nonlinear transformation is incorporated between each layer, allowing complex input-output transformation. In supervised learning, the network is trained using labeled input data. After iterative adjustments of its weights, the network learns to map every input to an output, which appears as a set of probabilities that the input lies in each possible label.

[1] We only discuss supervised learning in this paper, but we expect a similar conclusion for unsupervised learning. According to Cat (2022), algorithms in unsupervised learning are also non-unique, and choosing them involves contextual and subjective judgments.

[Figure 1]

We distinguish two approaches that scientists take toward supervised machine learning with ANNs. These approaches involve different practices, aims, and methodological implications.

One approach, which we call the *output-oriented approach*, regards the network as a black box and only uses its outputs. A trained network is used to correlate inputs, such as a qualitative description of a system or quantitative data, with relevant scientific attributes. For example, it could assign some degree of carcinogenicity to a molecule or decide the biological state of a cell in an image. Here scientists are not interested in *how* these mappings are performed within the network or *what* parts of the input are responsible for the output. ANNs are thus taken as an efficient and accurate replacement for humans or tools that perform the same mapping.

In contrast, a recently emerging approach to ANNs in science seeks to reveal *how* an ANN processes inputs to generate an output. This question is answered with the *features* that the network captures in data to perform the task. In the context of biological or artificial neural networks, a feature is a relational concept that involves the input, the network, the output, and the task: it is a reliable indicator *in the input* that is *used by the network* to *make certain output decisions for a task* (O'Reilly et al. 2012; Bishop 2006). For example, the human visual system contains neurons sensitive to input features from simple edges to shapes of body parts, so that it uses them to recognize what is seen. Likewise, features captured by the ANNs mentioned earlier would be the functional groups in a molecule that make the network attribute carcinogenicity and the shape of a cell image by which the network classifies it as diseased. As this approach expects to use such features beyond the network outputs, we call it the *feature-oriented approach*.

This approach carries a different aim from the output-oriented approach. Here ANNs are not merely expected to replace existing data-processing procedures but to be a unique source of discoveries. ANNs' accurate input-output mapping suggests that their internal representation might have captured important features in the target system for a given task. The network might employ distinctive properties of an object to label it accurately or use causally decisive factors to make a prediction. Because how the network performs the task is *not* manually hard-wired but an original product of its learning process, the network might have learned to process data in a novel way that avoids biases from human cognitive tendencies and existing scientific paradigms (Samek and Müller 2019). Moreover, many suggest a structural analogy between ANNs and the human brain and highlight their potential to form interpretable concepts and capture non-trivial regularities (Buckner 2018; Räz 2022). Therefore, ANNs may illuminate features omitted by human experts and extend our empirical knowledge.

The feature-oriented approach is thus anticipated to serve a unique role, providing both heuristics and evidence. The features suggested by ANNs can be used to refine representations of a phenomenon, identify starting points for future investigation, and suggest potential explanations (Zednik and Boelsen 2022).

Despite the widespread expectation of its contributions to science, this approach is still at the stage of exploration. ANNs have long been notoriously opaque (Lipton 2018; Boge 2022). In the last few years, the feature-oriented approach has been gradually enabled by a broader project of understanding how ANNs work, the explainable artificial intelligence (XAI) project (Adadi and Berrada 2018; Zednik and Boelsen 2022). XAI develops algorithms to answer how and why ANNs make predictions in general or given a specific input. Although XAI has been adopted in domains that involve high-stake decisions to understand and build trust over ANNs, not many scientists have used it to enrich scientific knowledge about the target system.

Existing tentative uses of the feature-oriented approach in science start with an ANN trained to process scientific data for a task. Next, by analyzing the network parameters or using XAI algorithms, scientists identify the features that the network employs to complete the task. Scientists then try to correspond them to features of the target system. The approach is used by neural scientists to find unknown indicators of Alzheimer's disease in brain fMRI images (Eitel and Ritter 2019), by climate scientists to detect physical drivers of climate change, and by cosmologists to detect signatures of certain ranges of cosmological parameters in lensing maps. In practice, scientists do not only take them as heuristics but also endow them with a justificatory role, so that the features are taken with some evidentiary power.

Next, we analyze this approach following three questions. What are the existing ways that scientists identify features captured by ANNs? How do these features become relevant to science and provide implications? From a critical and normative perspective, can the approach fulfill its expectations, and how should it be best used in science?

3. Features and Non-Uniqueness

In this section, we summarize existing practices of the feature-oriented approach. Its central step is to correspond features identified from ANNs to features of the target system. We provide a taxonomy and analysis of what scientists call a 'feature' in this process. We introduce parametric and diagnostic features according to how they are identified from ANNs and point out their fundamental differences with real-world features in the target system. We further argue that non-uniqueness is present in identifying the former and interpreting them into the latter[2].

3.1. Parametric Features

A neural network effectively performs a combination of linear and non-linear transformations on input variables. On each layer, these transformations map the input data into a new space in which they obtain a new configuration, so that the reconfigured data points on the final layer can be more easily classified (Fig. 2) or fit into a regression function. Within this context, computer scientists call the space that each layer represents data a 'feature space' and use 'feature' to refer to its dimensions (Bishop 2006, 192-195). On each layer, every node transforms the input through a different function and constitutes one dimension of this feature space, thus representing a parametric feature.

Take an example of a convolutional neural network (CNN) trained to classify pictures of animals into two labels: dogs and cats. CNNs mimic the biological visual system in that the layers capture input features hierarchically. After training, nodes on lower layers typically learn to represent simpler features like edges, and those on deeper layers represent larger features like motifs. One unit on the first convolutional layer, for example, represents a 45-degree slash by multiplying the input with a weight matrix (aka. kernel) that highlights regions consisting of the slash (Fig. 3). The numerical

[2] As features are always relative to a system and a task, different features identified from networks trained for different tasks are not problematic. Our discussion is constrained to ANNs trained with similar data for the same task.

configuration of the convolutional kernel is thus a parametric feature. Likewise, one can also obtain more complicated parametric features by extracting the functions applied by nodes on deeper layers.

[Figure 2]

[Figure 3]

Parametric features most faithfully reflect how the network manipulates input data, with the trade-off of being unintelligible. A network is highly distributed and holistic (O'Reilly et al. 2012). No one unit reacts uniquely to one property in the inputs, and each property can activate multiple units. Moreover, weights in a network are also contingent upon the training process. Many designs of the training process are not directly aimed at extracting intelligible or stable statistical patterns but are guided by the practical goal of balancing accurate mapping between the given data and their labels, on the one hand, and avoiding overfitting, on the other. The learned weight thus relies on multiple practical choices, such as the loss function, the length of the training period, and how one splits training and testing data (Bishop 2006). With this, a parametric feature may not be the statistical regularity in the input data but an 'expedient' choice that optimizes network performance under a training condition.

As a result, to have comprehensive knowledge of parametric features and their roles, one may need to analyze all units in the network and the training condition. However, the deep network configuration, non-linear transformations, and complex training history prevent such an analysis. Additional diagnostic strategies are needed to summarise these complexities.

3.2. Diagnostic Features

Diagnostic features emerge when scientists apply algorithms and methods external to ANNs, especially those developed in XAI, to summarise the behavior of the network. Because XAI is a developing domain home to ongoing debates and renewing algorithms (Lipton 2018), here we do not give an exhaustive definition of diagnostic features by specifying a set of legitimate XAI methods or processes of interpretation. In practice, what scientists identify from XAI methods as a feature is often a matter of agreement. Therefore, we take a naturalistic approach to explicate how scientists obtain features from XAI methods.

Two examples of XAI strategies are *saliency maps* and *activation maximization*. Saliency maps are images of the same dimension as the input. Each pixel in the map represents the importance of that pixel for a given output according to a certain algorithm for measuring importance. Activation maximization seeks to synthesise images that trigger the highest activation of one designated unit. When the unit is an output unit, this technique effectively shows the prototypical input for a label.

Diagnostic features are patterns found in the products of XAI strategies. In the earlier example of the CNN classifier for dogs and cats, one finds diagnostic features not by scrutinizing the network parameters, but by recognizing patterns from products of XAI strategies applied to the network. A saliency map may highlight pixels around the nose and ears of a cat in the input, together with some scattered pixels elsewhere that do not form a noticeable pattern. Here, scientists typically take the patterns of the nose and ears as features. Likewise, a synthesised input that maximizes the dog label may show repeating patterns of a dog's mouth, and scientists typically take the pattern of mouth as a feature. Note that identifying such patterns is already making interpretations of the products of XAI algorithms, which are technically only pixels of the input dimension. Scientists dismiss unintelligible pixels or motifs and only take patterns meaningful to them as features[3]. Diagnostic features are thus relational in a more complicated way. They provide intuitive summaries of parametric features according to XAI algorithms and human interpretation.

Diagnostic strategies are not unique but many. Many of them show discordant results, but there is no agreement on how to choose between them (Samek and Müller 2019). Network diagnosis involves a trade-off between faithfulness to network parameters and intelligibility to human subjects, and no algorithms can satisfactorily realize both aspects (Rudin 2019). For example, different methods to produce saliency maps involve different trade-offs between practical considerations regarding whether the highlighted parts agree with human intuition, on what existing networks they have shown stability, and whether they also support other techniques (Montavon et al. 2018). Products of activation maximization also differ in whether they appeal to human intuition, whether they faithfully present patterns unintelligible to the human eye, and whether they specifically activate one unit without activating others (Nguyen et. al. 2019). The variety of diagnostic strategies and their results have been reported by many practitioners as a source of confusion. Even if some of the features are not strictly competing, their multiplicity makes the approach less informative about what one should focus on in practice.

Choosing between XAI algorithms and diagnostic features faces additional difficulties in the scientific context. Computer scientists tune and evaluate these

[3] These interpretations are still a step away from identifying physical reality: diagnostic features are patterns that appear approximately as nose or ears but do not necessarily point to actual nose and ears. The network might be sensitive also to the environmental pixels around the ear, which are part of the ear pattern but not of a real ear.

algorithms on datasets of mainly everyday-life images. It is relatively intuitive there what diagnostic features are more pertinent, such as the wings of birds as opposed to the branches they are standing on. However, which diagnostic features are pertinent may not be straightforward for scientific images in immature domains, where scientists do not know what features they are looking for.

3.3. Real-World Features

What scientists want to discover with the approach are real-world features. By real-world features, we mean stable features of the real system that are *characteristic* of a class of phenomena and are *meaningful* to the relevant scientific domain. 'Characteristic' means that the features are not coincidental similarities only in the sample but are distinctions between classes of phenomena in the world. By 'meaningful' we mean that the features can be connected or explained with concepts or theories in the relevant scientific domain. In the earlier example of a CNN for classifying dogs and cats, the plants in the background and the general color of the fur would not count as real-world features, even when they appear as diagnostic or parametric features. Instead, the shape of the face or paws are real-world features for this task. This is because they are more characteristic of the distinction between dogs and cats and align with the common scientific practices of classification based on skeleton shape.

We take real-world features as elements in scientific representation and do not assign them any ontological weight. They do not have to indicate real entities, natural kinds, or causal regularities. As long as scientists recognise them as stable and pertinent elements in the target system, they could be either a mixture of the above categories or a more moderate version thereof. This concept is not aimed at stipulating what a feature in science should be, but rather to illustrate the distinction between features already meaningful to scientific practices and those extracted by networks.

There are some fundamental differences between features captured by networks and real-world features. For a start, real-world features are not relational to ANNs. In contrast, parametric features are also partly features of a trained ANN, and diagnostic features are further characteristic of an XAI algorithm.

Moreover, real-world features are constructed by scientists based on their research activities, refer to something in the real world, and are generalizable to a broader scope of phenomena. In contrast, parametric and diagnostic features are determined by algorithms, entailing real-world components only by human interpretation, and often not generalizable to situations very different from the training dataset. The weights of a unit are only a set of numbers that, when combined with all the network parameters, achieve the optimal outcome during the training process. A highlighted pattern on a saliency map only suggests how numerical changes in an input affect the output. These are very different from how scientists find real-world features. Scientists do not assign weights to individual features and calculate the outcome by combining all of them, but they reason with qualitative language. They do not carve their subjects in any way that optimises the ability to distinguish them but often introduce intuitive, conventional, and theoretical reasons. The contrast between these features is also analogous to that between data and the target system[4]. Machine-captured features may not have real-world counterparts, and features relevant to a task for ANNs may not be relevant to the task in the target system for human scientists. For example, ANNs may learn to be sensitive to certain pixels of a galaxy picture to optimise their

[4] We thank Claus Beisbart for suggesting this analogy.

classification performance in a dataset, whereas real-world features are the real galaxy parts that are connected to certain physical mechanisms or morphological characteristics that scientists know about.

Scientists thus need to interpret machine-captured features to connect them to real-world features and generate knowledge about the latter. However, as Boge (2022) suggests, there is no guarantee from machine learning and XAI algorithms for successful interpretation, but 'the *translation* of these into scientific concepts is up to scientists' ability and knowledge' (70). In the next subsection, I analyze the interpretation process in the feature-oriented approach and present its difficulties.

3.4. Non-Uniqueness

By non-uniqueness, we mean the existence of multiple equally applicable mathematical structures or algorithms to characterize or represent a target system for a scientific purpose, together with the epistemic difficulty in justifying the choice between them.

An analogy can be made between non-uniqueness and underdetermination. Proponents of contrastive underdetermination argue that there may exist multiple empirically equivalent theories for a given set of evidence (Stanford 2017). Because evidence alone cannot arbitrate the choice among alternative theories, the burden finally lies on other factors like conventions, contexts, and non-epistemic values.

Non-uniqueness occurs in the application of mathematical structures or algorithms for a target system and a scientific task. Mathematical structures often serve as the framework to assemble and derive propositional descriptions. Algorithmic tools specify how concepts are operationalized and measured in a system, such as using a network model to define and measure the degree of 'connection' between subjects or using Euclidean or non-Euclidean metrics to measure the similarity between objects in the feature space. These mathematical characterizations can also be ‘underdetermined’ if their alternatives can be used to establish a system of concepts and propositions that are similarly applicable to the target system for similar purposes.

Non-uniqueness has been discussed in various domains. For example, in modern physics, scientists can choose between multiple geometrical metrics for spacetime, and there is no uniquely correct choice either *a priori* or by empirical test (Heinzmann and Stump 2017; Hagar and Hemmo 2013). Non-uniqueness also occurs in the construction of mathematical models for a system or dataset. For example, one can develop multiple equally applicable network models for a system of connecting subjects, such as mobile phone users. These models can differ in whether their connections are characterized as one-way or reciprocal, or whether they need to meet a threshold value to count, but such differences may not affect how the models fulfill the present purpose of investigation (Baker 2013). One can also develop multiple algorithms for similarity measures in unsupervised data clustering (Cat 2022), or multiple different machine learning models trained from the same data for the same task (Li 2023).

These existing discussions also show that choosing between non-unique mathematical characterizations or models can face epistemic difficulties. There are often no *a priori* principles that favor one option over another, nor can they be directly tested by evidence. Because they are frameworks or operationalizations of propositional descriptions, those propositions are often amended first if discrepancies with evidence occur. Like in underdetermination, when multiple mathematical characterizations appear to be equally applicable to the system for a task, the decisions between them would often be subjective or conventional.

Non-uniqueness is not by itself problematic, especially when the different mathematical characterizations highlight different aspects of the system. However, failing to recognise the non-uniqueness of options can often lead to problematic consequences. First, one can be blind to alternatives when employing chosen characterizations. In practice, alternatives can provide discordant implications about the target system, and one may fail to recognize these alternative implications. Second, one may falsely attribute an evidential significance to the result of one characterization when it is falsely deemed to be 'the unique, objective tool'. For example, in the case offered by Hagar and Hemmo (2013), proponents of dynamical approaches to spacetime claim to *derive* the Riemannian metric from non-geometrical quantum gravity consideration, while in fact, this non-unique derivation relies on a contingent presupposition about counting. Here scientists claim to 'infer' a unique structure out of raw evidence, while what they perform is actually an *interpretation* of the evidence, which presupposes a desired non-unique theoretical framework.

Non-uniqueness deserves special attention, especially in the era of data-driven science. With the proliferation of metrics, algorithms, and mathematical models, failing to notice the multiple options and be acquainted with their respective properties can bias science with contingent choices.

Non-uniqueness of the feature-oriented approach appears as the multiplicity of parametric and diagnostic features. To uncover real-world features, one needs to choose from numerous network parameters, multiple diagnostic strategies, and several ways to interpret them. As we will show in the next section, a series of problems can arise from non-uniqueness. For a start, if one contingently adopts some features without realizing more options and their comparable applicability to the system for the task, one may be blinded by these choices or even go further to falsely attribute evidential value to them.

Second, even if scientists become aware of the multiplicity of machine-captured features, it is often unclear how one can identify those that point to genuine real-world features without being lost in numerous futile ones. Then, if multiple real-world features can be interpreted out of machine-captured features, it is still unclear which ones are significant and which are trivial for a task. Simply reporting multiple machine-captured features without interpreting them into a few real-world features and ascertaining their significance does not make up good scientific research, as it may fall trivially into "anything matters" and fail to contribute to novel discoveries as expected.

To make a choice between all these options and evaluate their significance, scientists need to put forward their rationalization with the help of background knowledge. A desired rationalization should show that *those network-captured features relate to genuine real-world features* instead of just being coincidental or unintelligible patterns. According to Sullivan (2019), this is usually achieved by linking model parts with stable elements in the target system via existing theories or empirical evidence that supports the correspondence between the two.

However, because the feature-oriented approach is expected to generate new knowledge, they are taken in immature domains that lack sufficient background knowledge. For example, scientists may not have established the vocabulary to describe certain phenomena or developed relevant theories. While scientists expect real-world features to *be constructed out of* or *warranted by* parametric and diagnostic features, they do not have sufficient knowledge about what those real-world features should be like. Rationalization, therefore, often cannot justify the correspondence between machine-captured features and any genuine real-world features.

Thus, the embarrassing situation of the feature-oriented approach is that while ANNs are expected to provide novel real-world features, to justify the choice between many network-captured features requires that one already knows the real-world features being pursued. We are not suggesting that this approach cannot contribute to scientific discovery at all, but this problem suggests a more realistic picture of what it can contribute to science.

4. Searching for Features with CNNs in Cosmology

We illustrate our analyses and arguments with a concrete case, showing how machine-captured features can be non-unique, how scientists make choices and rationalize them without sufficient background knowledge, and how this can lead to problems.

According to general relativity, light rays emitted from faraway galaxies are distorted by the gravitational field between them and the observer. On cosmic scales, this distortion is called weak lensing. Cosmologists measure this distortion and produce lensing maps, which are approximately 2D projections of the 3D matter density of the universe. Weak lensing offers evidence for the Lambda cold dark matter (ΛCDM) cosmological model. ΛCDM describes the expansion of metric space based on three components of the universe: dark energy, cold dark matter, and ordinary matter. The statistical properties of lensing maps are sensitive to two parameters of ΛCDM: the mean matter density of the universe, and the amplitude of the initial perturbations that served as seeds for the cosmic structure growth. In our case, cosmologists use CNNs to process lensing maps and constrain the values of these two parameters.

Traditional statistical methods such as power spectrum, peak counts, and Minkowski functionals have allowed cosmologists to constrain the parameters with summary statistics of the image. However, these general methods fail to capture features specific to lensing maps (Matilla et al. 2020). Because CNNs map the entire image directly to the output, they could extract more statistical information from data than traditional statistics and further constrain the two parameters.

In these studies, CNNs are trained with computer-simulated lensing maps as the input. These inputs are labeled with binned cosmological parameters that serve as the initial condition of those simulations. CNNs can predict cosmological parameters from lensing maps with higher precision than any existing statistical methods.

4.1 Why the Feature-Oriented Approach?

Beyond taking CNNs merely as black-box models to constrain the parameters, cosmologists expect the feature-oriented approach to contribute several aspects of novelty.

First, it can supplement an immature domain like weak lensing with the vocabulary to describe phenomena. Lensing maps are visual representations of matter distribution and are highly unintuitive. How would one describe a distribution of matter? What parts or patterns of it constitute the vocabulary to express its properties? Among numerous possible combinations of pixels that can summarise a picture, neural networks might be able to single out those corresponding to physically important features in the target system.

Moreover, features could be taken away from the networks to refine traditional statistical methods. Because of the widely known weakness in extrapolation, networks trained on simulation images may not be directly applied to real lensing maps that involve different types of noise. To take advantage of the local accuracy of CNNs while circumventing the problem of extrapolation, features could be extracted and used ‘as a hint’ to ‘build an easy-to-understand and robust estimation method’ (Ribli et al., ‘Inference Scheme,’ 2019, 93).

Finally, the feature-oriented approach also enables the analysis of systematic errors. Systematic errors can only be assessed by comparing multiple methods that extract non-identical features from data. By knowing what statistical features a CNN utilises, one can assess whether it processes data in a way independent of other existing methods. If there are discrepancies in the prediction between CNNs and traditional statistics, and CNNs do use a different set of features, they may indicate systematic errors in traditional statistics.

4.2 Non-Uniqueness, Selection, and Rationalization of Features

Cosmologists employ different methods to obtain features within a well-trained CNN.

Ribli et al. ('Inference Scheme,' 2019; 'Noisy Data,' 2019) started by inspecting the parametric features of the network, especially the convolutional kernels. As we illustrate in section 3.1, a kernel is a matrix of weights that is applied to pixels in the input image to highlight a pattern. A deep CNN has multiple convolutional layers. Each node on those layers applies a different kernel to the image to capture a pattern. Ribli et al. studied the first convolutional layer among multiple in the network and focused on 3 kernels among 32 on that layer.

To justify this choice, they need to explain why the first layer is important, what role those three kernels play in the network performance, and how they map onto any real-world features. First, they rationalise the choice of the first layer with assumptions about the physics underlying lensing maps and the properties of CNNs. They assume that what traditional statistical methods miss but CNNs capture are the small-scale patterns related to gravitational collapse. They also assume that small-scale patterns are captured mostly by the first convolutional layer. Then, they use a saliency method called layer-wise relevance propagation to identify diagnostic features ('Noisy Data,' 2019). This method highlights the high-signal peak areas. Interpreting from this, Ribli et al. believed that small-scale patterns around peaks are important. They further interpret that the convolutional kernels that are sensitive to the signal gradient around peaks are crucial to the network performance.

Ribli et al. further rationalise the choice of 3 kernels among all 32 based on their morphological similarity with known filters for processing other physical images. One kernel resembles a '2D discrete Laplace operator', which 'calculates the difference of the peaks and the surrounding pixel values', and others resemble the 'Roberts cross kernels', which 'approximate the gradient of an image' ('Inference Scheme', 93-94).

Beyond this, they suggest that the kernels may map onto certain previously hypothesized real-world features. Ribli et al. ('Noisy Data,' 2019) invoke existing simulation-based studies that suggest the correlation between peak steepness and cosmological parameters, indicating that peak steepness is likely a real-world feature for the task of inferring cosmological parameters with weak lensing. Because the 3 chosen kernels highlight peak steepness, they are suggested to map onto real-world physical regularities.

These rationalizations should not conceal that first, the chosen features are not unique, and more features could have been considered. Second, the choices are made under certain background assumptions and could have been different under other assumptions. Indeed, another study by Matilla et al. (2020) identifies pixels from non-peak areas as containing more statistical information with several other diagnostic strategies[5]. The emphasis on voids differs even from conclusions of mainstream

[2] This result from saliency maps conflicts with what is suggested by the algorithms used by Ribli et al. The conflict indicated by algorithms, however, does not dismiss the

traditional statistical methods, but Matilla et al. (2020) could rationalise this result by referring to existing studies that favor a different hypothesis stressing the cosmic voids. Based on those studies, they suggest that statistical properties of voids are more significant real-world features under certain physical assumptions:

> Large voids […] have previously been found to contain most of the cosmological information in simulated maps […]. These regions have also been shown to be less affected by baryonic physics, which are hard to capture accurately in simulations of growth of structure. On the other hand, these regions have been shown to be sensitive to neutrino physics and modified gravity theories. (11)

4.3. Limitations of Rationalization and Persistence of Non-Uniqueness

In our case study, non-uniqueness occurs as both research projects could have studied more features than those analyzed. Ribli et al. selected only 3 out of 32 convolutional kernels on the first layer of the network and only applied one diagnostic strategy. Matilla et al. selected only several diagnostic strategies. Moreover, because peak and void areas are continuous, there is no unique way to draw the line for those areas and no unique feature to identify.

The epistemic difficulty is shown in the insufficiency of cosmologists' background knowledge for justifying the choice of features and their physical reality. Ribli et al.'s assimilation of parametric features to existing kernels in image processing does not count strictly as a justification out of physical regularities. Without an explicit theoretical basis, the use of certain kernels may only be a mathematical maneuver.

possibility that both voids and peaks contain important statistical information and could be used together in some way to constrain the cosmological parameters. This suggests that further work needs to be done beyond simply adopting the results of algorithms.

Likewise, Matilla et al. only draw a loose connection between the manually delimited high-saliency regions and the relevant statistics proposed by existing studies they cite.

Furthermore, existing hypotheses in the domain also fail to justify the choices. In the above case, two hypotheses, one about the significance of peaks for predicting cosmological parameters and the other about the physical mechanisms that make voids crucial to such predictions, are referenced to select machine-captured features. However, due to the immaturity of the domain, scientists do not have agreed-upon knowledge about the credibility of these hypotheses. These hypotheses also do not suggest what specific patterns in the data contain the relevant statistical information, what the relevant real-world features specifically are, and whether statistical information from different regions is competing or complementary. One should thus be cautious about what the features identified from the approach can imply. As different features are chosen in light of different existing hypotheses, it would be a circularity if one takes the identified features as *evidence* in support of the hypothesis used to single them out. When used this way, the feature-oriented approach cannot support the hypothesis, and the hypothesis also cannot justify the choice of features.

Therefore, by reducing the number of plausible choices and suggesting potential connections to existing hypotheses, rationalization only temporarily steps away from non-uniqueness, but it does not provide epistemic security for those chosen features to genuine real-world features relevant to completing the task.

5. Instrumental Value, Transparency, and Plurality

Non-uniqueness of the feature-oriented approach is not unique to cosmology but also appears in other domains of science such as medicine (Eitel and Ritter 2019). In this section, we provide prescriptions for the legitimate use of the feature-oriented approach.

We suggest that the commitment to features should be instrumentalist rather than evidential, no matter how tempting the latter is. This means that features interpreted from ANNs should only be used for inspiration and guidance toward further studies, but not as concrete evidence to settle the reality of relevant real-world features or the credibility of hypotheses.

We also identify several possible instrumental uses of the approach from scientists' expectations and practices in our case: (1) integrating features into traditional methods and exploring their contribution to it, (2) referencing existing theories under dispute and highlighting promising ones for further study, (3) using features to serve as descriptors on the phenomenal level, and (4) suggesting possible sources of influences on ANN performances and sources of errors. While the last two have not been fully developed in weak lensing and remain legitimate expectations, our case demonstrates the power of the first two.

The first use appears in the successful integration of features into traditional methods by Ribli et al. ('Inference Scheme,' 2019). The filters are manually applied to lensing maps before applying traditional statistical methods. Extracting features from the network circumvents CNNs' potential problems with extrapolation and makes the features independently applicable. Moreover, with knowledge about the properties of different kernels, scientists can make modifications to them for different situations and goals. Despite doubts about whether ANNs capture real-world features, this further empirical success testifies to the usefulness of the features.

Secondly, the feature-oriented approach can highlight existing theories for further study and bring debates to the fore. Although the relation between the statistical information from peaks and voids is unsettled, the divergent features help to expose the community to the existence of such alternatives and highlight the hypotheses that stress different regions.

Lessons from the literature on non-uniqueness suggest the tricky consequence of taking a choice as unique and well-justified. This can also happen to the feature-oriented approach. Individual scientists may fail to notice the non-uniqueness of features, because they may not know all possible background assumptions and hypotheses in the domain or be capable of applying all diagnostic strategies. They also need to make a tradeoff between identifying many possible features and giving constructive suggestions about a few. The risk here is that rationalization can be interpreted as justification, and individual choices may blind the community from alternatives.

To mitigate this risk, we suggest transparency and plurality of feature selection and rationalization. Facing the unavoidable underdetermination of theories by evidence, Longino (1990) proposes that to retain the objectivity of science, it is crucial to ensure that a community performs critical scrutiny upon the background knowledge that is involved in individual research. This lesson can be transferred to addressing non-uniqueness in the feature-oriented approach. First, scientists should be transparent about the way they extract and rationalise features, expose what alternative features are omitted and what theories are invoked, and how neatly and uniquely the network-captured features can be interpreted into real-world features. Second, the importance of plurality is shown by the case study, as the co-existence of different results demonstrates to the community how many different real-world features may be rationalised to a similar extent. If the community can encourage scientists to use multiple diagnostic strategies and rationalize multiple features with competing theories, it would not easily be biased by contingent individual choices based on limited background knowledge.

6. Conclusion

ANNs have shown distinct strengths as an efficient tool in scientific research. A question thus arises as to whether their opaque statistical inferences and automatic reasoning could meet the needs of science, which is a domain originally guided and delimited by human cognitive abilities, values, and conventions. Answers to this question can illuminate constructive uses of ANNs in scientific practices.

The feature-oriented approach emerges from recent developments in XAI and the anticipation that ANNs may serve a creative role in identifying novel features of the target system for a scientific task. We analyze how this approach is performed, evaluate what can be reasonably expected from it, and highlight potential caveats. We distinguish three types of things that scientists call a "feature" and identify an issue of non-uniqueness: multiple alternative parametric and diagnostic features could stand out when scientists interpret them into features of the target system. With our case study in cosmology, we demonstrate that non-uniqueness cannot be eliminated by rationalizing and interpreting the features with existing background knowledge when one aims at novelty in an immature domain.

If our arguments hold in a broader scope, there is good reason to be critical and cautious about the approach. It is unlikely that ANNs will automate scientific discoveries by giving solid evidence and overcoming inductive uncertainties, but scientists are still the decision-makers in important steps. To guard science against over-enthusiasm about ANNs and mitigate the potential risks from non-uniqueness, we argue that scientists should take the feature-oriented approach only instrumentally, instead of using them evidentially to justify existing hypotheses or provide foundation for new theoretical frameworks. When using the approach, scientists need to be transparent about the multiplicity of features and on what ground they are rationalised and adopted. The community also needs to promote plurality by exploring various ways of feature identification and interpretation with competing theories.

References:

Adadi, Amina, and Mohammed Berrada. 2018. 'Peeking Inside the Black-box: A Survey on Explainable Artificial Intelligence (XAI).' *IEEE Access* 6: 52138-60.

Baker, Alan. 2013. 'Complexity, Networks, and Non-Uniqueness.' *Foundations of Science* 18 (4): 687-705.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. New York, NY: Springer-Verlag.

Boge, Florian J. 2022. 'Two Dimensions of Opacity and the Deep Learning Predicament.' *Minds & Machines* **32**, 43–75.

Buckner, Cameron. 2018. 'Empiricism without Magic: Transformational Abstraction in Deep Convolutional Neural Networks.' *Synthese* 195 (12): 5339-72.

Cat, Jordi. 2022. 'Synthesis and Similarity in Science: Analogy in the Application of Mathematics and Application of Mathematics to Analogy.' In *Words and Worlds: Use and Abuse of Analogies and Metaphors with Sciences and Humanities*, edited by Shyam Wuppuluri and Anthony C. Grayling. Springer, Synthese Library.

Eitel, Fabian, and Kerstin Ritter for the Alzheimer's Disease Neuroimaging Initiative (ADNI). 2019. 'Testing the Robustness of Attribution Methods for Convolutional Neural Networks in MRI-Based Alzheimer's Disease Classification'. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. ML-CDS 2019, IMIMIC 2019. Lecture Notes in Computer Science, vol 11797*, edited by Suzuki K. et al. Springer, Cham.

Hagar, Amit, and Meir Hemmo. 2013. 'The Primacy of Geometry.' *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 44 (3): 357-64.

Heinzmann, Gerhard, and David Stump. 2017. ‘Henri Poincaré.’ In *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2017/entries/poincare/.

Leonelli, Sabina. 2012. ‘Introduction: Making Sense of Data-driven Research in the Biological and Biomedical Sciences.’ *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43 (1): 1-3.

Li, Dan. 2023. ‘Machines Learn Better with Better Data Ontology: Lessons from Philosophy of Induction and Machine Learning Practice.’ *Minds & Machines* 33: 429–450.

Lipton, Zachary C. 2018. ‘The Mythos of Model Interpretability: In machine learning, the Concept of Interpretability is both Important and Slippery.’ *Queue*, *16*(3), 31-57.

Longino, Helen E. 1990. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton, NJ: Princeton University Press.

Matilla, José M. Z., Manasi Sharma, Daniel Hsu, and Zoltán Haiman. 2020. ‘Interpreting Deep Learning Models for Weak Lensing.’ *Physical Review D* 102 (12): 123506.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. 2018. ‘Methods for Interpreting and Understanding Deep Neural Networks.’ *Digital Signal Processing* 73: 1-15.

Nguyen, Anh, Janson Yosinski, and Jeff Clune. 2019. ‘Understanding Neural Networks via Feature Visualization: A Survey.’ In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, 55-76. Springer, Cham.

O'Reilly, Randall C., Yuko Munakata, Michael J. Frank, Thomas E. Hazy, and Contributors. 2012. *Computational Cognitive Neuroscience*, 4th ed. https://compcogneuro.org/.

Pietsch, Wolfgang. 2015. 'Aspects of Theory-ladenness in Data-intensive Science.' *Philosophy of Science* 82 (5): 905-16.

Räz, Tim. 2022. 'Emergence of Concepts in DNNs?' *arXiv preprint.* arXiv:2211.06137.

Ribli, Dezső, Bálint Ármin Pataki, and István Csabai. 2019. 'An Improved Cosmological Parameter Inference Scheme Motivated by Deep Learning.' *Nature Astronomy* 3 (1): 93-98.

Ribli, Dezső, Bálint Ármin Pataki, José Manuel Zorrilla Matilla, Daniel Hsu, Zoltán Haiman, and István Csabai. 2019. 'Weak Lensing Cosmology with Convolutional Neural Networks on Noisy Data.' *Monthly Notices of the Royal Astronomical Society* 490 (2): 1843-60.

Rudin, Cynthia. 2019. 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.' *Nature Machine Intelligence*, 1(5), 206-215.

Samek, Wojciech, and Klaus-Robert Müller. 2019. 'Towards Explainable Artificial Intelligence.' In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, edited by Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, 5-22. Springer, Cham.

Stanford, Kyle. 2017. 'Underdetermination of Scientific Theory.' In *The Stanford Encyclopedia of Philosophy* (Winter 2017 Edition), edited by Edward N. Zalta.

https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/.

Sullivan, Emily. 2019. ‘Understanding from Machine Learning Models.’ *The British Journal for the Philosophy of Science*: axz035.

Zednik, Carlos, and Hannes Boelsen. 2022. ‘Scientific Exploration and Explainable Artificial Intelligence.’ *Minds & Machines* **32**, 219–239.

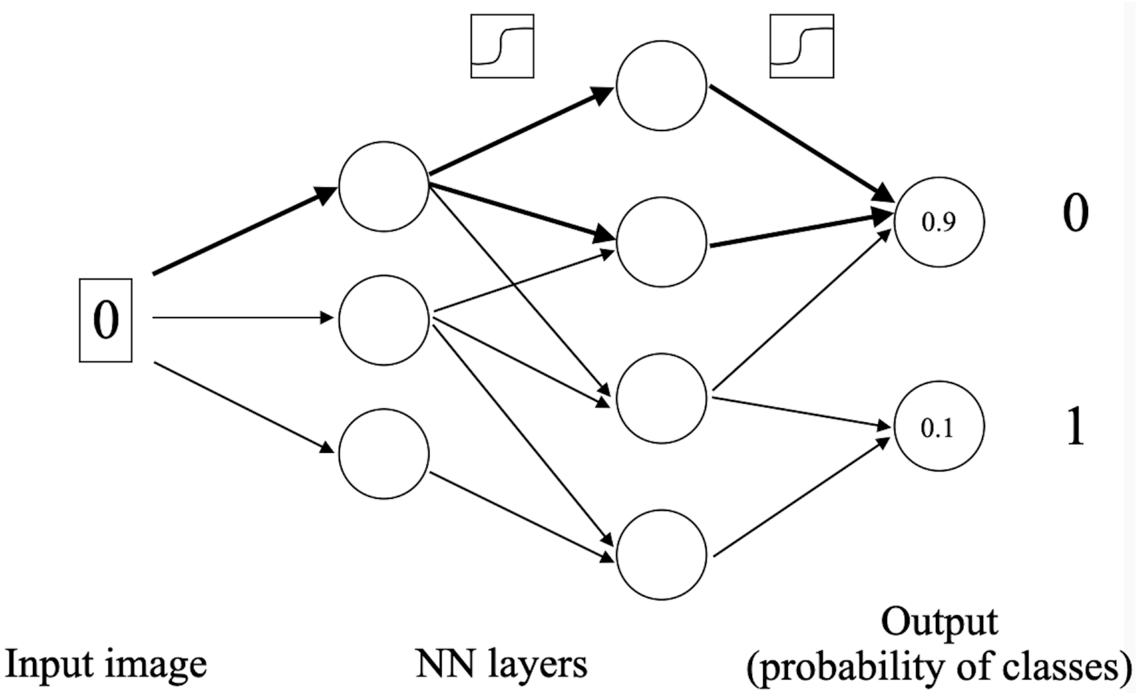

Figure 1. An illustration of the structure of an ANN.

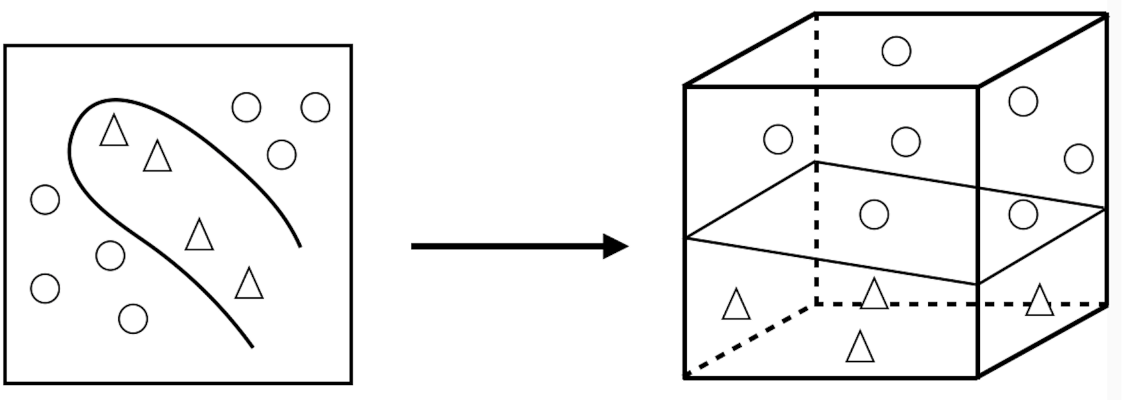

Figure 2. The data in the input space is transformed into a feature space. Linear classification can now separate two categories of data.

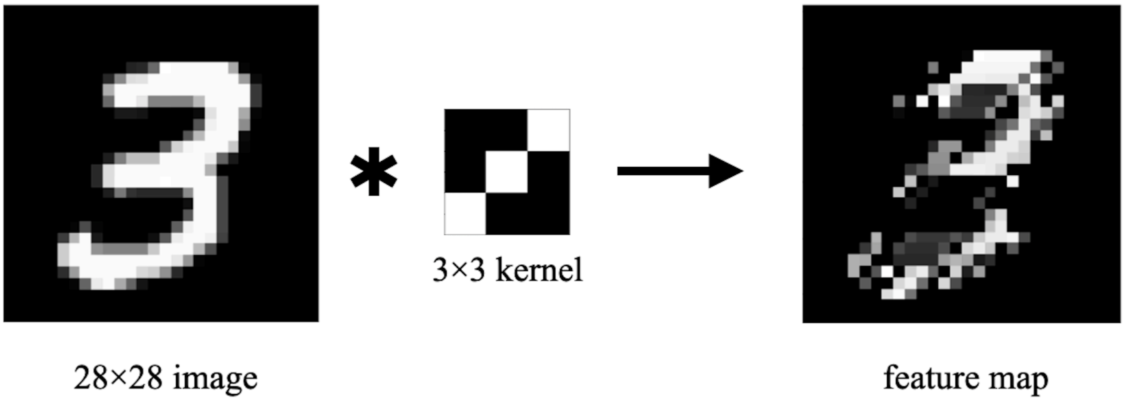

Figure 3. An illustration of how a kernel generates a feature map of a slash.