

Mapping Knowledge Representations to Concepts: A Review and New Perspectives

Lars Holmberg,^{1*} Paul Davidsson,¹ Per Linde²

¹ Department of Computer Science and Media Technology

² School of Arts and Communication

Malmö University, Sweden

lars.holmberg@mau.se, paul.davidsson@mau.se, per.linde@mau.se

Abstract

The success of neural networks builds to a large extent on their ability to create internal knowledge representations from real-world high-dimensional data, such as images, sound, or text. Approaches to extract and present these representations, in order to explain the neural network's decisions, is an active and multifaceted research field. To gain a deeper understanding of a central aspect of this field, we have performed a targeted review focusing on research that aims to associate internal representations with human understandable concepts. In doing this, we added a perspective on the existing research by using primarily deductive nomological explanations as a proposed taxonomy. We find this taxonomy and theories of causality, useful for understanding what can be expected, and not expected, from neural network explanations. The analysis additionally uncovers an ambiguity in the reviewed literature related to the goal of model explainability; is it understanding the ML model or, is it actionable explanations useful in the deployment domain?

Introduction

Digitalisation influences all parts of society. Even the meaning of the central philosophical term episteme has shifted from being best translated as knowledge, towards being best translated as understanding (Grimm 2021). The shift can be exemplified by an ever-present weather app that knows if and when it will rain but has no understanding, in a human sense, concerning why it rains or the subjective implications for an individual human being. Focus in this work is then on the tension between the prospective human understander and the third-person objectivising stance characteristic of the natural sciences (Grimm 2016), here represented by Artificial Intelligence (AI) and in particular Machine Learning (ML).

Explanations, typically answering a *Why* or *What if* question, is a common human way to bring understanding of the natural world or other people (Grimm 2021). Bridging the gap, between information processing systems, like AI/ML, and human understanding is important since AI/ML increasingly affect us and our decisions (Coudry and Mejias 2019; Webb 2019; O'Neil 2016; Vallor 2016).

*This work was partially financed by the Knowledge Foundation through the Internet of Things and People research profile. Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

We here view contemporary ML as limited to local generalisation within a single task or well-defined set of tasks that only holds when the training data used is independent-and-identically-distributed (i.i.d). ML is then limited when this does not hold or when it comes to causal inference and out-of-distribution (o.o.d) generalisation (Chollet 2019; Scholkopf et al. 2021).

The human capability to generalise knowledge builds, as a contrast, on that we can formulate explanations using causal relations and generalise via concepts. Concepts are then building blocks for thoughts, blocks that are connected via relations forming explanations that in turn can bring understanding (Margolis and Laurence 2021; Grimm 2021).

Human understanding and trust in ML concerns not only *understanding* promoted decisions¹, but also, evaluating these decisions in relation to limitations built into the ML model. Limitations are introduced in ML systems by humans during the design phase, for example; what to model, choice of algorithm, feature engineering, training data selection (Gillies et al. 2016). The need for explanations to convey understanding is pronounced in more complex ML models (Lipton 2016) and especially prominent in today's dominating technology: neural networks.

Our approach towards understanding ML decisions builds on connecting human understandable concepts to the ML models knowledge representations with the goal of making them explicable. Below follows an outline of the perspectives on explanations used in this paper.

We use Hilton's (1990) definition that explanations in a human context is a three-place predicate: *Someone* explains *something* to *someone*. A definition that focuses on the explanation as a conversation between the explainer and the explainee.

Additionally, the need for an explanation in a human context is often triggered by an event that is abnormal or unexpected from a personal point of view (Hilton and Slugoski 1986; Hesslow 1988; Hilton 1996). Research in Explainable Artificial Intelligence (XAI) (Barredo Arrieta et al. 2020; Guidotti et al. 2018; Biran and Cotton 2017; Gilpin et al. 2019; Adadi and Berrada 2018; Hoffman, Clancey, and Mueller 2020), on the other hand, are less concerned with

¹The output from an ML system in the form of classification, recommendation, prediction, proposed decision or action

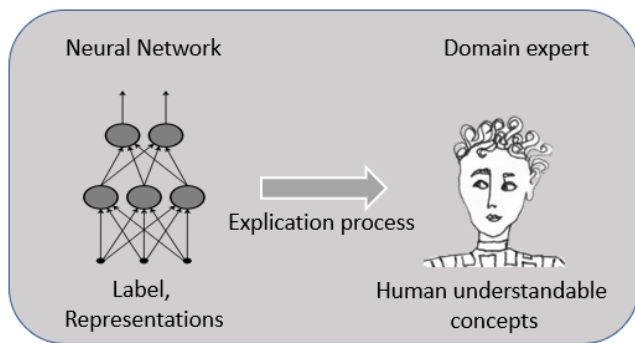


Figure 1: Approach to explainability used in this work

who gives the explanation, to *whom* it is given or *why* it is needed (Miller 2019) and has, in comparison, a more objectivising decontextualised stance to explanations.

We then aim towards a situation where humans, with domain knowledge (henceforth domain experts), can act as explicators and formulate an explanation based on human understandable knowledge representations extracted from the neural network as concepts in line with Hilton's (1990) definition.

Figure 1 is an overview covering the approach to explainability we aim for, which is a system where the neural network presents evidence for a decision in the form of knowledge representations in an explication process. The goal for the domain expert is to associate these representations to Human Understandable Concepts (HUC) and thus deepen their understanding of evidences for the decision and the models capabilities. The human can then be seen as the explaine in Hilton's definition, a person that aims to understand decisions, trust the model and use it to reach some goal.

Focus in the work presented here is a targeted review that lifts out examples of existing literature on methods that aims to extract knowledge representations from neural networks. We are guided by the following research questions:

- How do current methods extract internal knowledge representations in a neural network and map them to human-understandable concepts?
- Can Deductive-Nomological (D-N) explanation taxonomy and causal types of explanations be useful in order to analyse what can be expected, and not expected, from knowledge representations in neural networks?

We answer the first question by organising the methods as global and local to discuss how HUC are induced by humans, either as knowledge priors added in the form of conceptual understanding or, during analysis of the explanans provided by the methods. We also find the D-N taxonomy helpful, and that it, together with explanation types, opens up a generative path that makes it possible to better understand and discuss what we can expect from the type of machine learning we analyse.

The rest of the article is as follows: First, we present a more detailed description related to explanations and concepts, which is complemented by a description of the literature selection process followed by our review. The results

are then deepened using our theoretical approach followed by a discussion related to our results. The work ends with a conclusion section.

Background and foundational concepts

In this work we envision a situation with at least one domain expert that has the capability to understand and value knowledge representations extracted from a neural network. This prerequisite has the advantage of picturing a domain expert in the loop, a person that can be trained in scientific thinking and can relate to scientific explanations. The approach is also useful for persons not trained in scientific thinking since both mundane and scientific explanations aim at answering a *why* or *what if* question, with the difference that scientific explanations, in the D-N case, aim towards objectivity and adds rigour to the answer.

According to Murphy (2016) there are no exemplar theories of concepts. What can be generally agreed on is that concepts are named using a referent, for example, `GOLD`, and that a concept, like `GOLD`, can be grasped by believing that it has some properties as, a specific shine, value or its malleability. Peoples' beliefs, related to concepts and their properties, can be both false and incomplete and, additionally, contain both causal and descriptive factors (Genone and Lombrozo 2012). Ghorbani et al. (2019) deem that concepts, in relation to the domain modelled by a neural network, should be meaningful, coherent and important.

This then implies that the domain expert's understanding of the trained model's behaviour is built on concepts that are derived from internal knowledge representations in a process that can be viewed as parallel to Carnap's (1962) theory of explications. The explication process is then the transformation or replacement of an imprecise concept (explicandum) with new concept(s) (explicatum/explicata). The new concepts adhere to the criteria of being similar to the imprecise concept but more exact, fruitful and simpler (Justus 2012). These Human Understandable Concepts (HUC) can then be Disentangled (HUDC) if they are not confounded and they don't depend on spurious correlations.

For the work presented here, we imagine an explication process that refines and map the label, seen as a concept and internal knowledge representations, to human understandable concepts, with the goal of global understanding related to the model's behaviour. These explicated concepts then aim to bridge the gap between the model's knowledge representations and the fraction of the real world it models.

We use `kind`-related to refer to an abstract class of concepts, for example `SWAN`, `GOLD`, `DOG` or `DOTTED`. We use `entity` to refer to `kind` instances that are concrete particulars existing in time and space. For example, the `kind` concept `ZEBRA` can be explicated and refined by connecting it to the sub-concepts `HORSELIKE` and `STRIPED`. We then in this example, use two `kind`-related sub-kinds to create a causal explanation connected to the `kind` `ZEBRA`. We can then train a neural network using for example labelled images as `data`, so it can classify images containing `ZEBRAS` and `HORSES` as HUDC. If there are a sufficient amount of `data` the network will generalise and be able to classify unseen images picturing `ZEBRAS` correctly.

We can alternatively train a neural network using a core relation between `entities` to separate and classify, for example, individual ZEBRAS. The internal knowledge representations learned by the neural network will then relate to ZEBRA instances and, for example, be explicated as the HUDC SCAR, BLURRED STRIPES and MARE. We denote this type of concepts `entity-related` since they follow the instance.

In this work we, in line with Pearl (2019; 2020), define three types of explanations that answers to different types of *what if*-questions:

- Association that answers to *What if I see?*
- Interventional explanations answers to *What if I do?*
- Counterfactual explanations that answers to *What if I had done?*

At the first level we are only concerned with associations, e.g. regularities in the observation, and no understanding of cause and effect is needed. Interventional and counterfactual explanations builds on imagining alternative outcomes based on counterfactual causes introduced by consciously changing the prerequisites for the decision in question. This requires a causal model of the phenomena, a model that can be used to falsify a claim that make statistical sense. To use a classical example, a causal model that depicts why it is the sun that makes the roster to crow even if the crow precludes the sunrise.

The type of knowledge representations that can be created in a neural network are based on associations between data and a label, a label that then belongs to one of two D-N categories: `entity` or `kind`. The trained ML model is then built using inductive statistical data and can consequently only answer to a *What if I see?*-questions. This question is then answered by presenting the label and accuracy measurement. This can be sufficient in a static well-defined setting, but if the explainee wants a deepened understanding of how sensitive the decision is, for example, concerning the STRIPEDNESS concept, we need to contrast the decision by using alternative input data in the form of counterfactual or semi-factual causes (Akula, Wang, and Zhu 2020; Kenny and Keane 2021). Intervention or *What if I do?*-questions implies this type of doing and relies on causal understanding. On the top rung of Pearl’s (2018, p. 28) ladder of causation are counterfactual explanations (*What if I had done?*), that also builds on causality, but additionally also on a capability to imagine an alternative reality that would have manifested itself if another decision were taken in a given situation.

In this work, we are interested in answering *what if* questions, related to an ML decision and we use Deductive-Nomological (D-N) (Hempel and Oppenheim 1948) explanations to introduce rigour and structure to the answer. As outlined in Figure 1, we also leave it to the domain expert to account for or get insights into confounded features and causal relations.

In line with Overton (2012) we structure D-N explanation using the following categories: `theory`, `model`, `kind`, `entity` and `data`. Neural networks build knowledge inductively using data and labels as a referent to concepts

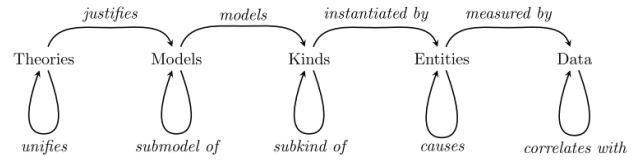


Figure 2: Structure of D-N-explanation, used by permission from Overton.

related to `entities` and/or `kinds`. The model trained in this process is then not a `model` in a D-N sense, since a D-N model is justified using a `theory` that articulate relations between `kinds`.

Overton (2012) outlined a generalised structure for scientific explanations (See Figure 2). Since we use neural networks only the categories `kind`, `entity` and `data` are involved and can be used to build explanations. The deductive part of the explanation is missing and this delimits the explanandum. Instead of a `theory` justifying a `model` the ML model is built using statistical data and we equal this to a law in a D-N sense. The structure of scientific explanation that builds on a core `kind-data` and `model-data` relation can be seen in Figure 3. In those figures the explanandum is denoted quality B and the explanans, that are the evidences for a decision are denoted quality A. A complete D-N explanation (`theory-data`) can be seen in Figure 4.

A majority of the research reviewed in this work uses raw input data in the form of images and we will not to any large degree discuss input data in other forms. In relation to the issue of classification of explanation methods we adhere to (Gilpin et al. 2019) that organise the methods in three categories:

- Methods based on the processing of data, which we denote Feature Based Attribution (FBA).
- Methods based on Internal Knowledge Representation (IKR).
- Methods that aims to automatically create understandable explanations.

We will focus our work on the first two categories and leave it to humans with domain knowledge to create explanations based on FBA and/or IKR explanans. Related to the discussion in the introduction we find it difficult to imagine automatically created explanations valid in a human context in general, without any restrictions related to the domain, the context or, as in our approach presume human explicators with domain knowledge (Se Figure 1).

Well-cited FBA methods are for example Grad-CAM (Selvaraju et al. 2017), LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2016). These methods are local in the sense that they are used to reveal evidence for a decision for a specific input (f.x. an image). LIME and SHAP rely on the creation of a local interpretable substitute model (for example a linear model) using perturbation on input features to infer which features the classification are sensitive for. For images, this approach can be used to grey out areas and expose ‘super-pixels’, e.g. areas

in the picture that the classification is sensitive for. Grad-CAM belongs to a category of methods that uses gradients to attribute model output to layers in the model or to input features. Grad-CAM specifically uses the last convolutional layer and therefore combines high-level semantic information with spatial information. The methods being local does not rule out that they over time, with usage in different situations, can result in a global understanding of the model and trust in its decisions. The analogy here being, for example, trusting a dog interacting with your kids, a trust that are built over time using singular specific situations. Related to concepts, these methods exposes the relation between the internal knowledge representations and a specific image. This implies that for an image that belongs to the assumed i.i.d training data the methods can reveal information on learned representations.

Methods that builds on extracting IKR are global since they reflect the neural networks overall learning process. A network is forced, during training, to learn disentangled representations in each layer in the form of vectors connected via weight matrices, these vectors are then generalisations on some representation level (Bengio, Lecun, and Hinton 2021). It is these generalisations that potentially can map to concepts and they tend to get more complex with the depth of the layers. Therefore more basic concepts like colours, patterns and shapes are represented in the early layers and concepts like GENDER and HORSE in later. An influential method, Concept Activation Vectors CAV (Kim et al. 2018), for extracting knowledge representations, or variants of it, is used in a majority of the reviewed papers. CAV can be used for example to expose images from a training set sensitive to the concept STRIPED or to reveal the correlation between vectors that represent simpler concepts and more complex concepts, for example, between the colour RED and FIRE TRUCK.

For this review we mainly use Webster and Watson (2002) and Knopf (2006) as guidelines. For our review we were interested in a representative selection of relevant research, useful to indicate the applicability of our theoretical approach. The XAI field is a vast area and we experimented with search criterion that were general enough to result in a, for our purpose, useful and manageable selection of research articles. We searched in the abstract, title or keywords in research published between 2018 and September 2021 that has the term ‘understandable’ within three words before the term ‘concept’ and that the abstract, title or keywords contained either ‘neural network’ or ‘deep learning’. By limiting to the search like this we could target papers that had the intended focus and still use more recent XAI methods. The challenge with the search is otherwise that our search terms are to general, especially the concept *concept* that is used in many ML related fields. We searched IEEE Xplore, Scopus, Web of Science and ScienceDirect and got 13 relevant hits. From these we removed work related to mathematically understandable and not human understandable.

In the end, we had nine papers targeting the area. We analysed the papers given our approach and a setting where a human domain expert explicates the internal knowledge representations exposed to create explanations. We then applied

our theoretical lens centred on concepts, the D-N-model and the types of explanations mentioned above. At the review phase, we organised the research concerning the type of explanation the work presented aimed it for.

Review

In this section, we present our review results that build on our targeted search question, theoretical lens and the methodological approach presented earlier.

As the search criterion is formulated the focus is on research that aims to connect internal knowledge representations in neural networks to HUC. Two approaches dominate in the reviewed papers, either unveiling HUC using FBA and local understanding, used in two papers (Wang et al. 2020; Natekar, Kori, and Krishnamurthi 2020), or more directly by using IKR (Yeche, Harrison, and Berthier 2020; Ghorbani et al. 2019; Elshawi, Sherif, and Sakr 2021; Mincu et al. 2021; Chen et al. 2020; Rabold, Schwalbe, and Schmid 2020; Lucieri et al. 2020), used in the remaining seven papers. In both cases, the goal is to better understand the model from a global perspective via concepts. The two approaches are in most papers presented as a dichotomy between global and local understanding which, of course, is relevant if the system is not used over some time, or that a user of the system can experiment with a combination of explainability methods. In Ghorbani et al. (2019) the approaches are to some extent combined in that an initial segmentation of the image is performed and then the image segments are clustered using IKR to calculate the segment’s importance in relation to the concepts they represent. This points towards interesting opportunities in combining methods to infer the best explanation where the explanandum reachable is represented by IKR and the explanans, as evidence for a decision, by FBA.

Related to popular explainability methods: six of the nine papers mention CAV (Kim et al. 2018), five mentions Grad-CAM (Selvaraju et al. 2017), one LIME (Ribeiro, Singh, and Guestrin 2016) and not any mentions SHAP (Lundberg and Lee 2016).

The research reviewed in the medical field (Natekar, Kori, and Krishnamurthi 2020; Lucieri et al. 2020; Yeche, Harrison, and Berthier 2020) sticks out compared to research reviewed in the non-medical field. One central goal in the medical domain is the search for alignment between human concepts and IKR to create trust in the model decisions. In Natekar, Kori, and Krishnamurthi (2020) an alignment between the human concept identification process and the same process in a neural network is highlighted and in Lucieri et al. (2020) an alignment between disentangled concepts identified by the neural network and concepts routinely used by dermatologists is unveiled. In Yeche, Harrison, and Berthier (2020) the consistency of a concept over the layers in the network is exposed.

Explanans useful to underpin a decision in this domain has their base in artificially created 2D images and we can hypothesise that there is a substantial overlap between the explanandum reachable for the ML system and the explanandum reachable for the human.

Explanation type	Core relation			
	theory-data	model-data	kind-data	entity-data
Counterfactual What if I had done?			*Expose temporal factors between correlated concepts (Mincu et al. 2021).	
Intervention What if I do?		*Expose decision tree that containing main concepts used for the decision (Elshawi, Sherif, and Sakr 2021). *Expose first order logic for a decision (Rabold, Schwalbe, and Schmid 2020).	*Expose typical vs -atypical images (Chen et al. 2020; Lucieri et al. 2020). *Relate decision to predefined disentangled concepts (Chen et al. 2020).	
Association What if I see?		*Expose human machine learning epistemic alignment (Natekar, Kori, and Krishnamurthi 2020; Lucieri et al. 2020; Yeche, Harrison, and Berthier 2020). *Expose necessary and sufficient concept attribution (Wang et al. 2020).	*Expose concept-typical images (Lucieri et al. 2020; Ghorbani et al. 2019).	

Table 1: Categorisation and structure of explanations the reviewed articles **aim** for.

The work by Ghorbani et al. (2019) briefly discusses how machine identified HUC can be misleading and include concepts not aligned with human understanding. For example, that the player’s jerseys in basketball is a more important concept than the ball for predicting the sport in question. The work by Wang et al. (2020) lifts similar concerns related to that concepts deemed by the ML system to be *sufficient* and/or *necessary* can be misleading if, for example, images used relates to complex situations or situations not reflected in the training data. The example made in that paper is: If training data for traffic signs only contains stop signs on poles, the pole can be deemed as *necessary*. Consequently, a stop sign without a pole can be classified as a false negative with potentially serious implications. These situations can be viewed as a lack of overlap between the explanandum reachable for the human vis-a-vis the explanandum reachable for the ML system.

In our work, we discuss explanations that build on association, intervention and on counterfactuals. In the work we reviewed, explanation types are not discussed in-depth, instead, they are treated more implicitly as a part of self-evident background knowledge. In our work we are interested in a deepened discussion on the role of explanations in ML to better understand if and how explanation types can be a useful and actionable tool to understand an ML model’s abilities and limitations. An initial step is to arrange the reviewed research in rows, reflecting the different types of *what if*-questions the research targets (See Table 1). Below follows a categorisation of the explanation types, starting with association being least complex and placed at the bottom row in Table 1 thus organising the table in line with Pearl’s (2019) causal hierarchy. Some of the reviewed articles appear in more than one place since they present the use of XAI methods with different goals in parallel experiments.

Association (What if I see?) One example of direct association, is research that exposes images similar to an input image for human comparison to a skin lesion concept using

IKR (Lucieri et al. 2020). Ghorbani et al. (2019) segments images and then uses IKR to align concepts for human comparison. In Wang et al. (2020) an algorithm is created to calculate if a HUC is sufficient and/or necessary for a decision. Three research papers find alignment between how humans and machines learn as their central explanans (Natekar, Kori, and Krishnamurthi 2020; Lucieri et al. 2020; Yeche, Harrison, and Berthier 2020). Natekar, Kori, and Krishnamurthi (2020) sticks out in that the explanans extracted aims at creating trust in the ML decision process by showing that the process is aligned with a human decision process. We arrange these research papers as an association since they do not offer any alternative decision and instead focus on presenting evidence for a human that can be used to increase trust in the system or a decision.

Interventions (What if I do?) In the reviewed research this type of question is addressed by: exposing typical and atypical images related to a concept (Chen et al. 2020; Lucieri et al. 2020), relate a decision to one of a number of predefined disentangled concepts (Chen et al. 2020) or building simplified models over the decision logic (Elshawi, Sherif, and Sakr 2021; Rabold, Schwalbe, and Schmid 2020). We arrange these systems as aiming for intervention in relation to their input data and their labels since they present both contrastive and non-contrastive explanans thus making it possible for a domain expert to construct *what if* explanations. Knowledge priors added are, for example, the selection of disentangled concepts or using a decision tree as the structure for a contrastive explanation.

Counterfactuals (What if I had done?) Since counterfactual explanations build on a capability to imagine alternative futures, from an historical point in time, there is a need for temporal data for this type of explanation. In the reviewed work there is one example based on electronic health records and a comparison between how a specific treatment at a certain situation, for example, antibiotics treatment, can be evaluated in comparison to alternative treatment (Mincu

et al. 2021).

Explanation categories If we view the reviewed research through the lens of D-N explanations the explanations aims for either a *model-data* or *kind-data* relation. Data are in all cases images except in Mincu et al. (2021) where temporal tabular data from electronic health records is used. Below we analyse the work reviewed in relation to their core relation.

In Figure 3, to the left, the structure of a *kind-data* explanation is presented. As discussed earlier we view the trained model as a law under scrutiny and not a *model* in D-N sense. For example, in Lucieri et al. (2020), that focuses on identifying skin lesions, the images presented as similar to the image to be explained, together, with the trained model create the explanans that are a selected part of the available explanandum. This explanandum then answers a question of the form: This instance belongs to the concept **quality a** (a specific skin lesion concept) presenting these similar images as evidence/explanans for the decision (**quality b**). If a human with domain knowledge finds that these evidence sufficient, perhaps by combining them with other factors as experience, known sub-kind concepts or data not included in the training data then an explanation that builds on causality (if C then D) valid in the real world can be formulated by the domain expert.

In one example by Chen et al. (2020) the correlation between five disentangled *kind* concepts: *BEDROOM*, *AIRFIELD*, *STABLE*, *BOAT DECK* and *INDOOR LIBRARY* and seven disentangled sub-kind concepts: *AEROPLANE*, *BED*, *BENCH*, *BOAT*, *BOOK*, *HORSE*, *PERSON* is part of a proof of concept experiment. Here the explanadum consists of the classified *kind*, the trained model and sub-kinds pictured and not pictured in the image. A contrastive explanation can then be formulated around the inner workings of the model for example that: since the sub-kinds identified are *PERSON* and *BED* the model classifies the picture as a *BEDROOM* and not a *STABLE*. This explanation is contrastive from the trained model's perspective in the sense that it explains the classification and that it is likely that images containing a bed and a person will be classified as bedrooms. Here, again, knowledge priors in the form of selection of kinds and sub-kinds, but also training data and model selection, together delimits the explanadum available. So even if the form of the explanation can be classified as intervention the explanation is not causal in the sense that it holds in a real world context, instead it gives insights in how the trained model associate input data with outputs. The same holds for the proof of concept in Mincu et al. (2021) where a system in the hands of a person holding medical expertise can be a tool useful to create counterfactual explanations. The explanans presented by the system, together with other explanantia, can open up to better understand the consequences of an alternative historical decision. For example that it is probable that using a different type of antibiotics on women than men will make women recover faster.

In Figure 3, to the right, the structure of a *model-data* explanation is presented. We placed the work that compares

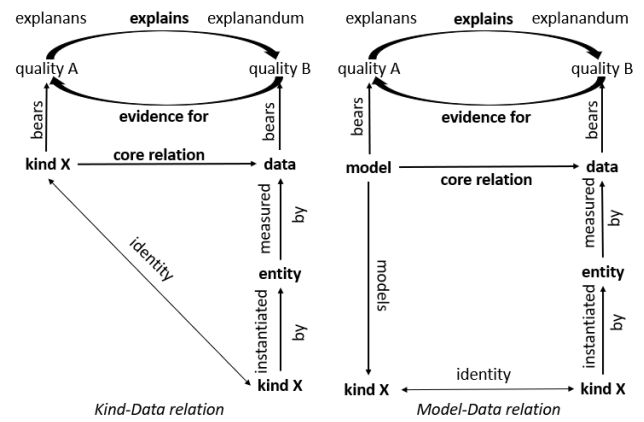


Figure 3: The two explanation structures the review articles aim for.

the human decision process with the ML decision process as a core relation between *model* and *data* since the aim is to use an overlap as an explanan and useful *model* over a trustworthy decision process (Natekar, Kori, and Krishnamurthi 2020; Lucieri et al. 2020; Yeche, Harrison, and Berthier 2020). In Wang et al. (2020) calculations of sufficient and necessary causes are used to explain a decision. Knowledge priors added here are then under which conditions a presupposed value of an alignment to human learning is useful as an explanan and under which conditions it is possible to calculate sufficient and necessary cause in an inductive learning process.

In two reviewed work a *model*, in a D-N sense, over the decision process is created. In the work by Rabold, Schwalbe, and Schmid (2020) sub-kinds are automatically identified and used to build first order logical rules covering spatial relations in images (the relative placement of *EYES*, *MOUTH* and *NOSE* useful to identify a *FACE*). In Elshawi, Sherif, and Sakr (2021) a decision tree based on sub-kinds is constructed and used to explain classification of pictures, answering contrastive questions like: Why is this image classified as a *COAST* and not a *MOUNTAIN*?

The work reviewed does not include any research that uses a *theory-data* relation or a *entity-data* relation (See Table 1). Related to *theory-data* it can be argued that using a surrogate *model* implicitly presumes that the proposed decision can be explained using, in this case, a decision tree or first order logic (Rabold, Schwalbe, and Schmid 2020; Elshawi, Sherif, and Sakr 2021).

There is in the reviewed work no trained model that focus on *entity-data* relations, relations that a neural network can learn well in a similar fashion as it can learn *kind-data* relations. *Entity*-related concepts follows the instance and can be related to ageing or wear and tear, for example, *SCRATCHES*, *SPLINTERS* or *MARKINGS* and can be useful to, for example identify objects and estimate ageing (Holmberg, Generalao, and Hermansson 2021).

In line with Tjoa and Guan (2019) we find a lack of user studies in the work we reviewed. The studies conducted are limited even if they address a mundane domain where

users are readily available (Elshawi, Sherif, and Sakr 2021; Ghorbani et al. 2019). The lack of studies in explainable AI is notable since the research heavily relies on human traits and abilities to, for example, compare images for similarity and infer disentangled sub-kind concepts. Concrete examples from the work we reviewed includes, look at images picturing beds from different environments and relate them to the concept `BED` (Chen et al. 2020) or infer that the tail is sufficient evidence to classify a `KOMODO LIZARD` but that the rocks it rests on and the body are necessary evidence (Wang et al. 2020).

The reviewed work exposes a wealth of undefined expectations on what the explainee can infer from the evidence for a decision exposed by the explainability methods. Specifically, the explainee is expected to understand the model's limitations and be aware that explanations produced reflect the training data, the architecture used and that it only is an incomplete overlay on the reality it models. The possibility for an explainee to fathom this difference is crucial if we are not only interested in the more introspective project of evaluating the model built as such but interested in applying proposed decisions in the real world. For example, to evaluate if a predefined named concept selection is relevant in relation to the domain targeted and the decision promoted or, alternatively, if the decision is due to exposure to o.o.d and non i.i.d data. For example, evaluate if the predefined concepts chosen: `BED`, `SINK`, `SEA`, `TREE`, `HIGHWAY` are the best one to evaluate the classification as a image picturing a `COAST` (Elshawi, Sherif, and Sakr 2021).

Discussion

In this section, we discuss the review results using our theoretical lens. Initially, the centrality of concepts is lifted followed by a comparison with the systems we aim for using scientific instruments as a comparison and base for the discussion. We then focus on the central notion of causality and the role it has related to explanations, this is followed by a discussion on training data distribution. The section ends with limitations and a summary section.

It is perhaps not surprising that methods that build on extracting IKR dominate the review papers since they aim for global understanding more 'by design'. FBA methods are local in that they compare one decision with the trained ML models internal knowledge representations. In our setup, FBA methods can be used to falsify the model and get a deeper understanding related to outliers in non-i.i.d data and unknown domain related concepts in o.o.d data. IKR methods, on the other hand, give insights that are more general and 'typical' for the trained model. This point towards the need for user-studies that combines IKR and FBA methods in studies that aim towards understanding the model from a global meta-perspective similar to how we 'understand' and trust companion species. It is somewhat surprising that in other XAI papers well-cited research, LIME and SHAP, are relatively invisible in the selected articles. One reason can be that they, in creating local surrogate models (f.x. linear), adds an extra layer that needs to be interpreted to get global understanding. In our setting these surrogate models can be faster to interpret since they simplify and can allow

for the domain expert to experiment and search for decision boundaries or get a general overview of the knowledge representations learned by the ML system.

Explicating and identifying concepts are in all cases in the reviewed research done using human knowledge. The incomplete world model in the trained models becomes apparent in, for example, Ghorbani et al. (2019) where images of ocean water that is `CALM`, `WAVY` and `SHINY` are identified as separate concepts and not as sub-concepts of `OCEAN`. Developing concept ontologies, concept formulation and reformulation are then seminal and support the position we take here, that a human with domain expertise interacting with the system needs to be part of any system, based on inductive learning, used in a non-stationary context.

In the radiology related research Yeche, Harrison, and Berthier (2020) and Natekar, Kori, and Krishnamurthi (2020) the approach can be seen as an incremental development of scientific instruments in line with previous technological progress within a well-defined usage domain in the hands of domain experts (Roscher et al. 2020; Karpayne et al. 2017). The research focuses on shared hermeneutic 2D images that have a substantial explanandum overlap between the ML system and the domain expert. The work reviewed that targets the medical domain focus on actionable explanations related to the decision and less on explaining the ML models inner workings. This is an indication that the current focus in ML on large static data sets needs to be complemented with datasets that has more overlap with human understanding of how the world is constituted, its diversity and contextual dependence. Increased focus on `entity-data` relations can then complement the current objectivising `kind-data` focus, and open up an interesting path towards more contextual and small-scale usage of ML systems, systems that then can add value to humans in context.

In this work, we pay special attention to what ML, in the form of neural networks, *cannot* do based on its statistical inductive learning approach. The epistemic consequences are originally formulated by David Hume as the problem of induction (Henderson 2020), popularised as the black swan problem, a problem that cannot be solved using more data. Since ML/AI of today is void of understanding, and only handles local domain generalisation (Chollet 2019), we have to be mindful of the black swans these systems do not see and can hide for us even if we as humans, and understanding seeking animals, are aware of them (Prasada 2017). To combine these information processing artefacts that ML systems are, with humans seeking understanding, we need systems that can explain themselves or, as the focus in the work presented here, find protocols so humans can understand and compensate for ML systems shortcomings.

The reviewed research that aims for interventional or counterfactual explanations in Table 1 rely on causal relations. These contrastive and counterfactual explanations are then only valid for the ML model in isolation. We find that there, in the reviewed work, is a lack of discussion related to this, if the goal is to create explanations applicable in the domain the ML system targets. For example calculation of necessary and sufficient causes (Wang et al. 2020) has to

be evaluated towards how well the training data reflects the context it will be used in, and if the training data carries this subjective information in a form that is not only statistical.

The approach promoted here that views the ML system as a tool in the hands of a domain expert is one path forward underpinned by limitations unveiled in the reviewed work. Especially the `theory` and `model` categories need to be part of the discussion so deductive reasoning can be included and implemented in the ML-system or by using human capabilities. This would then be an approach that makes it possible to understand and challenge a, by the ML system, promoted decision.

Our choice to use the D-N model and Overton’s (2012) schematic overview delimit the type of explanations we aim for and exclude pragmatic, inductive statistical explanations, teleological and historical. Also, we have not deepened the discussion around the difference between description, justification, argument and explanation (Woodward and Ross 2021; Salmon 1989). As for concepts, we see them as central building blocks for thoughts and we are aware of that our understanding of the concept `CONCEPT` is not here well-grounded from an ontological and epistemological perspective.

Our framing using scientific explanations and types of explanations illuminate an ambiguity in the reviewed work concerning the goal of the explanation. Is the goal understanding the ML model or understanding the domain modelled by the ML system? We can also see that this ambiguity is less pronounced when there is a large explanandum overlap between the reality the ML model models and the reality we as humans perceive. An interesting path forward is then to use scientific explanations, for example, using a complete general structure for D-N explanations (Overton 2012, pg. 17) (See Figure 4). Using this structure to formulate theories and falsifiable hypothesis similar to a traditional research process is one path towards evaluating how well a trained ML model models the usage domain. By doing this we move away from the idea that more data solves the problem of induction and instead treat ML systems as tools that can mediate better understanding. By complementing ML systems that builds on large data sets with `theories` that can be translated to `models`, in the form of, for example, causal graphs, algorithms and logic, we can add a needed `model` layer to these systems. Addressing these explanation structures is an important future focus that can create systems that can be challenged and possible to learn from.

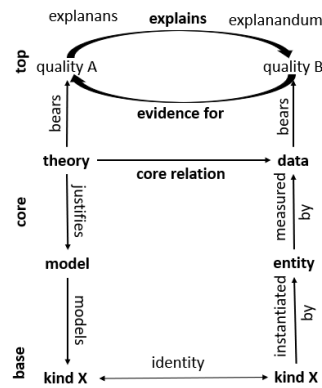


Figure 4: theory-data relation based on kind identity.

In this work, we take an outside perspective in relation to a trained ML system and we find similarities with a scientific process that uses hypothesis and theories that are possible to challenge, improve and refine in relation to the domain targeted. Additionally we lift out a number of areas, the importance of concepts, causality, data shift and explanation types and explanation categories, essential to make these systems falsifiable.

Conclusion

ML systems increasingly affect many aspects of human life, gaining trust in their decisions is a central and active research area. *What if*-questions and the centrality of concepts is the focus for this review where we examine how concepts are extracted from a neural network. We presuppose a situation where a human, with domain knowledge, use concepts to answer why-questions. In our review, we use the structure of D-N explanations and three types why-questions, *What if I see?*, *What if I do?* and *What if I had done?* as an analytic lens to deepen and detail what we can expect, and not expect, from the research reviewed.

This review raises important questions on *What is the goal for the explanation?* and *What type of knowledge can be extracted from a neural network?*. Related to the first question we see the importance of differentiating between explanations that focus on a better understanding of the trained model, void of context, and those that focus on actionable decisions, useful in the deployment context. Related to the second question we see that the reviewed work, in many cases, aims for explanations that build on causal relations, that are required for these types of explanations, without discussing how these are added to the system.

We believe that studies that actively involve users can emphasise contextual dependence and refocus research in the area more towards the limitations of ML systems and consequently open up for an awareness of the societal and environmental impact these systems have when they are deployed.

References

- Adadi, A.; and Berrada, M. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.
- Akula, A.; Wang, S.; and Zhu, S.-C. 2020. CoCoX: Generating Conceptual and Counterfactual Explanations via Fault-Lines. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(03): 2594–2601.
- Bareinboim, E.; Correa, J. D.; Ibelind, D.; and Icard, T. 2020. On Pearl’s Hierarchy and the Foundations of Causal Inference. Technical report, CausalAI Laboratory, Columbia University.
- Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Benetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; and Herrera, F. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58: 82–115.

- Bengio, Y.; Lecun, Y.; and Hinton, G. 2021. Deep learning for AI. *Communications of the ACM*, 64(7): 58–65.
- Biran, O.; and Cotton, C. 2017. Explanation and Justification in Machine Learning: A Survey. In *IJCAI Workshop on Explainable AI (XAI)*, volume 8, 8–14.
- Carnap, R. 1962. Logical foundations of probability.
- Chen, H.; Janizek, J. D.; Lundberg, S.; and Lee, S.-I. 2020. True to the Model or True to the Data?
- Chollet, F. 2019. On the Measure of Intelligence. *arXiv preprint arXiv:1911.01547*, 64.
- Couldry, N.; and Mejias, U. A. 2019. *The Costs of Connection: How Data Are Colonizing Human Life and Appropriating It for Capitalism*. Stanford University Press; 1 edition (August 20, 2019). ISBN 9781503609754.
- Elshawi, R.; Sherif, Y.; and Sakr, S. 2021. Towards automated concept-based decision tree explanations for CNNs. In *Advances in Database Technology - EDBT*, volume 2021-March, 379–384.
- Genone, J.; and Lombrozo, T. 2012. Concept possession, experimental semantics, and hybrid theories of reference. *Philosophical Psychology*, 25(5): 717–742.
- Ghorbani; Wexler, J.; Zou, J.; and Kim, B. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*.
- Gillies, M.; Lee, B.; D’Alessandro, N.; Tilmanne, J.; Kulesza, T.; Caramiaux, B.; Fiebrink, R.; Tanaka, A.; Garcia, J.; Bevilacqua, F.; Heloir, A.; Nunnari, F.; Mackay, W.; and Amershi, S. 2016. Human-Centred Machine Learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA ’16*, 3558–3565. ISBN 9781450340823.
- Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2019. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 80–89. IEEE. ISBN 9781538650905.
- Grimm, S. R. 2016. How Understanding People Differs from Understanding the Natural World. *Nous-Supplement: Philosophical Issues*, 26(1): 209–225.
- Grimm, S. R. 2021. Understanding. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 21 edition.
- Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Gianotti, F.; and Pedreschi, D. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5): 42.
- Hempel, C. G.; and Oppenheim, P. 1948. Studies in the Logic of Explanation. *Philosophy of Science*, 15(2): 135–175.
- Henderson, L. 2020. The Problem of Induction. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
- Hesslow, G. 1988. The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, 11–32.
- Hilton, D. J. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1): 65–81.
- Hilton, D. J. 1996. Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning*, 2(4): 273–308.
- Hilton, D. J.; and Slugoski, B. R. 1986. Knowledge-Based Causal Attribution: The Abnormal Conditions Focus Model. *Psychological Review*, 93(1): 75–88.
- Hoffman, R. R.; Clancey, W. J.; and Mueller, S. T. 2020. Explaining AI as an exploratory process: The peircean abduction model. *arXiv preprint arXiv 2009.14795*.
- Holmberg, L.; Generalao, S.; and Hermansson, A. 2021. The Role of Explanations in Human-Machine Learning. In *IEEE Transactions on Systems, Man, and Cybernetics*.
- Justus, J. 2012. Carnap on concept determination: Methodology for philosophy of science. *European Journal for Philosophy of Science*, 2(2): 161–179.
- Karpatne, A.; Atluri, G.; Faghmous, J. H.; Steinbach, M.; Banerjee, A.; Ganguly, A.; Shekhar, S.; Samatova, N.; and Kumar, V. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, 29(10): 2318–2331.
- Kenny, E. M.; and Keane, M. T. 2021. On Generating Plausible Counterfactual and Semi-Factual Explanations for Deep Learning.
- Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; and Sayres, R. 2018. Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *35th International Conference on Machine Learning, ICML 2018*. ISBN 9781510867963.
- Knopf, J. W. 2006. Doing a literature review.
- Lipton, Z. C. 2016. The Mythos of Model Interpretability. *Communications of the ACM*, 61(10): 35–43.
- Lucieri, A.; Bajwa, M. N.; Alexander Braun, S.; Malik, M. I.; Dengel, A.; and Ahmed, S. 2020. On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 1–10.
- Lundberg, S.; and Lee, S.-I. 2016. An unexpected unity among methods for interpreting model predictions. *arXiv preprint arXiv:1611.07478*.
- Margolis, E.; and Laurence, S. 2021. Concepts. In *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.
- Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.
- Mincu, D.; Loreaux, E.; Hou, S.; Baur, S.; Protstyuk, I.; Seneviratne, M.; Mottram, A.; Tomasev, N.; Karthikesalingam, A.; and Schrouff, J. 2021. Concept-based model explanations for electronic health records. In *ACM CHIL 2021 - Proceedings of the 2021 ACM Conference on Health, Inference, and Learning*, 36–46.

- Murphy, G. L. 2016. Is there an exemplar theory of concepts? *Psychonomic Bulletin and Review*, 23(4): 1035–1042.
- Natekar, P.; Kori, A.; and Krishnamurthi, G. 2020. Demystifying Brain Tumor Segmentation Networks: Interpretability and Uncertainty Analysis. *Frontiers in Computational Neuroscience*, 14.
- O’Neil, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books. ISBN 9780553418835.
- Overton, J. A. 2012. *Explanation in Science*. Ph.D. thesis, University of Western Ontario.
- Pearl, J. 2019. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3): 54–60.
- Pearl, J.; and Mackenzie, D. 2018. *The book of why: the new science of cause and effect*. Basic books. ISBN 9780465097616.
- Prasada, S. 2017. The scope of formal explanation. *Psychonomic Bulletin and Review*, 24(5): 1478–1487.
- Rabold, J.; Schwalbe, G.; and Schmid, U. 2020. Expressive explanations of DNNs by combining concept analysis with ILP. *German Conference on Artificial Intelligence*, 12325 LNAI: 148–162.
- Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. ”Why should i trust you?” Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, volume 13-17, 1135–1144. ACM.
- Roscher, R.; Bohn, B.; Duarte, M. F.; and Garcke, J. 2020. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, 8: 42200–42216.
- Salmon, W. C. 1989. The Fourth Decade (1978-87) A Time of Maturation. In *Four Decades of Scientific Explanation*, 117–179.
- Scholkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5): 612–634.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2): 336–359.
- Tjoa, E.; and Guan, C. 2019. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. In *IEEE Transactions on Neural Networks and Learning Systems*.
- Vallor, S. 2016. *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.
- Wang, Z.; Mardziel, P.; Datta, A.; and Fredrikson, M. 2020. Interpreting interpretations: Organizing attribution methods by criteria. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, volume 2020-June, 48–55. ISBN 9781728193601.
- Webb, A. 2019. *The Big Nine*. New York: PublicAffairs,U,S, first edit edition. ISBN 9781541724419.
- Webster, J.; and Watson, R. T. 2002. Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2): xiii – xxiii.
- Woodward, J.; and Ross, L. 2021. Scientific Explanation. In Zalta, E. N., ed., *The Stanford Encyclopedia of Philosophy (Summer 2021 Edition)*.
- Yeche, H.; Harrison, J.; and Berthier, T. 2020. UBS: A Dimension-Agnostic Metric for Concept Vector Interpretability Applied to Radiomics. In Suzuki, K.; Reyes, M.; and SyedaMahmood, T., eds., *Interpretability of machine intelligence in medical image computing and multimodal learning for clinical decision support*, volume 11797 of *Lecture Notes in Computer Science*, 12–20. ISBN 978-3-030-33850-3; 978-3-030-33849-7.