Frankfurt School

Master's Thesis

Putting the Horse back before the Cart: Evaluating Methods in Explainable Artificial Intelligence against Philosophical Accounts of Explanation

Christopher Ludwig Schmitz

Putting the Horse back before the Cart: Evaluating Methods in Explainable Artificial Intelligence against Philosophical Accounts of Explanation

Master's Thesis

at the Frankfurt School of Finance and Management

Supervised by

Gregory Wheeler

Sebastian Köhler

Submitted by

Christopher Ludwig Schmitz

Master in Applied Data Science 2021

m4203352

Glauburgstraße 83, 60318 Frankfurt am Main

+49 1601588375

christopherl.schmitz@gmail.com

Statement of Certification

I hereby confirm that this thesis constitutes my own work, produced without aid and support from persons and/or materials other than the ones listed. Quotation marks indicate direct language from another author. Appropriate credit is given where I have used ideas, expressions or text from another public or non-public source. The paper in this or similar form has never been submitted as an assessed piece of work in or outside of Germany. It also has not yet been published.

Christopher Schmitz

Frankfurt, 3.9.2021

Abstract

The past decade has seen rapid proliferation of artificial intelligence (AI) across industry and academia, but the workings of AI systems are largely unintelligible to humans. The nascent field of explainable artificial intelligence (XAI) aims to remedy this by making these self-learning models interpretable. However, a survey of the methods employed reveals that their development usually includes little reflection on what constitutes a ‘good’ explanation in a philosophical sense, instead following technical ‘paths of least resistance’. Work that approaches these questions at best poses requirements to what explanations ‘should be’, rather than evaluating existing methods. As such, it remains unclear whether they meaningfully explain underlying models.

This work aims to answer this question by deriving a rigorous, evaluable model of explanation, as applicable to XAI, from literature in the philosophy of science. It contains four core features, which are argued to be broadly consensual among differing accounts of explanation: answering why-questions, contextuality, causal connection, and being based on accurate facts. The model also contains three less canonical extensions from different accounts of explanation: goal-directedness from functional accounts, reliance on statistical laws from statistical accounts, and the generation of understanding from pragmatic accounts.

Six families of XAI methods, the selection of which is justified by a process that optimizes for representative coverage of most of the field while maintaining a feasible scope, are then evaluated against this model. Most of them are shown to fulfill some, but not all, of its requirements. Transparent models and counterfactual approaches perform best. Surprisingly, especially the core requirement of answering why-questions is often unfulfilled. Some implications of these results are discussed, both in the direction of how this model and its findings may influence XAI method development, and in the direction of what the ‘use case’ of XAI may imply for open debates in the philosophy of explanation.

Table of Contents

List of Figures

1 Introduction & Motivation

Across the last decade, advances in computational power and theoretical research have allowed artificial intelligence (AI) systems to propagate rapidly across a wide range of academic and industrial domains. However, the self-learning models which are employed, and the decisions they produce, are often not intelligible to humans. Accordingly, AI adoption is being held back by human-level concerns about its workings, for example lacking trust in the decisions AI models make, ensuring freeness from bias, and accounting for AI actions (Vilone and Longo, 2020: 7).

The field of explainable artificial intelligence (XAI) aims to address these issues. Methods in XAI look to ensure the explainability of otherwise uninterpretable black-box models. It is a young, but explosively growing field of research, as documented by the exponential increase in XAI papers published annually (Arrieta et al, 2019: 3). This underscores the mounting pressure generated by otherwise feasible AI use cases which are held back by lacking interpretability.

Owing to its youth, XAI as a subject area has very little in the way of canonical research. Some methodologies, however, are seen as standard, and taxonomies of XAI also tend to represent the field quite homogeneously. Though a canon appears to be evolving, a survey of its methods reveals a tendency to work downwards from the available technology rather than upwards from a well-defined concept of explanation. The result is a wide field of methods which have been created along technical ‘paths of least resistance’ without rigorous consideration of whether they improve model explainability to humans in any meaningful way. Where this is attempted, it is usually worked from the model towards the explanation, which results in unintuitive methods and inconsistent evaluations thereof. Remedying this is the purpose of this work.

The research question guiding this work is “To what extent do current approaches in XAI provide ‘good’ explanations, as measured by philosophical models of explanation?”

This work attempts to invert the cart-before-the-horse nature of current XAI research by first deriving a rigorous, evaluable model of explanation from literature in the philosophy of science. To accommodate the active debates in the field, this model has a ‘core’ set of requirements based on the commonalities between different accounts of explanation, and is extended by additional requirements some individual models of explanation set. Families of XAI methods are then evaluated against the model, to mixed results; no method family provides fully satisfying explanations, though transparent models and counterfactual generation come close. Accordingly, the final section contains a discussion and recommendations for further directions in research, both in XAI as guided by the findings from philosophy and in the philosophy of explanation as informed by the ‘use case’ of XAI.

Given the liveliness and broadness of the XAI field, it is intractable to evaluate every method individually. As such, the research objective is subdivided into three parts: (1) the identification and description of families of XAI methods which collectively exhaust the field, are each internally cohesive, and are each suited to comparison to a philosophical model of explanation, (2) the development of a philosophical model of explanation as applicable to XAI, and (3) the evaluation of the identified method families against it and the derivation of implications for future research.

2 Literature Review

To the end of the three separate tasks above, the literature review is divided into four sections. The first summarizes the state of XAI research. It aims to derive a holistic view of the approaches taken to explain models, allowing for an intuitive grouping as described in the first aim. As such, beyond taxonomizing existing techniques, it highlights the major axes along which models are tendentially divided, such as global vs local explainers, model-agnostic vs model-specific explanation, and so on. The second section reviews philosophical work on the theory of explanation, with the aim of translating this work into an operable model. It also highlights some of the major debates in the field, as justification for a more cautious approach to the aim of deriving a singular model of explanation in XAI. The third section summarizes research on human interaction with AI, specifically with XAI methods, seeking to establish the place of this work in the literature, a more holistic view of the motivation of XAI research, and a more complete idea of how concepts of explanation ‘goodness’ are translated from the social and natural sciences into the AI space. Finally, the fourth section summarizes attempts to standardize or quantify the evaluation of XAI techniques, including proposals for evaluation frameworks, to establish commonalities and differences with this work’s approach.

2.1 A Taxonomy of XAI Methods

The task of collecting and sorting XAI techniques has been undertaken by a significant number of authors, which has resulted in a flurry of surveys of the field (eg. Arrieta et al., 2019; Adadi & Barreta, 2018; Guidotti et al., 2018; Mueller et al., 2019; Vilone and Longo, 2020; Burkart and Huber, 2020; Das and Rad, 2020). These efforts are not duplicated in this literature review – not only is there an intractable number of methods, but the goal of this work is not to name and categorize them all. Rather, a meta-review identifies the most common rules by which models are frequently categorized, with an eye to the later analysis of internally similar groups of models. Four generally agreed upon key divides of models are listed in section 2.1.2, and a number of less canonical additional divides are listed in section 2.1.3. As quickly becomes clear, some of these divides lend themselves well to a hierarchical classification of the field, but many also generate overlapping groups, so that a strictly hierarchical classification is not realistic.

To illustrate this point, and more generally to concretize the approach this literature review takes, this section begins by introducing five prototypical XAI methods in detail.

2.1.1 Key XAI Methods

This section introduces five prototypical XAI methods in detail. The purpose of this is twofold: firstly, to present illustrative examples of what current XAI methods generate, and secondly, to concretely demonstrate the difficulty of creating an exhaustive taxonomy of methods, because equally reasonable division rules can create vastly differing groupings. As such, they were chosen to differ one another, and to straddle some of the divides presented in sections 2.1.2 and 2.1.3.

1) SHAP (Lundberg and Lee, 2017) is the singular most-implemented XAI method (Molnar, 2020: 5.11). It generates local feature importance values, which measure the impact each input feature of the given model has on the prediction made for one particular datapoint. These values are derived from Shapley values from coalitional game theory; specifically, progressively larger ‘coalitions’ of input features are generated, and each feature’s contribution to the coalition’s score – its predictive strength – is measured (Molnar, 2020: 5.11). As the number of possible coalitions grows exponentially with the number of features, only a small fraction of the possible coalitions is evaluated (Lundberg and Lee, 2017: 5). SHAP is a model-agnostic black-box method, which is to say it can operate on any supervised model, regardless of its internal structure, requiring only access to its prediction function. There are, however, slightly differing model-specific versions of SHAP for some model architectures, specifically trees and neural networks (Lundberg and Lee, 2017:6). It also generates global explanations by combining local ones (Lundberg and Lee, 2017: 5).

2) LIME (Ribeiro et al., 2016) is a similarly popular XAI method. Like SHAP, it generates local feature importance values, but it follows a two-step process: the generation of a synthetic dataset, and the training of a locally linear surrogate model around one datapoint using this synthetic dataset (Ribeiro et al., 2016: 3). This surrogate tends to be quite accurate around the datapoint surveyed but generalize very poorly (Ribeiro et al., 2016: 4). Its commonalities with SHAP are thus that it is a feature importance method, local, and model-agnostic. Its big

distinguishing factor is its use of a surrogate model. Like SHAP, there is an extension of LIME to provide global feature importances called SP-LIME, which instead of generating a synthetic dataset samples from the original (Das and Rad, 2020: 10).

3) Partial Dependence Plots (PDP) are a feature importance visualization method which measures the marginal contribution one or two features have to a model's output (Molnar, 2020: 5.2). To generate a PDP, the marginal effect of varying a feature across its domain is calculated from data points sampled using Monte Carlo. The average predicted output value (for regression) or the average probability of a given class (classification) for each value of the feature is then plotted. One key assumption that is made is that the target input variable(s) is independent from each of the variables that are held constant, which of course is an unfulfillable requirement in practice (Molnar, 2020: 5.2). This means that, depending on the correlation of the target feature with other features, PDPs can depict highly improbable datapoints. A number of iterations on PDPs aim to correct this; notably, Individual Conditional Expectation (ICE) plots forego averaging and represent each data instance with its own line, and Accumulated Local Effects (ALE) plots use conditional instead of marginal distributions (Molnar, 2020: 5.2).

4) Anchors (Ribeiro et al., 2018) is a method that is frequently cited to be representative of rule extraction approaches (Vilone and Longo, 2020: 18). It generates decision rules to which an analyzed model fully adheres; specifically, where an 'Anchor' rule holds, any variable not named in the rule can change without the prediction changing, to a certain probabilistic standard (Ribeiro et al., 2018: 3). These rules are textual and take the form of IF/THEN/AND/OR statements. They are model-agnostic, only applicable to classifiers, and post-hoc. Rather than strictly local or global, though, they are *scoped*: an 'anchor', once identified, can be applied to all other datapoints, but only some of the dataset is covered by it (Molnar, 2020: 5.9).

5) GradCam (Selvaraju et al., 2016) is included representatively for the research field of gradient-based saliency maps, which is growing rapidly (Vilone and

Longo, 2020: 24). These techniques analyze neuron activations in neural networks for image classification, with the aim of discovering which area of a given image is particularly relevant for its classification. GradCam in particular derives importance values for neurons for a given decision using the gradient information available in the last layer of a CNN (Selvaraju et al., 2016: 4). It is model-specific to CNNs, local in that it only applies to one prediction at a time, and post-hoc in that it is applied after the model is trained.

2.1.2 Key Divisions of XAI Techniques

This section introduces four key fault lines along which XAI techniques are classified. As opposed to the more detailed divisions outlined in the next section, these are broadly considered fundamental, in that they can be found in virtually all taxonomies of the field. This section also includes a non-exhaustive summary of the techniques found in the categories created by these fault lines, and key points of the discourse surrounding them.

(i) Transparent Models and Post-Hoc Techniques

Firstly, all surveys distinguish between post-hoc explainability techniques and inherently interpretable models (see for example Molnar, 2020: 4; Arrieta et al, 2019: 10; Guidotti et al., 2018: 8, Mueller et al., 2019: 5). Where the former attempt to explain fully uninterpretable models, the latter are described as inherently not requiring any further explanation because their workings and output are entirely comprehensible to humans (Arrieta et al., 2019: 12). There is even a consensus on what families of models are inherently interpretable: linear and logistic regressions, decision trees and random forests, and rule-based methods (Guidotti et al., 2018: 8). Rule- and tree-based models are considered interpretable because they inherently provide if-then decision rules, such as “if age > 18 and income > 100000, then return 1”, which have a clear meaning to humans (Guidotti et al., 2018: 8). In turn, linear and logistic regression, in their coefficients, return a quantification of the importance of each feature in the input data, which is comparable across features if the features are normalized (Guidotti et al., 2018: 8). Beyond these, there are some other models that are frequently, but not unanimously mentioned as interpretable, including k-Nearest Neighbors and Naïve Bayes (Burkart and Huber, 2020: 13); also, there is a class of research attempting to create more powerful natively interpretable models (Molnar, 2020: 4.7).

Interestingly, despite these models being referred to as interpretable almost unanimously, what exactly makes them interpretable is rarely addressed. Burkart and Huber define it as having a “training process that is not explicitly optimized for interpretability” (interpretable by nature), or “including interpretability in the design of the training” (interpretable by design) (2020: 20-22). Lipton (2016: 3) states interpretability is satisfied if a model’s “functionality can be comprehended in its entirety by a person”. The implicit definition of interpretability, which is confirmed by the definitions Burkart and Huber and Lipton give, appears to be that a model natively provides output which it otherwise requires post-hoc methods to produce, such as feature importances in linear and logistic regression and decision rules in tree and forest models. As such, evaluating whether they really are natively interpretable may reduce to evaluating whether these post-hoc methods generate good explanations.

Some authors subdivide or qualify transparent models further. Lipton (2016: 9) defines three classes of transparency: simulatability, decomposability, and algorithmic transparency. Molnar (2020: 4) provides three further criteria interpretable models should ideally satisfy: linear relationships between input and target variable, monotone relationships between input and target variable (ie. strictly increasing or decreasing), and native support for interaction between input variables. Burkart and Huber (2020: 15) also mention sparsity, because human cognitive ability is limited.

There are also some more formal attempts to quantify model transparency in the vein of what Doshi-Velez and Kim (2017: 5) call the functionally-grounded approach, the search for proxy values that imply interpretability, which is expanded on more generally in section 2.4. There are a handful of intuitive proxy indicators for transparency: Tibishani (1996: 267) highlights that sparse models with fewer features are more interpretable than dense ones; Burkart and Huber mention the depth of decision trees, the number of trees in a random forest, or the number and complexity of decision rules (2020: 28).

Calls to use interpretable models rather than explaining black boxes are becoming more frequent. Rudin (2019:3-5) highlights that especially in high-stakes environments like justice and healthcare, only fully interpretable models satisfy the requirements users set or should be setting of models, most notably faithfulness to the model, detailed understanding of its workings, and the ability to convert model decisions into action easily. Molnar et al. (2021: 6), too, refer to using unnecessarily complex models as a general pitfall of XAI in practice, highlighting that at *best*, post-hoc techniques provide insights that interpretable models provide natively. Both Rudin (2019: 2) and Molnar et al. (2021:6) also stress that the trade-off between interpretability and model performance is much less drastic than it is often assumed to be, meaning natively interpretable models are suitably performant for many contexts where black-box models are used.

A final note on transparent models is that the implicit consensus that transparent models are more interpretable and should be used wherever possible to maximize explainability does find some dissenters. Lipton (2016: 25) claims human reasoning is inherently post-hoc, specifically that explaining the decisions other humans make requires interpolation between inputs and outputs. Lipton also makes the claim that linear models are not strictly more interpretable than neural networks, for example, because neural networks process training data directly, while linear regression usually relies on heavily engineered features to be performant (2016: 7). Similarly, Paez (2019: 14) highlights there is not much experimental data that confirms transparent models outperform post-hoc explanations, implying this may be another case of researchers being guided by their intuition on what is explainable.

None of the five methods presented in section 2.1.1 are attributed to the category of transparent models.

(ii) Model-Specific and Model-Agnostic Techniques

The second distinction that is all but universally employed is the splitting of post-hoc XAI methods into model-specific and model-agnostic techniques (see for example Molnar, 2020: 2.2, Guidotti et al., 2018: 11, Vilone and Longo, 2020: 17). The difference is generally that model-agnostic techniques require access only to the ‘predict function’ of a model, and no knowledge about its inner workings or structure, whereas model-specific techniques build on assumptions about model interna (Arrieta et al, 2019: 18). Guidotti et al. (2018: 11) specify this by distinguishing between the ‘reverse engineering’ of model-specific explanations and the ‘design’ of model-agnostic ones. Potentially because each of these groups hosts such a breadth of models, one doesn’t find a particularly lively discussion about their benefits and drawbacks comparable to that about inherently interpretable models; the distinction appears to be largely taxonomical (Arrieta et al, 2019: 18). Indeed, many model-specific techniques are variations on model-agnostic ones that exploit some computational shortcut made available through the model's structure, but do not fundamentally differ in explanatory approach or output (Molnar, 2020: 5). An exception is model visualization, which is expanded upon below, and obviously varies entirely with the underlying model being visualized (Arrieta et al., 2019: 13).

Model-specific techniques have the benefit of lending themselves well to an intuitive further subdivision by the type of model they address. One division that is expanded upon below is that into 'deep' and 'shallow' methods, that is, methods that explain neural networks and methods that explain all other models (Arrieta et al., 2019: 21). The field of model-specific post-hoc methods for shallow models is not particularly large, among others because shallow models tend to be more interpretable by design, creating overlaps with the area of inherently interpretable models (Molnar, 2020: 4). Nonetheless, there are a handful of model-specific shallow techniques, which frequently either wrap decision trees or linear models around either the whole model or a subsection thereof, to then extract decision rules and/or feature importances, such as ExtractRule for SVMs or EBI for Bayesian and hierarchical networks (Arrieta et al., 2019: 22; Vilone and Longo, 2020).

Subdivisions of model-agnostic techniques are a lot less canonical; some attempts to so are listed in section 2.3.

Like transparent models, model-agnostic techniques have also been subject to attempts to formulate explainability requirements more stringently. Ribeiro, Singh and Guestrin (2016: 5) find three desiderata: model flexibility, the applicability to different models; explanation flexibility, the ability to produce different sorts of explanations based on requirements; and representation flexibility, the ability to create feature representations that differ from those required by the model where this increases explanatory value, such as presenting the importances of words instead of word vectors, even if the model was trained on the latter.

The techniques presented in section 2.1.1 are categorized as follows: SHAP is model-agnostic, though it does have versions that are optimized computationally for decision trees/random forests and neural networks (TreeSHAP and DeepSHAP, respectively) (Lundberg and Lee, 2017: 8). LIME, PDPs, and Anchors are model-agnostic methods, and Grad-CAM is model-specific for neural networks (Ribeiro et al., 2016: 2, Molnar, 2020: 5.2, Ribeiro et al., 2018: 2, Selvajaru et al., 2016: 1). This is still fairly intuitive, but the ambiguity of SHAP's classification hints at some of the far less clear-cut classifications to come.

(iii) Global and Local Explanation

Among post-hoc techniques, authors distinguish between techniques that provide global explanations and techniques that provide local explanations (see for example Molnar et al., 2020: 2.2, Arrieta et al., 2019: 10, Vilone and Longo, 2020: 15). Global explanations aim to explain the behavior of models as a whole, whereas local explanations restrict themselves to explaining one datapoint, or a solution subspace (Arrieta et al, 2019: 17).

Global explanation is rarely subdivided further; local explanation is subdivided frequently. For example, Molnar (2020: 2.3) distinguishes between local interpretability for a single prediction and local interpretability for a group of predictions. Burkart and Huber (2020: 16) distinguish between local explanation, counterfactual local explanation, prototype local explanation, and criticism local explanation. Many families of techniques do not strictly fall into either of these categories, but some are decidedly more at home in one of them: prototype-based methods (eg. Rüping 2006), counterfactual generation methods (eg. Wachter et al., 2019), and feature importances for one datapoint, such as those generated by SHAP (Lundberg and Lee, 2017) or LIME (Ribeiro et al., 2016), are all decidedly local methods. Even so, as detailed below, some of them have been generalized to explain models globally, complicating a clear distinction.

While there is generally some consensus that local and global explanations either work in unison to create understanding, or have different realms of applicability, as section 2.2.3 describes, some research actively pits them against each other. Adadi and Barreta (2018: 136) highlight that example-based explanations, even model-agnostic ones, are analogous to human reasoning, which is frequently prototype-based. They derive from this that they are preferable for the generation of understanding. This is echoed by a larger canon that states that past a specific point of model complexity, there is no use in generating global explanations, because humans are incapable of understanding systems that nuanced (Paez, 2019: 19).

The techniques introduced in section 2.1 are categorized as follows: PDPs are decidedly global techniques (Molnar, 2020: 5.2); SHAP and LIME are local techniques which can be expanded (in the case of SHAP through averaging predictions, and in the case of LIME through an extension called SP-LIME) to be global (Lundberg and Lee, 2017: 2; Ribeiro et al., 2016: 8); GradCam is a local technique that only explains one image being processed at a time (Selvajaru et al., 2016: 1); Anchors are ‘scoped’ in that each rule is applicable to some, but not all, of the dataset, but on the binary they would fall under global techniques (Ribeiro et al., 2018: 8).

(iv) XAI for Supervised and Unsupervised Models

A distinction that has gone implicit in this work thus far is that between XAI methods for supervised and for unsupervised learning. Specifically, the work summarized in this literature review only addresses the former. This is not to say that there is no research on XAI for unsupervised learning: for example, Puiutta and Veith (2020) summarize explainable reinforcement learning, and Zhang and Chen (2020) summarize explainable recommendation systems. As expanded upon in section 3 below, however, these make up a minority of research in XAI, to the extent that their omission is justified in this work. Of course, this means that the other three rules identified in this section are subsidiary to this one.

Overall, these four rules lend themselves well to hierarchical classification of XAI methods, each dividing a subgroup created by a superior rule. This is summarized in figure 1.

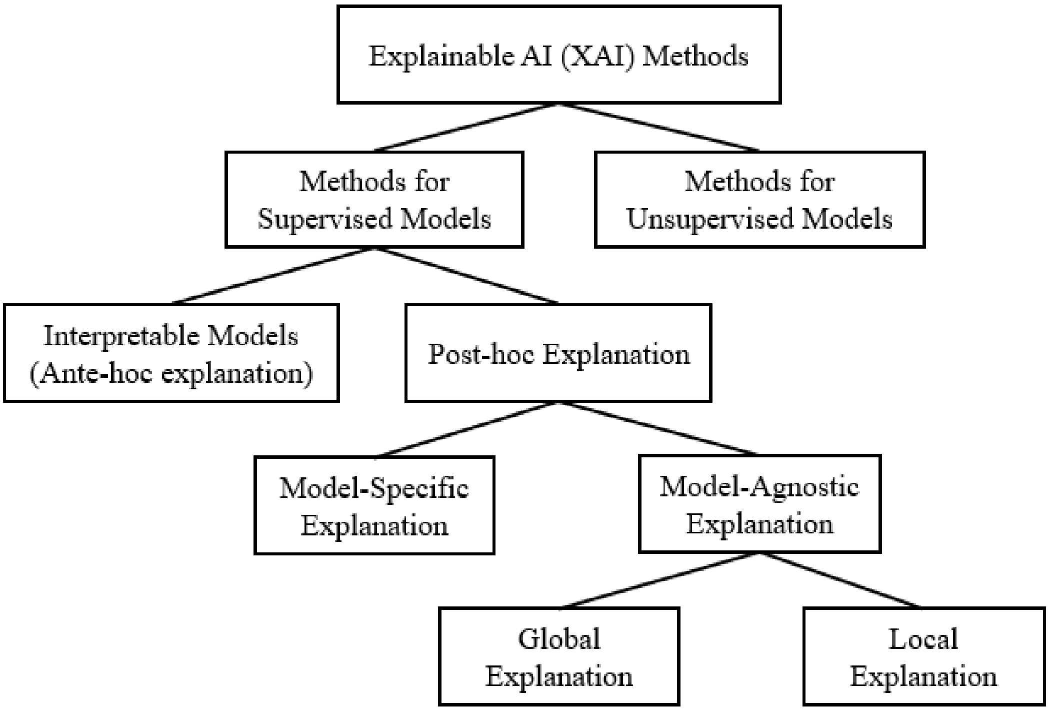
Explainable AI (XAI) Methods
Methods for Supervised Models
Methods for Unsupervised Models
Interpretable Models (Ante-hoc explanation)
Post-hoc Explanation
Model-Specific Explanation
Model-Agnostic Explanation
Global Explanation
Local Explanation

Figure 1: The Canonical Part of the Taxonomy of XAI Methods

2.1.3 Further Divisions of Models

Beyond these four clear-cut and functionally universally implemented divisions, the taxonomies differ much more strongly. Several of the approaches taken to classify models more finely are detailed in this section. It is worth stressing that there is little questioning of the validity of any of the below divisions – that is, each of them provides a sensible binary into which to categorize methods. However, they are largely incompatible, in that the groups they create often splice each other, and a hierarchical categorization that respects each of their definitions would be entirely intractable. An attempt to synergize them and generate largely cohesive families of methods for analysis is made in section 5.1 below.

One observation of note to accompany this taxonomy is that the literature usually presents techniques as largely independent and free-standing, whereas in practice, they are often presented within the same package, for example IBM's AIX360, DrWhy, IML, or Alibi Explain (Das and Rad, 2020: 18). In each of these cases, techniques from multiple of the below families are available together, which greatly lowers the barrier to using them jointly, as each of these packages typically only requires model and data to be in one format. It is therefore at least worth considering that ensembled XAI techniques may be applied, and where they provide different types of explanation, there may be synergies between them.

(i) Explanatory Approach

One approach that is common, but handled too differently in each paper as to provide one canonical form, is to divide models by their explanatory approach, i.e. the type of explanation they produce. Favel et al. (2020: 8) provide three classes in their framework for XAI evaluation: feature importances, recognition of patterns, and causal relationships. They also cite key findings in the social science of XAI, such as Miller (2017), to imply they are ranked in ascending order of informativeness. This classification appears very guided by the methodological workings of the major classes of XAI method available today. In contrast, Burkart and Huber (2020: 11) define three classes of methods, based on their relation to the underlying model: explanation generation, which creates explanatory materials based on a knowledge of a model's structure; surrogate model learning, which creates an explainable but ideally accurate surrogate around a model; and learning interpretable models, which forgoes the need to append explanative methods at all. Arrieta et al. (2019: 18) divide post-hoc techniques differently yet, half by approach and half by output, into text, visual, and local explanations, explanation by example, by simplification, and by feature relevance. Gilpin et al (2019: 4) taxonomize post-hoc methods into those that emulate the processing of the model, explain a representation of the model, and generate an explanation-producing network. Das and Rad (2020: 10) divide techniques by methodology, into perturbation-based (eg. SHAP and LIME) and gradient-based (eg. GradCAM) methods.

(ii) Type of Output

Another approach, taken for example by Molnar (2020: 2.2), is to divide techniques by the type of their output. Molnar distinguishes between feature summary statistics like importance and pairwise interaction, feature summary visualization, model internals like learned weights or tree structures, and datapoints like counterfactual explanations. A more general approach is to distinguish strictly by the output format, such as text, numerical, rules or visual (Vilone and Longo, 2020: 15).

Particularly the distinguishing of visual explanations is worth highlighting, as they are often considered separately. Vilone and Longo (2020: 15) split visual explanations into those that are fundamentally visual, and those that are supported by visuals, such as PDP plots, and their derivatives like ICE plots, ALE plots, and partial importance plots (2020: 20). There is also a significant amount of discourse about whether visualizations aid explanation. Offert (2017: 4) claims that visualizations may help, but as they often require some technical pre-interpretation, such as dimensionality reduction, we should not be overly reliant on them. Hohman et al., acknowledge that visualizations often fall short of producing full understanding, even though they claim they are the most human-centered method of explanation (2019: 4).

Among the techniques presented in 2.1.1, only GradCam, which produces saliency maps, is a fundamentally visual method. SHAP, LIME, and PDPs output numerical explanations but are frequently supported by plots thereof. Anchors provide textual rules, which do not usually have visual aids.

(iii) Feature Importances, Surrogate Models, and Rule Extraction

Before listing a few more dividing rules, it is worth highlighting three groups of methods that appear frequently under the above rules. One such group is feature importance quantifications. Feature importances, both global and local, appear frequently in XAI, and there are a large number of algorithms that generate them: SHAP, as already introduced, derives them from coalitional game theory (Lundberg and Lee, 2017: 4); LIME derives them from a fitted surrogate model on synthetic data (Ribeiro et al., 2016: 10); Distill-and-compare (Tan et al., 2018) also trains a transparent model to source feature importances from; Koh and Liang (2020) derive them from the application of influence function, which are a comparatively vintage statistical technique. Notably, there is overlap with some of the groups introduced above and below: feature importances are often local but sometimes abstracted to be global, such as SHAP and LIME, most are model-agnostic but there are model-specific versions, specifically for tree-based and deep models (Lundberg and Lee, 2017), some generate explanations ‘directly’ and some fit a surrogate model, and of course, linear and logistic models have ‘built-in’ feature importances (Molnar, 2020: 4). As such, this group of models is simultaneously very distinctive and very connected to other groupings.

Another grouping which is frequently found throughout these taxonomies are surrogate-based methods. These are techniques which extract insights from models by fitting a simpler, interpretable model to a black-box model's input-output pairs (Molnar, 2020: 5.7). They operate either globally or locally, though this distinction is blurred somewhat in operation, as is highlighted below. Though surrogate models are particularly well-suited to being model agnostic, as they can be generated using only input-output pairs, they are notably very common in the category of model-specific post-hoc explanation (Guidotti et al., 2020: 11). Burkart and Huber (2020: 27) divide global ones into 'pedagogical' and 'decompositional', whereas the latter makes use of the underlying structure, so is model-specific. They provide an exhaustive listing of global surrogate-based models and their class on page 29, which is not reproduced here. Local surrogate-based methods include SHAP and especially LIME (Lundberg and Lee, 2017; Ribeiro et al., 2016). LIME and its successor Anchors (Ribeiro et al., 2016; Ribeiro et al., 2018) blur the line between local and global somewhat, but in different ways. LIME builds a model using synthetic representations of the entire dataset, but explains a singular prediction – and beyond this, can be weighted to be more or less locally accurate at the price of generalizability (Ribeiro et al., 2016: 6). Anchors are also rules based on global analysis of the dataset, but they only apply to the subset of the dataset they happen to cover (Ribeiro et a., 2018: 1). As such, a clear distinction is difficult.

The third frequently occurring grouping are rule-based methods. These were already encountered under transparent models above, but model-agnostic rule-based methods exist too. These include G-REX (Johansson et al., 2014), Anchors (Ribeiro et al., 2018), or PALM (Krishnan and Wu, 2017). The former two generate if/then rules, the latter creates a tree first and then explains its branches with sub-models (Krishnan and Wu, 2017: 2). In general, many of these rule-based methods work by generating decision trees in one form or another, then extracting rules from them – as such, there is considerable overlap with surrogate-based methods.

(iv) Explanations for Neural Networks

Another approach worth highlighting is the differentiated treatment of methods aimed at explaining neural networks (NN). There is ample justification for this approach: the interna of neural networks are decidedly different from most 'shallow' techniques, which has spawned a set of purpose-specific methods. The techniques are further subdivided in a handful of different ways: for example, Molnar (2020: 7) distinguishes between learned features, pixel attribution, and concept activation vectors; Gilpin et al. (2019: 3) distinguish between explanations that focus on the processing of data by the network or on the representation thereof; Arrieta et al. (2019: 19) distinguish between techniques for deep, convolutional, and recurrent neural networks. One class which appears frequently are saliency maps, which were introduced in section 2.1. Especially gradient-based saliency maps are a very active area of research (Vilone and Longo, 2020: 24). Beside Grad-CAM (Selvajaru et al., 2016), algorithms to generate them include Smoothgrad (Smilkov et al., 2017), DeepLIFT (Shrikumar et al., 2017), and DeepResolve (Liu, Zeng, and Gifford, 2019).

Beyond saliency maps, another large class of neural network explanations are other visual explanations, which either generate visualizations of model interna or graphics that attempt to synthesize some of the information available about neural networks. Examples of the former are TensorBoard (*TensorFlow Developers*, 2021**)** or visualkeras (Garikov, 2021), examples of the latter are cnn-inte (Liu et al, 2018), which generates a scatter plot of neurons based on dimensionality reduction techniques, or TreeView (Thiagarajan, 2016), which adopts the hierarchical feature-space partitioning of decision trees. Another class of NN explainers are rule-based methods, which Hailesilassie (2016: 3) splits into ‘decompositional’ and ‘pedagogical’, which in more recent terms are, broadly speaking, model-specific and model-agnostic methods. By and large, these produce the same output as model-agnostic rule-based methods (Vilone and Longo, 2020: 29).

A final set of NN explanations are those that are to some extent ‘native’ to the model. The first type of these are explanations based on attention mechanisms, and the mechanisms themselves, which have seen plenty of research over the last few years owing to their superior predictive performance (Vaswani et al., 2017: 1). Though a body of work including Vaswani et al. (2017: 1) argues that the attention weights outputted are model and interpretation in one, moving neural networks closer to the domain of interpretable models, Jain and Wallace (2019: 2) posit that the attention weights outputted are much less expressive about feature importances than they are touted to be, and in general this field is still developing very actively. The second notable type of ‘native’ explanations are concept activation vectors (Kim et al., 2017). These allow the quantification of the sensitivity of the model against user-defined concepts, such as the dependence of a prediction of ‘zebra’ on the concept ‘stripe’ (Kim et al., 2017). While these are most intuitively applicable to image processing, they can also be used on NN’s processing text or tabular data.

(v) Example- or Data-based Explanations

A further division of methods is that of example- or data-based explanations, which may seem similar to local explanations, but it specifically refers to methods that select one instance of the dataset to explain either the data or the model more broadly (Molnar, 2020: 6). This includes counterfactual-based explanations, such as those Wachter et al. (2019) posit, adversarial examples, as presented by Fidel, Bitton and Shabtai (2019), which are counterfactuals with the aim of deceiving, not interpreting; the generation of prototypes and criticisms, as Gurumoorthy et al. (2019) do with ProtoDash, and the identification of influential instances, such as those found by Koh and Liang (2020). These approaches are united by their appeal to Adadi and Barreta’s (2018: 136) findings that humans like to reason in terms of examples, and all of them represent singular datapoints in one way or another. Beyond this, however, they differ considerably, both in explanatory approach and in output format. Some come close to feature importance quantification, like Koh and Liang’s (2020) application of influence functions; others present little other than a representation of the dataset through prototypical datapoints, like ProtoDash (Gurumoorthy et al., 2019: 6). As such, evaluating the group as a whole appears difficult.

Two subgroups of example-based methods warrant further discussion: the generation of counterfactual-based explanations and the generation of prototypes. There exist a handful of different algorithms for the selection of prototypical data samples: Bayesian teaching, set cover optimization, case-based reasoning, and ProtoDash (Gurumoorthy et al., 2019: 6), to name a few (Vilone and Longo, 2020: 22). Collectively, a set of prototypes is meant to represent the entire space of the data, and criticisms are meant to represent the data worst represented by the prototypes (Molnar, 2020: 6.4). The output of prototype-based methods is generally some representation of the datapoints, be that visual or tabular, and potentially their path through the model (Gurumoorthy et al., 2019: 8). One superior category of prototype and criticism methods are data generation methods, because they rely on the creation of datapoints that are not in the original dataset. This slightly larger umbrella also includes LIME, which learns a locally interpretable model based on a synthetic dataset it generates (Ribeiro et al., 2016: 1).

(vi) Counterfactual Explanations

The other sub-group of example-based explanation is counterfactual explanation. Counterfactuals are a plentifully discussed topic in the philosophy of explanation, as section 2.2.2 clarifies, so it comes as no surprise that methods aiming to generate them are gaining in popularity too. It is posited by Karim et al. (2018: 2) that counterfactual interpretability is the highest level of explanation attainable for ML. There is some consensus on how counterfactuals should be generated and evaluated. Moraffah et al. (2020: 6) state that counterfactuals should "answer why-questions by performing minimal changes to the data". They also posit additional evaluation criteria for 'good' counterfactuals: sparsity of perturbation, interpretability, proximity to the original instance, computational speed, and that counterfactuals for two different points are not identical (2020: 10). Overall, though Stepin et al. (2021: 11997) lament the lack of standardized measurement, they too agree that the definition of a counterfactual in the context of XAI, that of a minimal set of result-changing feature modifications, is fairly well-accepted. Wachter et al. (2019: 23) provide a much-cited approach, phrasing the finding of a counterfactual to a given datapoint as a constrained optimization problem, where the outcome of the model should be different, but the counterfactual datapoint should be minimally different from the original, as measured by Manhattan distance.

There is a deep link between counterfactuals and causality, but the creation of causal models, and the search for causality in supervised models in general, is contested. Molnar et al. (2021: 19) posit that XAI methods "do not generally provide" causal interpretations into the underlying data, and that supervised learning methods are "not designed to model causal relationships but to merely exploit associations". Paez (2019: 19), too, states plainly that "machine learning is the kind of context in which one can say that, in principle, it is impossible to satisfy the factivity condition for understanding-why", owing to the extreme complexity of the black-box functions implemented. More practically, Van der Waa et al., (2020: 17) find in a study that while contrastive rule-based explanations may help in identifying a situationally critical factors, they are not helpful in predicting model behavior in novel situations, and posit this is because they fail to clarify the underlying rationale of system behavior. In a similar user study, Sokol and Flach (2020b: 240) find that interactivity is key, in that user discovery of counterfactuals through self-posed what-if questions allows a deeper understanding of the model.

2.2 Explanation in the Philosophy of Science

This section summarizes the major models of explanation as devised in the philosophy of science, with no regard – at this stage – for acute relevance to XAI. Synthesizing these models into a concept of ‘explanation’ against which to evaluate XAI approaches is the major contribution of this work. As quickly becomes clear below, there are several hurdles to this task posed by incompatibilities between the models, let alone by the active debate around the very idea of unifying many accounts of explanation under one model.

Section 2.2.1 presents a handful of notable accounts of explanation in their historical order of appearance: deductive-nomological (DN), statistical, functional, mechanical, and context-pragmatic models of explanation. Section 2.2.2 covers accounts of explanation more intimately concerned with causality and counterfactuals: causal explanation as first presented by Salmon (1984), its extension into counterfactual explanation by Woodward (2003), and the attempts to develop a monistic counterfactual model of explanation. Section 2.2.3 addresses unification accounts and the state of the monism-pluralism debate more generally. Finally, section 2.2.4 highlights two recent attempts to move past this debate: reconciliation approaches, and a pragmatic reframing of explanation through the lens of understanding.

2.2.1 Notable Accounts of Explanation

This section provides a historical perspective of notable accounts of explanation: Hempel and Oppenheim’s (1948) deductive-nomological (DN) model, Hempel’s (1965) Inductive-Statistical (IS) model, Salmon’s (1971) statistical relevance model, and functional and new mechanist models. Salmon’s later causal-mechanical approach is grouped with other causal models in section 2.2.2. The models are introduced, their continued relevance to the theory of explanation is highlighted, and some more recent work that derives from them is presented.

(i) The Deductive-Nomological Model

Hempel and Oppenheim (1948) presented the deductive-nomological (DN) model of explanation. This has made a lot of people very angry and been widely regarded as a bad move. The model poses two innocuous-sounding requirements of explanation: one, that it takes the form of a deductive argument, where the explanandum follows from the explanans, and two, that the explanans include at least one "law of nature" as an essential premise (Woodward and Ross, 2021: 2.4). They posit that explanation, specifically, "to answer the question 'why?'", is one of the core objectives of rational enquiry (Hempel and Oppenheim, 1948: 1).

The model has been largely discredited and iterated on, which has itself been a fruitful process because many more recent models are rooted in criticisms of DN (De Regt, 2011: 2). There are two main avenues of criticism. Firstly, the imprecise definition of a law makes it impossible to distinguish one from an accidental generalization, and arguably impossible to truly define a law at all (De Regt, 2011: 2). The second avenue of criticism is the non-directedness of the model, allowing some effect-to-cause reasoning to pass as explanation under it. Salmon (1989: 47) famously exemplifies this with a flagpole casting a shadow: the length of the shadow can be explained by the height of the flagpole and the position of the sun, but the height of the flagpole cannot be explained by the shadow and the sun, though both are considered explanations under the DN account.

(ii) The Inductive-Statistical and Statistical Relevance Models

Two early influential models of explanation focus on the incorporation of statistical laws. The first is the inductive-statistical (IS) model posited by Hempel (1965). It is a relatively intuitive extension of the DN model, in that it incorporates not just 'static' laws but statistical ones, attributing not the certainty that deductive reasoning provides to events but a high likelihood, as generated by statistical laws which apply to the explanandum (Hempel, 1965: 331).

Salmon (1971, in Salmon, Jeffrey, and Greeno: 29) proposes an iteration on this called statistical relevance (SR). According to Salmon, a factor is considered statistically relevant in an explanation if, conditioned on its appearance or non-appearance, the probability of an event changes (1971: 31). This definition produces two key contrasts to the IS model. Firstly, the SR account does not conceive of explanation as argumentation, but as an assembly of information (Woodward and Ross, 2021: 3.1), which Salmon (1989: 53) claims changes what is demanded of its structure, specifically that irrelevancies are fatal in explanations but harmless in arguments. Secondly, Sober (2020: 19) describes that Hempel's IS model requires that a hypothesis explaining an event must assign it a probability greater than 0.5, which Salmon's model sensibly does not require. An example that is cited frequently is that of a syphilis patient developing a very rare side effect, which would be reasonably explained by referring to the disease, despite its probability being very low (see De Regt, 2011: 4).

(iii) Functional and Mechanistic Explanation

Functional explanation has strong roots in Aristotelian philosophy, specifically in teleology, the idea that systems in nature are goal-directed (De Regt, 2011: 9). As such, many authors in functional analysis restrict themselves to domains of science where such systems are readily found, such as biology. Functional analysis aims to explain the presence of elements of systems in terms of the goals of these systems: as Cummins (1975: 741) puts it, "the point of functional characterization is the to explain the presence of the item that is functionally characterized". Under the DN model, Hempel equates this to physical explanation, as subsumption under laws that govern the system are required (1948: 12). Both Hempel and Nagel (1961: 403), who is another key author on functional explanations, demand that the functional presence of items is explained deductively. Cummins (1975: 764) argues for a loosening of this approach.

Mechanistic explanation focuses, instead of on goal-directedness, on the organized interaction of parts of systems (De Regt, 2011: 11). Both mechanistic and functional explanations focus on the study and understanding of systems of explanans, however mechanistic explanations rigorize this somewhat. Paez (2019:19) states the latter "relies on an appreciation for functions, goals, and purpose" while the former "relies on an appreciation of parts, processes, and proximate causal mechanisms". Mechanistic explanation is not necessarily causal, but as becomes clear in section 2.2.2 below, distinguishing cleanly between strictly functional, mechanistic, and causal explanations is difficult at best and, more centrally, actively counterproductive for the aims of this work.

Compared to other explanatory approaches, mechanistic explanation has been subject to considerable iteration in recent years. A frequently cited source for the definition of mechanistic explanation is *Thinking about Mechanisms* by Machamer, Darde, and

Craver (2000) (see de Regt, 2011: 11; Rice and Rower, 2020: 9). In this work, the authors define mechanisms as structures consisting of entities and activities, the latter producing change and the former engaging in activities (2000: 3); the parallel to functional explanation is clear. Before them, Bechtel and Richardson (1993) framed explanation as ‘decomposition and localization’, with ‘decomposition’ referring to the piecewise understanding of systems and ‘localization’ referring to the assigning of tasks to the decomposed elements of systems. Craver and Darden (2013) and Glennan (2017) develop the theory of mechanisms further, but constrict themselves to the life sciences, which again is frequently the domain functional explanation is restricted to too. One issue with mechanistic approaches is that they are too descriptive and lack normative power (De Regt, 2011: 12).

(iv) Contextually Pragmatic Explanation

Pragmatic explanation may be best described as a “parallel strand” to the interconnected web of models, iterative improvements, and counterexamples that spans from DN over statistical, mechanical, functional, and causal models of explanations. This section introduces the original school of pragmatism presented by van Fraassen (1980). In section 2.4, a new iteration of pragmatism focused on explanation is detailed.

Van Fraassen (1980) develops the pragmatic theory of explanation around the key concept that what exactly an explanation should be depends on the background of its audience. This contextual reframing is guided by the view that explanations are evaluated vis-à-vis questions, which request information, but what information is requested varies between contexts (1980: 156). Two notes on this view’s interaction with the previously discussed models bear making. Firstly, van Fraassen’s audience focus is not inherently incompatible with Hempel’s and Salmon’s DN, IS, and SR accounts, nor with functional nor mechanical explanation – however his subsequent use of this focus to reframe what explanation ‘is’ does arguably disqualify attempts to build models of it at all (Woodward and Ross, 2021: 6). Secondly, the contextuality van Fraassen describes evokes the concept of explanatory relevance that Salmon’s SR model introduces, which threads through some of the subsequent accounts of explanation too.

Knowing that van Fraassen’s account is characterized by its contrast to the rest of the field, it goes without saying that there are numerous fundamental objections to it. One of these is posed by Kitcher and Salmon (1987:315), whose central complaint is that the three-point relevance relation between theory, fact, and context which van Fraassen posits is unconstrained, which means any pair of true propositions “explain” each other so long as they are related.

2.2.2 Explanation, Causality and Counterfactuals

This section aims to highlight models that establish or acknowledge the deep link between explanation and causality, and further, the link between causality and counterfactuals. This is a rich category. As Paez (2019: 10) highlights, understanding ‘why’, according to Aristotle through Salmon, is simply knowledge of causes.

The first and potentially most influential account in this category is Salmon’s (1984) causal-mechanical (CM) model. Its key concept is that causal processes transmit a ‘mark’, a persistent local modification in the process’ structure, and explanation is the description of the causal process that produced it (De Regt, 2011: 6). As is tradition, this model has been thoroughly criticized. Two key issues are raised: firstly, it distinguishes between causal and non-causal processes, but not between a process’s explanatorily relevant and non-relevant elements; secondly, there are a number of ‘action at a distance’ examples, such as key laws in physics, that are clearly explanations but do not involve anything representable as a ‘mark’ (Woodward and Ross, 2021: 4).

Despite these criticisms, the CM model served as the baseline for iteration, particularly into the domain of counterfactual explanation. In his influential paper *Making Things Happen*, Woodward (2003) formalizes the relation of causality and counterfactuals, and specifically a characterization of counterfactual explanation. He states (2003: 6):

“An explanation ought to be such that it can be used to answer what I call a what-if-things-had-been-different question: the explanation must enable us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways.”

This itself has been iterated upon in various directions; as described in the next section, Reutlinger (2016) evolves it to claim counterfactuals can be used as a basis on which to unify all theories of explanation. Strevens (2008) iterates further upon Woodward (2003) to instead provide the ‘kairetic’ account of explanation, which focuses on optimization of explanation to identify the ‘difference makers’ for an explanandum, which he claims can act as a ‘standalone explanation’.

In general, the relationship between counterfactuals, highlighting important causes, and the ‘background’ context of the explanation is strong. Potochnik (2015: 1163) claims causal patterns make sense of many aspects of how explanations are used in practice, including the use of equilibrium explanations and the relationship between explanations of events and explanations of causal regularities – specifically, that the causally regular ‘equilibrium’ state of a system is frequently contrasted against. Botterill (2010: 5) also highlights that selecting for relevant, adequate factors is particularly important in causal explanation, because all events have an infinite causal history. Northcott (2008: 15) formalizes the notion of selectiveness further. He derives an explicit theory of explanatory weight from a rigid definition of contrast classes and counterfactuals, from which he derives a few key insights: underlying causes need not be more important than proximate ones (2008: 9); several different causes can each explain most of one effect (2008: 11); and small causes are not necessarily less important than big ones (2011: 14).

More generally, Woodward and Ross (2021: 7.1) summarize how failing to account for causality has been the Achilles’ heel of many of the other presented models of explanation. The example of a flagpole casting a shadow Salmon (1989) used to discredit Hempel’s DN approach for being non-directed has already been mentioned. Additional instances Woodward and Ross (2021: 7.1) mention include Salmon’s earlier SR account being defied by different causal structures underlying identical statistical relationships, Salmon’s CM account failing because sometimes causal relationships hold without a connecting causal process, and rejection of Kitcher’s (1989) unificationism, which is introduced below.

2.2.3 The Monism Debate

Explanatory monism or unificationism, the idea that there is one theory of explanation that either encompasses all others, or allows one to disregard all others, is a very actively debated ideology. Two early prominent proponents were Friedman (1974: 5), who argued the mechanism by which science increases our understanding of the world is the reduction of the number of independent ‘ultimate’ phenomena, and Kitcher (1989), who rephrases this in terms of the ‘explanatory store’, the “maximally unifying systemization of statements accepted by the scientific community” (De Regt, 2011: 4).

Nothing about Kitcher's account strictly requires causality or even deduction, though he does claim all explanation is deductive "in a certain sense" (1989: 448), and that "the 'because' of causation is always derived from the 'because' of explanation" (1989: 477), meaning causal claims are derived from efforts at unification. Yet, the treatment of causality has been an avenue of criticism of Kitcher, specifically, that his unificationism doesn't specify why effects have to be explained in terms of their causes and not vice-versa (Woodward and Ross, 7.1 :24). A more general criticism is levied by Woodward (2003: 9), who disagrees with the implication that in any domain only the most unified theory is explanatory, and further argues that Kitcher's account does not distinguish between explanatory and non-explanatory information.

While Kitcher's model was moved on from, the debate around unificationism continued, characterized by the same rhythm of iterative advancement that defines the field at large. An example of a later monist is Nickel (2010), who notably rejects the idea that a unified theory of explanation can be causal, based on a claim called 'contrast fixes relevance': that the same contrast can be referenced from different domains, implying the 'relevance relation' by which this contrast is addressed from each of them must be identical – and some domains don't feature causal explanation, so causality can't be this relevance relation (2010: 15). He does not, however, argue this constitutes a case against unificationism through some other lens (2010: 24). Diez et al. (2010) do: they reject Nickel's argument based on multiple flaws, not least that he provides "no compelling illustrations of domain-invariant theories of explanation", and conclude that domain-specific explanation remains justified (2010: 14).

An avenue that has been explored more in recent years is unificationism around counterfactuals. A key text in this category is Reutlinger (2016), who argues that counterfactual explanation unites causal and non-causal accounts of explanation in a singular framework. Again, this has not gone uncontested; Roski (2021: 1972) argues that this is disproved by grounding explanations, metaphysical explanations that hierarchically relate an entity to a more fundamental level, such as a table to the atoms it's made of. Roski however argues this mustn't be the "death knell for monism", highlighting that 'backing models' of explanation may be able to accommodate both Reutlinger's (2016) counterfactual approach and grounding: these posit that information about determination or dependence between entities is 'backed' by a relation such as causality or grounding (Roski, 2021: 1989).

2.2.4 Moving Past the Monism Debate: Reconciliation and Pragmatism

(i) Reconciliation

Two strands of work attempt to reconciliate between pluralism and unificationism. The first, posited by Hochstein (2017), claims that explanation necessitates the combination of multiple methods. Hochstein (2017: 5) argues that this is because there are five explanatory goals, which are not jointly fulfillable by any one model: understanding how an event occurs, allowing the anticipation of this event, identifying patterns this event adheres to, identifying the mechanism that generates it, and providing information needed to intervene in the event. As some of these goals are, to borrow XAI vernacular, local and others are global, there is always a trade-off between which of them a model can address, so the only solution is to rely on a collection of models which jointly exhaust these aims (2017: 27).

The second strand, presented by Rice and Rohwer (2020), is reframing of explanation as a 'cluster concept', where multiple subsets of features are sufficient for a statement to be an explanation, but no one feature is sufficient in itself. They arrive at this conclusion through an exhaustive analysis of different models of explanation, including all the ones presented in this literature thus far, and distillation of them to their core requirements – which reveals that the only true commonality is that all accounts require their explanans be true (Rice and Rohwer, 2020: 16). As such, they propose that different clusters of requirements – such as temporal order, citing of counterfactuals, connection of disparate features, or generality – must be fulfilled jointly for a statement to pass as explanation (Rice and Rohwer, 2020: 18).

(ii) Understanding-Based Pragmatism

The second school of papers that attempts to move past explanatory monism, and to some extent past a model of explanation in general, are those that attempt to reframe explanation through the lens of understanding. Traditionally, understanding has been treated as a result of explanation: for Salmon, for example, understanding "results from our ability to fashion scientific explanation" (Paez, 2019: 5). It is this relation that is reversed by understanding-centered pragmatism. Wilkenfeld (2014: 3368) describes explanations as 'just' those things that create understanding, abandoning the formal structures much of the field creates around them. He also provides a reason why this school of thought has only bloomed recently, with most papers appearing in the last decade: the theory of understanding has only now been developed thoroughly enough to be able to make these sorts of claims (Wilkenfeld, 2014: 3368). It is clear why this sidesteps the iterations of models introduced above – as Wilkenfeld (2014: 3368) puts it, these were merely “adding epicycles to a fundamentally misguided approach”.

Exactly defining and measuring understanding is a field of research of its own; some notable accounts, especially those relevant to models of explanation, are selected here. De Regt and Dieks (2005: 137) identify as understanding the ability to ‘work with’ the understood, that is, to make extractive insights about relevant factors circumstances to an event. Knuuttila and Merz (2009: 8) argue that understanding is more broadly possessing an ‘appropriate model’ of the understood. Rancourt (2016: 77) introduces the ‘management account’ of understanding in the vein of de Regt and Dieks (2005), which defines understanding as the ability to extract relevant information and exploit it to answer questions about a subject. One commonality that Wilkenfeld (2014: 3397) highlights is that understanding is subject-sensitive, but not subjective – it may require different processes for different people to generate understanding, but whether they then possess it is clearly measurable. Further, Khalifa (2011: 1153) reduces understanding to the ability to generate explanations, which Strevens (2013: 18) echoes in saying that understanding is to grasp, and be able to produce, an explanation.

The monikers of ‘context-centered’ and ‘understanding-centered’ pragmatism, to distinguish pragmatism in the spirit of van Fraassen (1980) from this new approach, are introduced in this work and are not canonical. The two are not necessarily compatible: Paez (2019: 6), a strong proponent of the latter, argues context-centered theories of explanation like van Fraassen’s lack objectivity.

2.3 Humans and XAI

While there are few attempts to formally reconcile XAI with philosophical theories of explanation, a growing body of work does attempt to understand and measure what humans expect of XAI methods, and why it is so desirable to have explanations in the first place. This section summarizes the elements of this research that are relevant to this paper’s focus, in two parts: the reasons for finding explanations, and attempts to concretize what a ‘good’ explanation is.

2.3.1 Why Do We Want Explanations?

The most-cited reason we want explanations is ‘trust’ (Vilone and Longo, 2020: 7; Mueller et al., 2019: 3). From this broad notion, many more specific reasons for explanation are derived: ensuring fair and ethical decision-making, justifying the decisions made by a system, evaluating the robustness of a system, evaluating the transferability of a system, and gaining insight into how decisions would change were the input different (Adadi and Barreta, 2018: 143; Vilone and Longo, 2020: 7; Arrieta et al., 2019: 5).

The importance of these requirements is occasionally exemplified in some pertinent domains: in the medical domain, a diagnosis must come from a trusted expert; in the judicial domain, AI that predicts or judges criminality, or informs legal decisions, must be able to explain its judgements; in the financial domain, an AI must be trusted to behave rationally when investing (Rudin, 2018: 1). As the first law to integrate a ‘right to explanation’ (Doshi-Velez and Kim, 2017: 1), the European General Data Protection Regulation (GDPR) also warrants brief mention. It attributes three goals to explanation of AI: informing about the process that led to a decision, giving grounds for the contest of a decision, and understanding what needs to change for the decision to change (Wachter et al., 2019: 23).

2.3.2 What Makes a Good Explanation?

Miller (2017)’s *Insights from the Social Sciences* is as canonical as a four-year-old text can be, having arguably co-initiated, or at least revitalized, the entire sub-field of XAI research concerned with practicality and evaluation of methods. Based on an exhaustive survey, Miller (2017: 2) cites four concise insights about what humans expect of explanations: explanations are selected, meaning they focus on the most important features of the explanans; explanations are contrastive, in that they compare an instance to similar instances that led to different results; explanations are social, taking the background knowledge of the explainee into account; and probabilities do not matter in explanations, in the sense that we accept low-probability events as true if we perceive they are reasoned well.

Because this is not the scope of this paper, a large body of research that iterates on Miller’s findings, largely validating them, is not covered here. That said, some intuitive notions about how his goals are achieved may be less grounded than we assume. Adadi and Barreta (2018: 15) mentioned that some studies posit that transparent and fewer-input models don't lead to higher trust or lower prediction error, but better simulation of models' predictions. They also cite some work that questions whether explanations are necessary at all, because generally opaque systems are positively perceived, and it's costly to view and internalize an explanation (Adadi and Barreta, 2018: 160). This is refuted entirely by de Graaf and Malle (2017: 20), who find that humans tend not to trust opaque systems even where they perform with 100% accuracy.

One point Miller takes up is whether one should distinguish between scientific and 'everyday' explanation. Miller (2017: 5) states everyday explanation addresses 'why particular facts occurred' rather than general scientific relationships, which he argues is what XAI should strive to emulate. Mittelstadt, Russel and Wachter (2019: 5) also make the distinction, but in scope more than in structure; to them, the divide lies in the 'degree to which the entire causal chain and necessity of an event can be explained'. This mirrors the consensus found in the philosophical literature, which also states that scientific and 'everyday' explanations share structure (Woodward and Ross, 2021: 2.1). As such, for this work this distinction is best framed as a scoping problem, analogous to many of the ones presented in the philosophy of science.

2.4 Evaluating XAI

Having established the motivation of the search for explainability, the notion of 'good' explanation, and the main families of tools this has spawned, what follows is a brief summary of the attempts that have been made to formalize and quantify the evaluation of XAI solutions. The lack of consensus on this topic must be restated before these accounts are presented. As Vilone and Longo (2020: 51) phrase it: "there is no agreement among scholars on what an explanation exactly is and which are the salient properties that should be considered to make it effective and understandable for endusers, in particular non-experts".

An oft-cited work attempting to formalize this is Doshi-Velez and Kim's (2017: 4-5) presentation of three categories of explanation: application-grounded explanation, which poses humans real explanatory tasks, and compares how an XAI method's explanation compares to a human-generated one; human-level evaluation, which measures how well humans accept explanations made by XAI methods; and functional-level evaluation, where methods are measured against some proxy attribute already known to correlate with human understanding, eg. the depth of a decision tree.

There are a good number of varyingly concrete requirements of XAI methods that are repeated across evaluation frameworks and methodologies. These include performance of the underlying model (eg. Fauvel et al., 2020), comprehensibility – the comparison between white- and black-box models (eg. Fauvel et al., 2020), fidelity – truthfulness to the underlying model (eg. Alvarez-Melis and Jaakkola, 2018: 2), or consistency of explanation – for one model between similar instances, or for one instance across models (eg. Molnar, 2020: 2.4). The approach of this work does not require an exhaustive listing of these, as these visibly tend to be derived from the realities of XAI, whereas this work goes the exact opposite direction – but some implications about their efficacy in ensuring good explanations, from the perspective of the model developed in section 4, are discussed in section 6.

3 Scope and Approach

As the literature review makes evident, both areas this paper considers – active XAI research and theories of scientific explanation – are far too broad to cover exhaustively in the frame of this work. As such, this section serves to clarify both the scope of the paper and the approach with which the analysis was conducted.

For two reasons, the paper only considers explainability methods for supervised models. Firstly, while there is some research into explainability for goal-based or reinforcement learning, section 2.1.2 highlights that it only constitutes a minor portion of the field. As such, it is neither a gross misrepresentation of the field to disregard it, nor does it significantly diminish the relevance of the work presented. Secondly, and more importantly, unsupervised learning contains several structurally very different methods (Puiutta and Veith, 2020: 1), the dismissal of which allows for the tightening of some assumptions and requirements of the derived models of explanation. As quickly becomes clear below, numerous concepts from the theory of explanation are translated much more naturally onto ML when factors such as the type of output, input-output relationship, and forms of interaction with the model are standardized. In short, limiting the work to supervised models makes the aim of this paper far more tractable with an admissible loss of generality.

Secondly, it must be restated that the paper aims to evaluate explanation goodness strictly from the philosophical perspective of explanation. In doing so, it may well contribute to some of the other angles from which this issue is addressed, as described in section 2.3.2, but it does not seek to integrate social scientific or psychological considerations in its methodology. Work in these fields tends to be quite experimental (Miller, 2017: 2), introducing user and expert polls, which adds another level of complexity. If explanation is seen as the bridge between models and the goals of XAI outlined above, such as more trust in models and deeper acceptance of their workings, experimental research tends to create a direct connection between the former and the latter, ‘skipping’ the formalized notion of explanation (Miller, 2017: 1). This work instead focuses only on the connection between model and formalized explanation. Of course, this implies a host of further research questions, such as how adherence to formal models of explanation interacts with user trust and the other goals of XAI, which are elaborated further towards the end of the paper.

Finally, the direction in which the paper reasons about XAI and theories of explanation must be clarified. As indicated in the literature review, it is at some points enticing to make judgements about the efficacy of models of explanation based on their applicability to ML and XAI techniques. The modeling and analysis sections of this paper do not seek to make these sorts of judgements. Instead, they look to adhere strictly to what the leading types of explanation model require of explanations and evaluate XAI techniques from there, aiming to avoid falling in the trap of replicating the cart-before-the-horse structure. The discussion then does reverse the direction of analysis to consider some of the findings' implications for topics in the theory of explanation, such as explanation monism, the power of counterfactuals, and the applicability of functional explanation.

The approach of the analysis this paper conducts from here is straightforward: the broad background of the literature review on explanation is condensed into a model of explanation which has evaluable requirements of XAI methods. The literature review of XAI is condensed into a set of families of explanation to evaluate. Each of these processes of condensation are justified, the former more exhaustively than the latter, as it is the core of this work. The families are then evaluated against the model, without much of a formal structure of evaluation. The insights for XAI and for explanation are discussed. Though the research goal of this work is the evaluative process in section 5.2, much of the value produced stems from the condensational processes that precede it, especially the building of the model.

4 Modeling

The literature review covers existing accounts of explanation quite exhaustively. This section aims to synergize the many models presented and translate them into requirements against which XAI can be measured. As the literature review makes clear, there is a somewhat distinct ‘canon’ of accounts, which has evolved in an interconnected way, and beyond this a handful of ‘fringe’ accounts that either do not mesh well with the canon or define themselves through being opposed to it. The modelling reflects this: section 4.1 finds and justifies four generalities among accounts, and section 4.2 introduces three extensions to the model that derive from specific accounts of explanation.

4.1 Finding Generalities

This section establishes generalities among models of explanation and translates them into requirements for XAI. Considering the active debate around explanatory monism outlined in section 2.2.3, this objective appears ambitious. I argue that four factors, each of which is expanded upon where relevant throughout this section, justify it. Firstly, there are a handful of accounts of explanation that do not aim to account for the type of explanation ML models require, and misusing them to fulfill this task is largely fruitless. Dismissing these accounts of explanation makes unification of the remaining ones easier. Secondly, several of the objections and counterexamples to the remaining models are not applicable in the domain of supervised learning. As these objections tend to be anti-reconciliatory, justified ignorance of some of them greatly expands the scope of what can be argued as consensual. Third, many of the more recent attempts to move past the monism debate justify, or even depend on, either reconciliation between accounts of explanation through finding commonalities (eg. Rice and Rohwer, 2020), or an understanding-based pragmatism (eg. Wilkenfeld, 2014) that justifies the abstraction of not fully canonical requirements. And fourth, many of the more recent stringently monist approaches require little to nothing of explanation – ‘backing’ models, notably, require only entities and some form of connection between them (Roski, 2021: 1989) – such that there is flexibility to formulate requirements more concretely within them.

With this in mind, four general requirements for XAI are abstracted, for each of them (1) arguing that sufficient consensus exists to make them ‘core’ requirements of the model, (2) translating them to be measurable in the context of XAI, and (3) formulating them as a concrete requirement. To enable this, section 4.1.1 lays some groundwork for the translation of accounts of explanation into XAI.

4.1.1 The ‘Boilerplate’: Explanandum and Explanans in AI

Each of the requirements introduced below benefits from a formal translation of the concepts of explanandum and explanans onto AI. I argue that there are three different approaches to this, of which one does not apply, and that the other two reduce to different instances of the same concept. The approach I argue does not apply is the ‘intuitive’ one that leaves XAI methods aside, and considers some machine learning model M to be the explanation that connects the data D to a prediction P. Through the lens of functionally any of the philosophical accounts of explanation, the disqualification of this approach is obvious: each account poses at least *some* requirements of the explanation, and the model, black-box as it likely is, certainly cannot fulfil any.

Yet I argue that it is this conception of explanans and explanandum that is most commonly, if implicitly, applied. The expectation that is levied is that a black-box model explains the *relationship between input data and output prediction* that it finds and exploits, and because it does not do so natively, we require XAI methods to do this for us. The important insight is that these explanatory methods, X, have no place in this conception’s ‘ground’ level; rather, they serve to illuminate M. The impossibility of this task leads to the sort of defeatist statement Paez (2019: 19) makes: “the use of arbitrary black-box functions to make decisions in machine learning makes it impossible to reach the causal knowledge necessary to provide a true causal explanation”. It is lucky, then, that I argue this formulation misrepresents how data, model, explanatory method, and model output interact.

The second and third approaches both integrate the explanatory method X on the ‘ground’ level, but differ in that they are broadly scoped for global and local explanations respectively. Though section 4.1.2 argues in depth that the parallel consideration of global and local explanations is justified, this section concludes that these two approaches can readily be combined into one.

The second approach, considering ‘local’ explanations, scopes the explanans as the data D, the model M, and the explanatory method X, and the explanandum as a singular prediction P. This reflects the reality of how local explanations of AI are scoped: the input and training data, the model itself, and the method and output of an XAI technique are all explanans, and jointly from them, a singular prediction's explanation is derived. This account is far more sensible than the first when considering the sorts of interactions between explanans that accounts of explanation usually describe. Specifically, it affords space for the interaction between the model and the XAI method – even if this is just a call to a predict function – within the context of both of them being explanans, allowing (at least formally) the application of requirements accounts of explanations set of these interactions, such as causality, directedness, or explanatory relevance.

The third approach, considering 'global' explanations, scopes the explanans as the data D and the explanatory method X, and the explanandum as the model M. Notably, this approach integrates the model's interna and the model's output, the collection of its predictions, within M. An entire model is a much larger scope for an explanation than a single prediction, arguably too large for many models of explanation, which focus on singular events or phenomena, so this approach should also allow scoping of the explanandum to some subset of model or output. Where this is done, the elements of model and output that have been removed from the scope of the explanandum can be added to the explanans. Pushing this to its extreme reveals that it was not entirely truthful to present the two approaches as distinct: if all but one element of the model's output are removed, the second approach is identical to the first.

Regardless of whether the rigidly local, global, or flexible version of this divide is used, one commonality is worth highlighting: the explanatory method X is found on the side of the explanans in both instances, though one may intuitively expect it nowhere on either side, as it constitutes the explanation itself. Ideally, this would be the case. However, the explanatory method is far more than an explanation. It is a system, in some cases quite a complex one, that interacts with the model deeply or shallowly in ways that are not reflected by considering the entirety of X an explanation.

To summarize, the unified mapping of explanans and explanandum onto issues in XAI derived and justified in this section is as follows: $D + M\backslash P + X \rightarrow P$, where D is the input data, $M\backslash P$ is the model, including its output, excluding some prediction P, and X is the explaining method; P is a prediction (specifically a subset of the model's output space), whereas the scope of P can be varied between one prediction of the model (fully local) and the entire model and its output (fully global). In the latter case, the mapping reduces to $D + X \rightarrow M$. Having established this, the following sections return to utilizing it to find generalities across models of explanation.

4.1.2 Explanations Answer Why-Questions

I would posit that ‘explanations answer why-questions’ is the most agreed-upon consensus in the literature. The vast majority of the works reviewed make some reference to the concept of why-questions: van Fraassen (1980), Salmon (1969), and Kitcher (1989) frame their accounts around them; Botterill (2010: 6) highlights how causality is the search for answers to why-questions, Paez (2019: 10), too, reduces understanding-why to knowledge of causes. This prominence is likely owed at least in part to their role in Hempel and Oppenheim’s account of the DN model, which states that “to answer the question ‘why?’ rather than only the question ‘what?’ is one of the foremost objectives of all rational enquiry” (1948: 1).

Even so, the consensus is not absolute: Nickel (2010: 24) highlights in his conclusion that not necessarily *all* explanations answer why-questions, but some may answer how-questions, and further that why-questions imply causal explanations, the universality of which he spends his paper denying. This is addressed in the context of XAI below.

Turning to XAI, I argue that all why-questions that can be asked about supervised models fall into one of two categories: ‘local’ why-questions, which pertain to one datapoint, such as “why did my loan application get rejected?”, and ‘global’ why-questions, which pertain to more general behavior of the model, such as “why does the model reject more applications than a human?”. Of course, the latter category encompasses questions that are not truly global in the sense described in the XAI summary above, in that they may constrict themselves to only a subset of the input space (“why are demographic variables so important in this model?”), the output space (“why does this image classifier think it’s seeing a dog so often?”), or a set of input-output combinations (“why are people above 40 less likely to get accepted for a loan?”).

Two potential counterarguments may merit brief discussion. Firstly, the most salient counterpoint to binary classification may well be why-questions about model interna, such as “why does that node have such a high weight?”. Indeed, though there is room for a labored argument that model interna apply to all datapoints and are thus global, this does make binary classification feel unwieldy. There are two reasons I believe this approach is justified nonetheless: firstly, it reflects how we reason about models, which is often in terms of prototypes and sample datapoints (Adadi and Barreta, 2018: 136); secondly, the ability to interpret model interna supposes some prior knowledge of the model's workings that cannot reasonably be attributed to most people. The second counterargument is Nickel's (2010) hesitation about how-questions. I accommodate this by arguing that any reasonable 'how' question about a model can be rephrased as a why-question, either because it takes a global form ('how does this model work?') or a local one ('how was my classification decided'); in general, Nickel's objection appears too pedantic to rescope a pragmatic work like this for.

The question, then, is whether the model should demand of a given XAI technique to address one or both of these types of question. Here, it is yet again tempting to set the cart before the horse and realize that XAI techniques divide themselves quite neatly into global and local explainers, as section 2.2 showed, meaning requiring an answer to only one of these classes of questions would be a considerably easier criterion. I also argue for this criterion, but not for this reason. Rather, it is worth recounting what Hochstein (2017) claims about the aims of explanation. As the literature review detailed, he argues that all explanation has global and local aims, which cannot be simultaneously fulfilled by the same model. The insight this gives for our case is that no explanatory approach can be realistically expected to give both local and global explanations, and that this is not necessary anyways due to the possibility of combining models.

Having established the duality between global and local why-questions, what is left is to concretize what 'answering' a why-question is in the XAI space. This is particularly difficult given this usually encompasses the crux of the account of explanation. Leaning on 4.1.1's definition, a DN explanation may require a logically sound deduction of P from D, M\P, and X. A causal explanation may require a causal connection of factors within them. Van Fraassen's (1980) pragmatic account may require that the most relevant elements of D, M\P, and X to P and the explainee are highlighted, and so on. Also, the way in which these factors are connected – how deeply X interacts with the model, for example, or how the input data is processed through it – varies tremendously. Seeing that both the model-side and XAI-side approaches to this question defy meaningful distillation, it appears counterproductive to define the concept of ‘answering’ too strictly in the model, and makes more sense to instead appeal to the ability to analyze this question contextually for each case.

The requirement that is set, in short, is as follows:

R1: The explanation addresses either a local why-question, pertaining to one datapoint, or a global why-question, pertaining to general behavior of the model.

4.1.3 Explanations are Contextual

The second general criterion I posit is that explanations require some awareness of the context in which they are being presented. This takes different shapes in different models, but two common trends emerge: contextuality requires some consideration of what background knowledge the recipient of the explanation has, and contextuality requires that, against this background knowledge, particularly relevant elements of the explananda are highlighted.

Several accounts of explanation centralize this criterion. Salmon’s (1969) statistical relevance account requires the aggregation of features that affect the conditional probability of an event; van Fraassen (1980)’s pragmatic account is defined by exactly this focus on what the explainee already knows contextually, to the extent that he argues for the abolition of other accounts; Northcott’s (2008: 15) formalization of the notion of explanatory weight shows that the epistemological requirements we set of contextuality are very high, and we must content ourselves with approximation thereof. Even Wilkenfeld’s (2014) understanding-based pragmatism is contextual, even centrally so, in that it does not define anything as an explanation that does not generate understanding in the explainee, meaning it must of course be tailored to the explainee’s prior understanding. Beyond accounts that centralize this contextuality, most accounts with a ‘non-pragmatic core’ also are also broadly compatible with contextual awareness; this includes Hempel’s DN/IS and Salmon’s CM/SR models, or Woodward’s counterfactual account (Woodward and Ross, 2021: 6).

There have not been attempts to strictly translate this criterion onto XAI, however some formulation of it is frequently mentioned in the XAI literature. Miller (2017: 2), of course, does mention selectiveness and contextuality, but does not specify in depth how these should be measured and translated in the context of XAI. Arrieta et al. (2019: 7) also request that explanation be audience-focused, but also do not clarify what this requirement demands of XAI techniques. In general, work formulating these requirements largely describes how explanations ‘should be’, rather than evaluating whether current approaches fulfill these contextuality requirements.

In general, though this criterion is far too present in the philosophical literature to be ignored, it invites deviation from the core aim of this paper towards the inclusion of social scientific and psychological research. Scoping this criterion to the definition in 4.1.1, it only requires that elements of D, M\P, and X are highlighted to the explainee if they sit in the intersection of things that are particularly relevant to explanation of P, and things that the explainee does not already know.

This allows the formulation of R2:

R2: The explaining method demonstrates awareness of the background knowledge the explainee has about the model and its context, and against this background knowledge, highlights particularly salient features of the explanans for the explanandum.

4.1.4 Explanations Create Causal Links

The third general criterion I posit is that every relevant theory of explanation either demands or tolerates some form of causal relation between a set of events or circumstances and an explanandum. As Woodward and Ross (2021: 2.4) put it: “virtually everyone, including Hempel, agrees that many scientific explanations cite information about causes.” Explicitly causal models of explanation are exhaustively detailed in section 2.2.2; beyond this, I argue that many accounts of explanation do not necessarily require the establishment of causal relations but tolerate their inclusion. This includes Hempel’s (1948) DN and IS models. It includes Machamer, Darde, and Craver’s (2000) mechanical approach, which requires not causality in name, but the creation of directed relations between explicanda and explanandum, which amounts to a very similar concept. It also includes ‘old school’ context-centered pragmatism as presented by van Fraassen (1980) and ‘new school’ understanding-centered pragmatism as presented by Wilkenfeld (2014): why should the process that generates explanations not uncover causal relationships? Likewise, Potochnik (2015: 1182) claims that explanation of events and the description of causal regularities are closely related, creating a link between this criterion and the contextuality criterion above. Specifically, she argues that explaining events often requires explaining how causal chains usually operate to highlight the anomalies that caused an event.

Two objections to including causality in a general model arise from the literature review. Firstly, if "*many*" explanations are agreed by virtually everyone to cite causes, do all of them? I argue that all the ones relevant to XAI do. Reutlinger (2016: 41), for example, explicitly claims to unite causal and non-causal explanation under counterfactuals, with the latter exemplified by Euler's explanation and 'renormalization group explanations of universality'. Whether or not Reutlinger's counterfactual unification theory is valid, these examples are strictly mathematical explanations, which have no obvious parallel in XAI, justifying their ignorance. More recently, Roski (2021: 1971) claims the class of grounding explanations is a set of noncausal explanations that are not covered under Reutlinger's (2016) monist account. As highlighted in the literature review, grounding is a highly specific concept from metaphysics referring to hierarchical connections between different 'levels' of understanding of a concept. Parallels to XAI are again extremely labored at best; one could argue for example that a model's architecture is 'grounded' in computer science or the language it's written in, but this doesn't result in a good-faith counterargument. In general, I argue that one does not find an objection to explanations uncovering causality that is both general and accredited or is specific to a type of explanation relevant to XAI and accredited.

The second objection arises from the contested nature of the search for causality in XAI. However, I argue the doubts expressed by, for example, Molnar et al. (2021) do not prevent the translation of the requirement of causality onto XAI, for the key reason that one must distinguish between causal relationships in the underlying data and causal questions about the model itself. Even taking Molnar et al. (2021: 19)'s statement that supervised models simply do not represent causal relationships in the data as a given, it is reasonable to interpret why-questions such as "why did the model decide positively" as questions about the inner workings of the model, rather than as questions about causal relationships within the data. This conception is nicely embedded in the definition of explanans given above; as D, M\P and X are all elements of the explanans, it is reasonable to answer the above questions by referring to causal relationships between, say, M and X, or even M and D, which both provides a passable explanation and does not violate Molnar et al.'s criticism.

In summary, the requirement set by this section is as follows:

R3: The explanation illuminates a causal connection between some combination of data, model, and explaining method, which pertains to the explanandum.

4.1.5 Explanations are Factive

The final general criterion I posit is that all canonical accounts of explanation require their elements to be truthful. In this sense, it is not particularly important what exactly these elements are. Rice and Rohwer (2020: 16) establish that requiring the components of an explanation are true is the singular thing all accounts of explanation explicitly or implicitly have in common. In fact, a handful of evolutions of models of explanation arguably came from this requirement, most notably the rejection of Hempel's covering law and DN accounts because of the difficulty of defining and measuring a law's truthfulness that section 2.2.1 details.

Sections 2.3 and 2.4 of the literature review showed that requiring truthfulness is also a common component of the social scientific XAI literature, and of XAI evaluation frameworks. As opposed to the other criteria, the work of translating this criterion onto XAI has thus largely been done. Authors tend to apply the requirement of truthfulness in two ways: ensuring the XAI method's truthfulness to the model's behavior, and ensuring the model's truthfulness to the data. One could extend this to the data's truthfulness to reality, which also sees plentiful analysis across domains through to classical statistics, but this is well beyond the scope of this paper. Either way, what we aim to analyze for this criterion is simply whether what the explaining method presents, its output, is demonstrably accurate to model and data.

Despite it being arguably the most intuitive of the requirements, and enjoying inarguably the greatest consensus, factivity also sees the most principled pragmatist objection. Paez (2019: 7) has been cited a handful of times in this work for his belief that black-box models simply are not explainable in themselves, from which he deduces that at best, central propositions of an explanation must be true, and beyond this, we have no choice but to learn by analogy (2019: 11). If an uninterpretable model is not represented accurately, but an analogical version nonetheless generates understanding, then it appears reasonable to state factivity is not required from a pragmatic point of view. Pragmatism here is very at odds with what all above theories of explanation claim – as such, I would argue it is not justified to remove it as a criterion based on Paez's doubts alone. How this conflict resolves may well prove insightful for the debate around explanatory pragmatism in general. One attempt at reconciliation may be that clearly identifying the analogy as such avoids the violation of factivity.

Unlike the other three requirements, which apply to D, M\P, and X symmetrically, factivity applies to how X reflects D and M\P. As such, the focus lies not on the explanation but on the explaining method:

R4: Information depicted by the explaining method’s output is true.

4.2 Extending for Specific Models

Some models of explanation that were covered in section 2.2 are not entirely compatible with the others presented. Whereas the evolution from DN and IS accounts through statistical relevance and causal-mechanical models to modern causal and counterfactual approaches provides the broadly consensual criteria presented in 4.1, the three criteria outlined below are not posed often enough to be canonical. Nonetheless they represent influential accounts of explanation, so warrant being evaluated against; further, the inclusion of non-canonical criteria is fruitful for discussion of models of explanation in section 6.2 in that it provides information on XAI as a ‘use case’ to weigh in on active debates in the field.

4.2.1 Functional Explanations Rely on Goal-Directedness

The inclusion of functional explanation as a requirement is a largely investigative exercise. It is motivated by the realization that despite – or maybe because of – its focus on explanation of systems in the natural sciences, functional explanation appears to fit beautifully onto the domain of supervised learning. To recall two insights from section 2.2.1: functional explanation explains the presence of entities through the function of the system they are a part of, and a historical root of functional explanation is goal-directedness. Supervised learning systems are sectioned-off closed systems, with inputs and outputs, and a wealth of individual components. Furthermore, they are distinctly goal-directed, even quantifiably so, thanks to the loss function they aim to minimize.

The explanandum functional explanations target is some subset of the model M itself, not an aspect of the model’s output. The specific set of inquiries that functional explanation appears to address thus are why-questions about model interna, such as weights of nodes, coefficients, interactions of features, heights of parameters and so on:

E1: The explaining method illuminates what function an element of the model performs towards the generation of predictions.

4.2.2 Statistical Explanations Utilize Statistical Laws

This criterion aims to integrate the reliance on statistical laws found in inductive-statistical and statistical relevance accounts of explanation as detailed in section 2.1.1. Statistical grounding of explanations is far from a universal requirement across accounts of explanation, not least because many do not necessarily lend themselves well to translation into either Bayesian or frequentist terms. The inclusion of those that do in a criterion of this model is justified by the deep dependence of AI on statistical techniques. It goes without saying that each type of AI model centrally integrates statistical and probabilistic concepts in the training and prediction process, to the point where it can be taken as a given that some statistical law is embedded in any model being analyzed. The interesting question, then, is how this interacts with the way statistical laws are called upon in philosophical accounts of explanation.

To reintroduce the precise way statistical laws are embedded in these accounts: inductive-statistical explanation requires that at least one of the explanans is a statistical law that applies to the explanandum (Hempel and Oppenheim, 1948); statistical relevance requires of each explanans that its presence, absence, or value affects the nature of the explanandum (Salmon, 1969). These are notably different utilizations of statistical concepts, with the first appropriating them for a more structured deductive approach, obviously mirroring the DN model, and the second clearly setting the groundwork for what would later become more focused concepts of weighted explanation (eg. van Fraassen, 1980, Northcott, 2008), through to the selective contextuality formalized in the second criterion of this model.

Reconciling between these two approaches to phrase a general criterion of XAI is eased by thinking in terms of explanans and explanandum: both accounts place D, M\P, and X among the explanans, and both also subsume statistical laws under them. IS requires the statistical laws to simply be an explanans, possibly as part of the model M or as part of the explanatory method X; SR requires the statistical laws take the role of qualifying the parts of D, M\P, and X that are explanatorily relevant. I argue this does not contradict their placement among the explanans, either in model or in explanatory method; rather, this is their intuitive position in this binary, allowing them to be called upon in the process that generates the selections SR requires.

E2: Among the explanans, in model or explaining method, is a statistical law that is relevant to the explanandum.

4.2.3 Pragmatic Explanations Generate Understanding

Finally, the research strand of pragmatic explanation is not well-represented in the above criteria. As section 2.2.4 of the literature review made clear, it may by definition not be well-represented in *any* model of explanation. Its motivating concept is, after all, that modeling explanation by setting requirements of the explanation itself is condemned to failure. Including it as a criterion is thus an exercise that invariably will provide some insight on the strengths and shortcomings of pragmaticism.

The key reason I think inclusion of pragmaticism is justified is that despite its rejection of the over-definition of models of explanation, pragmaticism does not reject explanations that do adhere to these accounts as non-explanatory, so long as they generate understanding (Wilkenfeld, 2014: 2). As such, there is no reason a given explanation cannot both be adherent to the above model and understanding generating.

The question that is unanswered in the context of XAI is what ‘understanding’ a model, a decision, or generally the answer to a why-question is. The literature review highlights accounts by De Regt and Dieks (2005), Knuutilla and Merz (2009) and Rankourt (2016), the commonality of which is some sort of ability to identify, extract, and work with relevant information from the subject. An obvious parallel in AI is the ability to predict future model decisions, which incorporates local understanding – of those future datapoints – and global understanding of the model’s mechanisms. More locally, understanding may include intuition about how changes in the value of input variables affect the output of a model. Clearly, there is the potential for the blurring of the line between understanding the model and understanding relationships within the underlying data. As such, it may be a futile exercise to attempt to characterize understanding concretely in the XAI context, especially in a work not centered on this aim. Further, it appears counterproductive to attempt to characterize model understanding too rigidly without experimental measurements thereof – which already exist in the social sciences (see Mueller, 2019: 22) – especially considering that pragmatism aims to accommodate exactly this understanding.

Understanding-based pragmatism also does not translate well onto the conception of explanans and explanandum developed above; a labored attempt to do so would see much of the model and the data on the side of the explanandum. As such, this is not developed further here. The requirement, instead, is phrased generally, leaving room for interpretation in each instance:

E3: The explainee's understanding of the model's decision process is measurably improved by the explanation.

4.3 Summary: The Model of Explanation

The model of explanation established in this section has two types of requirements, core requirements (R1-R4) and extensions (E1-E3), which it levies of XAI methods. These were deduced rigorously from consensuses in the philosophy of explanation, which were introduced in the literature review and justified in this section. The requirements are re-listed here.

R1: The explanation addresses either a local why-question, pertaining to one datapoint, or a global why-question, pertaining to general behavior of the model.

R2: The explaining method demonstrates awareness of the background knowledge the explainee has about the model and its context, and against this background knowledge, highlights particularly salient features of the explanans for the explanandum.

R3: The explanation illuminates a causal connection between some combination of data, model, and explaining method, which pertains to the explanandum.

R4: Information depicted by the explaining method's output is true.

E1: The explaining method illuminates what function an element of the model performs towards the generation of predictions.

E2: Among the explanans, in model or explaining method, is a statistical law that is relevant to the explanandum.

E3: The explainee's understanding of the model's decision process is measurably improved by the explanation.

5 Analysis

5.1 Selecting Families of Models

As section 2.1 of the literature review established, it is neither feasible to analyze every technique in XAI, nor is it a particularly valuable contribution to analyze groups that are too broad, such as all post-hoc methods. This section aims to find the optimum between these points by proposing groups of methods that are feasibly evaluable against the model developed above while remaining representative of the field. As such, it addresses point (1) of the research aim outlined in the introduction.

The choice of groups was loosely framed as an optimization task with three constraints:

1) All techniques in the group have similar or comparable outputs, ideally in format (visualization, text, numbers, etc) but certainly in substance (eg. feature importance, rules, model interna).
2) All techniques in the group have a similar explanatory approach.
3) There is no superior group to which the group belongs which also fulfils criteria 1) and 2).

The first two criteria aim to ensure homogeneous evaluability against the model established above; the third criterion aims to ensure tractability by avoiding duplication of efforts. Below, the groups are listed, and their selection is briefly justified.

(i) Transparent Models

Section 2.1.2 of the literature review shows in depth that the separation of transparent models from post-hoc explainability techniques is canonical and justified. The ‘explanatory approach’ and substances of these models are suitably comparable; while they may differ in what exactly their model interna are, be it a linear regression with coefficients or a decision tree with rules, they are united by the idea that a human can reasonably comprehend the process by which the model makes decisions in its entirety. As the smallest superior group to which transparent models belong contains all XAI techniques for supervised models, which is excessively broad, constraint 3 is also satisfied.

(ii) Feature Importance Methods

As the literature review highlights, feature importance methods appear at various places in the taxonomy of XAI. Feature importances appear in transparent models such as regressions and trees, they sometimes build surrogate models like LIME, they sometimes directly work with the model like SHAP, they are sometimes local and sometimes global. Nonetheless grouping them together is amply justified according to the criteria listed above. All feature importance techniques (1) return a relative quantification of the importance of each feature in the input data. They each do so (2) through some variation of measuring the sensitivity of the model to changes in the input around a point. And due to their appearance across the taxonomy, their next-largest superior group are all XAI methods for supervised learning, or all post-hoc methods if model-internal feature importances are disregarded, which is far too broad to be tractable (3).

(iii) Surrogate Methods and Rule Extraction

Like feature importances, surrogate model fitting is ubiquitous in XAI. The literature review highlights how many explanatory techniques have a surrogate model as an underlying method. It also mentions that, despite their well-suitedness to model-agnostic explanation, they are frequently employed in the domain of model-specific post-hoc techniques, particularly for shallow models. Despite their ubiquity, their inclusion as its own group is not entirely trivial.

Surrogate model-based methods are united by the key feature of their explanatory approach, however they appear somewhat sporadically across the XAI taxonomy and, further, differ in output. That said, the limiting factor in the explanatory capability of surrogate-based methods is the interpretability of the secondary model they fit, which as the literature review highlights usually delivers either decision rules or feature importance values. Criterion (1) is thus satisfied only sparsely through generous interpretation of the ‘substance’ of the explanation: it must be considered to be the interna, or insights, of the surrogate model. Criterion (2) is satisfied handily; criterion (3) is satisfied because of surrogate models’ appearance across the XAI taxonomy.

(iv) Model Visualization

This category encompasses those techniques that specifically aim to represent interna of models, especially diagrams thereof such as decision trees, neural network charts, or plots of decision boundaries against data. It does not include secondary visualizations such as plots of the output of numerical XAI methods like feature importances. It also does not include other methods that produce visualizations natively, but do not primarily visualize the model, such as image saliency maps (see below).

While criteria (1) and (2) are readily fulfilled by this scoping, the fulfillment of criterion (3) requires some elaboration. The directly superior domain is visualization, but analyzing the efficacy of visualization for explanation in general is, as the literature review showed, an intractably large field of research, largely rooted in social and cognitive science, and scoped far beyond XAI (Offert, 2017). Because of the diversity of visualizations, there is no meaningful commonality in explanatory approaches other than the format of the output. As such scoping to model visualizations is justified, especially because visualization of model interna is uniquely possible in the domain of AI, and perhaps less so in other domains of explanation. The other types of visualization mentioned above are also overwhelmingly covered by the remaining groups.

(v) Counterfactual Generation

Techniques that generate counterfactuals are very homogeneous in approach and output. The literature review covers in detail that counterfactual generation is one of the better-defined fields of XAI, with quantifiable criteria for what constitutes a good counterfactual (Moraffah, 2020: 6). The output is generally one or multiple synthetic datapoints that vary minimally from a sample datapoint, but generate a different prediction. The way this grouping satisfies criteria (1) and (2) is clear.

How counterfactual generation separates itself from the other types of example-based explanation introduced in the literature review is a lot less intuitive. Each of them, including influential instances, prototypes and criticisms, and adversarial examples, either chooses or generates datapoints that in one way or another explain some facet of the data or the model. I argue it is correct to distinguish between these methods because their explanatory approach is sufficiently different. Prototypes, criticisms, and influential instances are all approaches that select existing datapoints, or sets thereof, to explain or represent the data space or model decisions. Counterfactuals on the other hand are synthetic datapoints which are not represented in the training data. This creates several divisions when considering them against the model of explanation. As such criterion (3) is fulfilled.

(vii) Neural Network Attention and Saliency Methods

The final group of methods are attention and saliency maps in neural networks. This is the only family among those identified that is not present in the literature in its exact form. I argue that the grouping of these two approaches is very intuitive: both apply only to neural networks; both refer to model interna, particularly neuron activations in the later layers; both read these neuron activations using gradient-based methods; both address only one instance at a time; and most importantly, the output of both is a highlighting of the parts of the input that were most relevant for a given prediction.

While they both fit under the umbrella of neural network specific methods, they differ from other NN XAI methods listed in the literature review considerably in approach and type of output. Neither visualization tools like TensorBoard (*TensorFlow Developers,* 2021**)**, nor other methods like the neuron-grouping visualizations cnn-inte (Liu et al, 2018) produces, are comparable in how they generate or present explanations. As such, criterion (3) is satisfied.

(vii) Justifying the Selection

The groupings of models presented are neither mutually exclusive nor are they collectively exhaustive, which briefly warrants justification. The first criterion, mutual exclusivity, is virtually impossible to fulfill, as much of the literature review demonstrates. Methods such as surrogate fitting and outputs such as feature importances simply appear too broadly across the field, and the canonical divisions figure 1 shows create groups that are far too large and diverse to analyze collectively. Either way, overlap between the groups does not hurt the aims of this paper; if anything, it allows models that appear in two groups to be analyzed from multiple angles.

The non-exhaustivity of the groups is owed to the breadth of the field. While I posit that the vast majority of methods in XAI can be found in one of the groups presented, a handful of splintered methods definitely exist that do not fit into one of these groups. Probably the largest class of these are model-specific methods that exploit some specificity of the model they are analyzing to produce, and produce neither feature importances nor decision rules as their explanatory outputs, such as Concept Activation Vectors (Kim et al., 2017). Another class that is not considered are prototype selection algorithms, the counterpart to counterfactual example-based explanations. The justification for their omission is simply the scoping of the paper. Nothing about these methods fundamentally prevents them from being compared against the model presented.

5.2 Evaluating Model Families

We now have the components required to perform the core aim of this paper: a rigorously deduced model of explanation with concrete requirements of XAI, and a broadly representative set of method groups to test against it. This section merges these two components by analyzing each family against the model. No formal structure is followed in this analysis, for two reasons: firstly, as the analytical part of this work was front-loaded into the literature review and condensational processes resulting in the model and the families, most of the interactions between families and criteria are inane and resolved intuitively. Secondly, many of the justifications for whether or not a family fulfills a criterion are repeated or similar across families, and their duplication is avoided. Where a criterion's fulfillment is not intuitive, this is addressed in the appropriate detail.

(i) Transparent Models

R1 Why-Questions: Transparent models inarguably answer 'why' a prediction was made as well as any method can. Two things become very clear from the example of transparent models: grouping model and data into the explanans is sensible, and separating local and global 'why'-questions is sensible. Local why-questions find satisfying answers in transparent models, as a singular predictive output can be expressed precisely in terms of the interna of the model, be that the interaction of feature values and their coefficients, the path along a decision tree or set of trees that the prediction takes, and so on (Guidotti et al., 2018: 8).

Global why-questions pose more of a problem; rather, they open the pandora's box of why-questions about model training that is discussed further in section 6. One can provide global why-answers about the behavior of the model only in terms of specifications about the model's interna, which generates somewhat empty explanations (for example, 'why does the model classify so many claims as fraudulent' is poorly answered by 'because the threshold value for fraudulence is low'). What is implicitly asked is how the model interna came to be in the state they are now in, and the answer to this lies in the training process, and with this in the training data. Simply having access to the model is thus not enough to answer these types of questions.

Luckily for transparent models, R1 only requires one type of why-question be answered, and local ones are.

R2 Contextuality: I argue transparent models are not inherently contextual, as simply having access to the entire model does the opposite of highlighting which parts of a model were particularly relevant to a given prediction. Yet one will struggle to find a transparent model which does not provide or at least allow some sort of feature importance quantification; for example, linear and logistic regressions have feature coefficients and decision trees allow an intuitive derivation thereof in the form of Gini importance (Menzel et al., 2009). As such, transparent models are contextual with respect to D but not with respect to M. As both D and M are in the explanans, this criterion is satisfied.

R3 Causality: Transparent models do satisfy causality in local cases in the sense in which it was derived in section 4.1.4: they allow the exact reconstruction of the causal chain between input data and elements of the model that led to a certain prediction. They do not allow causal interpretation into the dataset in the inputs-explain-outputs sense in which it is often attempted. As transparent models which provide feature importances are more interpretable than any of the following groups, this conclusion applies to all subsequent groups too. Additionally, they fall short of satisfyingly elaborating the causal chain that led to their parameters taking certain values, which is a criticism in the same vein as that of their 'answering' why-questions. Nonetheless R3 is satisfied.

R4 Factuality: Transparent models are inherently fully factual.

E1 Functional: Transparent models appear to be the ideal case for functional explanation. A question posed about model interna can be answered with full reference to the role of the part of the model that is asked about, the function the model was optimized for, and the general interaction of parts of the model. It is somewhat surprising, then, that functional explanations that follow this pattern are entirely empty. One cannot argue that the explanation of a part of a model – a feature weight, a neuron, a rule in a decision tree – is completed by saying that it was chosen to optimize the model's loss function or, in the case of decision trees for example, some secondary choice function. These sorts of explanations are profoundly empty and to some extent circular, in that they refer to the overall purpose of the model, which in turn is achieved by applying it to each part of the model, including the one being explained. This more or less closes the book on functional explanation as a concept in XAI in general, as again transparent models are the gold standard which post-hoc techniques generally try to replicate.

E2 Statistical: E2 is satisfied in that all transparent models rely on statistical laws, or at least loss or optimization functions like L1- or L2-norms or Gini coefficients.

E3 Understanding: Understanding is inherently satisfied by transparent models if the definition for transparent models Lipton (2016: 3) uses, that the "functionality can be comprehended in its entirety by a person", is applied. However, transparency is also taken to refer to the theoretical interpretability of a model's structure, which may not correspond to Lipton's definition, especially if the number of input features is large. In those cases where it does not, a loose application of the accounts of understanding that phrase it as an ability to 'work with' the understood results in a favorable evaluation of transparent models: seeing a transparent model, any reasonable explainee would be able to interpret how an unseen instance would be treated, for example. As such the criterion is satisfied.

(ii) Feature Importance Methods

R1 Why-Questions, R3 Causality: These two criteria are addressed in the same line of argument, which results in the conclusion that feature importances do not answer why-questions in a meaningful way. As a starting point, considering first that feature importance measures provide in a singular output a combination of interpretations of data and model, it is difficult to localize the part of the explanans that they target. They do not fully target the model because correlations in the data exist and will be exploited to some extent across models; they do not fully target the data because obviously models will differ in how they fit it.

That said, the literature review highlights in depth that why-questions are considered by a number of scholars to be strictly counterfactual questions, notably Woodward (2003). It was also made clear that counterfactuals are strongly correlated with causality. There is then some argument to be made that feature importances are counterfactual explanations, and therefore causal explanations, in that they illuminate what would need to change for the output to change. However, as Paez (2019: 15) puts it, these are not 'explanations' because "true counterfactual reasoning is purely theoretical, based on knowledge about how the model works”. Through this lens, feature importances merely convey *that* the prediction would change given a certain input change, not *why* it would. This, combined with the dubious clarity of the explanans, leads to the conclusion that feature importances fail these criteria.

R2 Contextuality: Feature importances are the prototypical selective method, and I do argue that contextuality is fulfilled, however with three asterisks, one about background knowledge and one about the explanans that mirrors the above. Firstly, feature importances do provide, in the traditional input-explains-output sense, some quantification of which areas of the input the model focuses on. They obviously do this more strongly if the importances attributed in any given case are somewhat disparate, with a handful of features being highlighted above others. What they don’t do is incorporate any background knowledge of the explainee; for example, a loan application will likely weigh current income very highly, and seeing it be attributed a high feature importance is accordingly unsurprising.

The second asterisk is that again, feature importances focus on some combination of data and model, and do not necessarily work to highlight for example what elements of the model were particularly relevant to a prediction (decision rules in trees, or sets of neurons in NNs, for example). Nonetheless, they do fulfill contextuality as posed in R2.

The final asterisk is that many, but not all, feature importance quantifications are derived from surrogate models, the disqualification of which is expanded upon below.

R4 Factuality: The factuality of feature importances, especially of post-hoc feature importances, stands and falls with the factuality of surrogate models, which is addressed below. In general, feature importances are factual if they are highlighted as the product of a surrogate approach, or directly calculated from a transparent model.

E1 Functional: Feature importances, as described above, do not focus on certain parts of the model and thus do not provide functional explanations.

E2 Statistical: All feature importance methods incorporate statistical laws, most obviously in the case of surrogate model fitting, as the surrogate models incorporate statistical laws themselves. ‘Direct’ feature importance approaches like SHAP, too, depend on mathematical concepts like coalitional game theory.

E3 Understanding: Understanding is a strong argument for feature importance methods. Generally, people perceive that they understand predictions better if they see feature importances for them; this knowledge is also to some extent transferable to unseen predictions (de Graaf and Malle, 2017: 3). Again, though, there remains the issue of whether understanding values that were generated from a surrogate model or a post-hoc technique counts as understanding the model itself, which is addressed below.

(iii) Surrogate Methods and Rule Extraction

R1-R4: Every core criterion of the model, with respect to surrogate methods, requires an answer to the question of whether insights based on an analogical model count as insights into the model itself. Given the prominence of surrogate models in XAI, this has rippling implications for the evaluation of much of the field. It is particularly damning, then, that the answer is quite decidedly no, despite pragmatic objections.

Understanding analogical models simply is not understanding the model; as such explaining analogical models is not explaining the model, and this is something all accounts of explanation but pragmaticism agree on. The weight of this claim justifies a brief deviation from the funneling methodology of this work to consider whether, or how, a few of the root models treat analogy. DN explanation (Hempel and Oppenheim, 1948) obviously rejects it, as it requires the explanans to be generally true laws or circumstantially true facts, which an analogical model is not. IS explanation (Hempel, 1965) does leave some room, in that the surrogate model does invoke statistical laws in its training process, but IS too rejects surrogates because it requires strictly inductive reasoning from the knowns, and the predictions and interna of a surrogate model are not strictly induced by the root model. None of the causal or counterfactual models accept it for the same reason; the relationship between input and output of the surrogate model, which is what is presented in the explanatory method’s output, is merely correlational as in every other supervised model. Given this consensus, any insights generated from a surrogate model, transparent as they may be, cannot be taken to be insights into the underlying model.

Having established this, the condemnation of surrogate models in the context of the requirements is easy: why-questions about the original model are not answered by surrogates, contextuality is not produced, causality is not considered, and factuality is not satisfied either. However, it is ironically the requirement of factuality that escapes this condemnation, to the extent that an explanatory method's output, if it highlights that it is merely the result of a surrogate model, does not contain anything untrue. This does not redeem analogical models, however – what does redeem them somewhat is understanding-based pragmaticism.

E1 Functional: Surrogate models do not make any attempt to even give insight into the original model, so there is no attempt at functional explanation either.

E2 Statistical: As highlighted above, surrogate models do make use of statistical laws when they are being trained, so this requirement is satisfied. In the light of the above, this raises some questions about the efficacy of this requirement in identifying explanations.

E3 Understanding: Understanding is what redeems surrogate-based approaches as XAI methods despite their failure according to the model. Paez' (2019) argument that analogy is 'all we have', in that black-box models are simply too complex to comprehend in themselves, has been stated at various points in this work. It is backed by the justified presence of analogical reasoning, though not analogical explanation, throughout scientific and human history (see Bartha, 2019). It is further backed by the gateway to experimental studies of understanding which pragmatism opens. These sorts of studies are already providing some evidence that techniques based on surrogate models can generate a deeper, transferable understanding of the model (eg. van der Waa et al., 2020). As such, the rejection of surrogate models by this model of explanation is all but a death knell for them, and in fact poses some questions in the other direction, which are addressed in section 6.

(iv) Model Visualization

R1 Why-Questions: Visualizations indubitably make an attempt at answering why-questions, but, where they are fully accurate to the underlying model, they fall short in the same way transparent models fall short of being fully explainable, in that at best they reduce the question to being about the training process. Furthermore, a key feature of many model visualization techniques, notably visualizations of neural networks, is a vast simplification of the interna of the model. Unfortunately, this makes the broad objection to analogical methods summarized in the above section relevant to visualization techniques as well. As such it cannot be said that they fulfill the requirement. The subdivision into accurate and abstracted visualizations, though, greatly simplifies the evaluation of the remaining criteria too.

R2 Contextuality: For completely accurate visualizations, visualizations of only the model, without visualization of, for example, a datapoint being processed by the model, are not contextual and not selective, mirroring fully transparent models. Abstractive visualizations fail this criterion too, as surrogate models do.

R3 Causality: I argue that accurate visualizations do illuminate causal relationships the same way that transparent models do, leveraging the fact that data and model are both in the explanans. In this case, the objection to surrogate models cannot be translated easily onto abstractive visualizations, as they do not construct a model based on input-output pairings, but do represent a strictly simplified version of the underlying model, or even the model's structure without its exact values. Nonetheless, without the ability to fully trace a datapoint's path through the model, there is no argument to be made that abstractive visualizations provide causal information.

R4 Factuality: I argue that both abstractive and full visualizations are factual. Abstractive visualizations may be missing some details about the model, but what they do represent is not untrue.

E1 Functional: For non-abstractive visualizations, the same line of reasoning applies as to transparent models; the visual component does not alleviate the emptiness of explanations generated by a functional approach. Abstractive visualizations do not generally permit functional explanation.

E2 Statistical: Nothing about visualizations, abstractive or not, is distinctly statistical in any way beyond what the underlying model is. As such, while dependencies on statistical laws in the model may be represented in the visualization, the explanatory method does not introduce new ones or require them to produce its output. This criterion is not satisfied.

E3 Understanding: Visualizations are, as the literature review highlights, hugely helpful in generating understanding. This understanding is fungible and translates onto unseen cases (de Graaf and Malle, 2017: 3). It also intuitively supports the ability to 'work with' the model that the accounts of understanding highlighted in the literature review require.

(v) Counterfactual Generation

R1 Why-Questions: Uniquely among the reviewed model-agnostic post-hoc techniques, counterfactual explanations do answer why-questions. This is owed to their rephrasing as what-if-things-were-different questions in Woodward (2003: 3) and the body of work evolving the concept of counterfactuals since then. Counterfactuals, as explained in the literature review, enjoy a close relationship between the theory of explanation and the requirements set of them in the practice of XAI, down to justifiable and canonical quantifications of their requirements, such as sparsity of perturbation and proximity to the original instance (Moraffah et al., 2020: 10). They are thus a class of method that very closely approximates what theories of explanation require of them. As such, evaluation of counterfactuals in XAI comes very close to evaluation of counterfactuals as a theory of explanation – which has been positive to the extent that they have been the basis of monist approaches (Reutlinger, 2016).

R2 Contextuality: Counterfactuals are contextual in the sense that a counterfactual instance presupposes the explainee's understanding of the original instance that is being compared to. Moreover, counterfactual instances by nature highlight the elements of the input that have the most leverage over the output, satisfying selectiveness.

R3 Causality: Thanks to their close approximation of counterfactuals in the theory of explanation, counterfactuals in XAI do provide causal interpretation.

R4 Factuality: Whether or not counterfactuals are factual is not intuitively obvious, because the actual counterfactual instance is fictional. However, it is also clearly highlighted, even central to the theory, that the counterfactual instance is not real; the key element of the output is the perturbation that has been performed to change the instance. As such, counterfactuals are factual (contradictory as that sounds). The above argument about the identification of analogy guaranteeing factuality also applies here.

E1 Functional: Counterfactuals are local methods, so they are not functional.

E2 Statistical: Counterfactuals and measurement of perturbations, as highlighted above, rely on fundamental statistical concepts like distance norms, so E2 is satisfied.

E3 Understanding: Whether or not counterfactuals generate understanding of the model is less intuitively answered. While they are doubtlessly the most effective method at developing understanding around a local prediction, whether they generate an overall understanding of the model is far less clear, and not studied experimentally. The former is sufficient, however, for saying counterfactuals satisfy this requirement.

(vi) Neural Network Attention and Saliency Methods

R1 Why-Questions: Because of the similarity of their insight to feature importances, methods in this category can be seen as analogical to them in terms of answering why-questions. It was argued there that feature importances simply show *that* certain elements of the input were relevant, but not why. Despite a deeper connection to the underlying model, methods in this category provide functionally the same output, a highlighting of parts of the input space, and therefore cannot be said to answer why-questions any better.

R2 Contextuality: This family of methods absolutely satisfies contextuality. These methods highlight exactly the parts of the input that are relevant to a given prediction, and do so through a deep understanding of the model, eliminating the hesitations about this judgement for feature importance methods. They also presuppose the explainee's understanding of the nature of the input.

R3 Causality: Causality is satisfied more closely than by model-agnostic feature importances, but not enough to claim the criterion is fulfilled. There is again a difference between *whether* a part of the input was relevant, and *how* it was relevant. There is no meaningful connection made in the method's output between the data and the interna of the model, meaning we cannot translate their output in this way (Das and Rad, 2020: 20)

R4 Factuality: Factuality is absolutely satisfied by these methods, as the weighting of the inputs' importance they undertake is extracted entirely from model interna by a transparent, if complex, process.

E1 Functional: As local methods, these approaches make no claim at functional explanation.

E2 Statistical: Gradient-based techniques, which dominate this group, rely deeply on statistical and mathematical concepts, especially those from stochastic calculus.

E3 Understanding: Whether or not understanding is generated again mirrors feature importances, as it is largely independent of the explaining method's internal process. As there are no hesitations about ‘fake’ understanding being generated by a surrogate model, these methods absolutely satisfy this criterion.

6 Discussion

6.1 Implications

This section addresses some of the implications of the analysis presented in this work. It is categorized into two sections: insights that models of explanation have for the further development of XAI, and insights about models of explanation that the ‘use case’ of XAI provides.

6.1.1 From Explanation to XAI

(i) An Evaluation Framework for XAI

Firstly, it is frequently lamented that XAI lacks rigorous standards against which to evaluate methods. This work may serve as a step in this direction. Of course, it does not answer the request many authors pose for an evaluation framework that is quantitative and comparable across models (Stepin et al., 2021: 11997). If anything, the level of nuance and customization for each family of models that was required may act as an argument that such a framework is unrealistic, as it ignores much of the variance within the field. Nonetheless, I argue the seven-requirement model presented in this work is well-suited to a systematic evaluation of any given XAI approach on its own, and as the analysis section showed, does allow some comparison across drastically different approaches.

(ii) Meta-Evaluation of XAI Evaluation Frameworks

Beyond providing its own framework against which to measure XAI, this work may also help in the meta-evaluation of other frameworks that aim to measure XAI as presented in sections 2.3.2 and 2.4. Specifically, it generates a surprisingly clear distinction between criteria that do or do not further the aim of providing explanation as defined by a purely philosophical model. Some criteria that do are Miller’s (2017: 2) selectiveness and socialness, which closely match the contextuality criterion the model poses; Miller’s (2017) claim that explanations are contrastive, which can be subsumed under the causality criterion; fidelity (Molnar, 2020: 2.4), the closeness of an XAI method to its underlying model, a lack of which often causes a method to fail the truthfulness criterion; and possibly even model performance (Fauvel et al., 2020: 6), which also affects an explanation’s truthfulness criterion. Sample criteria that fail to further this aim are decomposability of transparent models (Lipton, 2016: 9), as is expanded upon below, or the portability of methods across different models (Molnar, 2020: 2.5), which was shown to be inversely correlated with explanatory strength by the poor performance of global methods in the evaluation.

A more detailed examination of the success and failure of extant criteria in generating explanations, as defined by this work’s model, may thus also further the aim in (i), in that it may inform the development of more sensible concrete frameworks for measuring the performance of XAI.

(iii) The Hurdle of the Training Process

A core observation that jeopardized the performance of many of the groups, even transparent models, against many of the criteria, especially the answering of why-questions, is that they attempt to explain only the model as-is, that is, the form of the model that is making predictions. However, a large contingent of reasonable and commonly asked why-questions about models, which includes but is in no way limited to why-questions about model interna, is unsatisfactorily answered without some reference to the process that resulted in the model taking the form it has at the point at which it is being explained. The training process appears to be a colossal blind spot in XAI literature. Of course, model-agnostic post-hoc methods cannot be expected to know anything about the training process, considering they do not know anything about the model itself either. But the development of model-specific methods that elucidate how the training process resulted in certain parameters taking certain values may prove very fruitful.

(iv) The Observable Value of Reconciliation

An observation with implications in both ‘directions’ is that the results of this paper back those voices which call for reconciliatory multi-model approaches, such as Hochstein (2017) or Rice and Rohwer (2020). No model family alone fulfilled all seven aims, and those that came close – transparent models and counterfactual explanation – were tendentially global and local, respectively. No family of models produced fully satisfactory explanations that were both globally and locally applicable. Hochstein’s (2017: 5) claim that no model of explanation can fulfill all explanatory goals is thus supported by this work’s analysis. Combining multiple explanatory methods may thus prove a fruitful line of enquiry.

(v) Global and Local Explanation

The categories of explanation that were most successful in the analysis tended to be the ones that addressed single prediction explanations, especially counterfactuals, but also local feature importances, especially neural network-specific ones such as saliency maps and attention. This is not a death knell for global explanation; especially transparent models, which are by nature global, performed well. However, for further research in XAI, it is worth considering that human reason being largely in terms of prototypes (Adadi and Barreta, 2018: 136) appears to be reflected in what methods performed well against the model of explanation.

(vi) The Use of Transparent Models

The literature review cites authors like Rudin (2019) that posit that explanation of black-box models is not sufficient for transparency, and that interpretable models should take their place wherever possible. This is confirmed to a surprising extent by the analysis. Transparent models adhere well to the model of explanation that was presented, better than any other type of model except perhaps counterfactuals; even then, transparent models are global where counterfactuals are local, and nothing prevents these two approaches from being used in conjunction. In high-stakes decision environments like those Rudin (2019: 1) describes, her demand for interpretable models is thus – at least at the current stage of research – justified.

6.1.2 From XAI to Explanation

Reasoning about models of explanation from the perspective of XAI has been an enticing prospect at various points in this work, but would have come at the price of methodological consistency; risking redundancy, it bears restating that the aim was to work from a model of explanation towards XAI without leaking influence in the opposite direction. As such, insights in this direction that the work provides are collected in this section.

(i) XAI as a Datapoint in the Monism Debate

Firstly, XAI provides some insights about the monism debate. The core insight is that the division in the types of questions that can reasonably be asked about XAI systems, this division throws a spanner into some attempts at unification, most notably the counterfactual account advanced by Reutlinger (2016). This division, into global and local why-questions, was expanded upon in section 4.1.2. Counterfactuals in XAI, at least at the current stage of research, are strictly local methods, pertaining to one datapoint at a time. There is an argument to be made that global why-questions are not counterfactual at the scale of the entire model, but either way, there are no counterfactual techniques that answer global questions. Notably, the sort of global questions one may ask about a model and expect a counterfactual answer to could reasonably be portrayed to require grounding answers – contextualization in things like the architecture the model is run on, the broader data environment, or the training process – which Roski (2021) claims act as counterarguments to Reutlinger (2016). As such, counterfactual monism may find principled opposition in the realm of XAI.

(ii) The Inapplicability of Functional Explanation

Secondly, across the categories of methods evaluated, the functional criterion was either not fulfilled, or it was fulfilled but the ‘explanation’ generated was nonetheless profoundly empty. This does not indict functional explanation at large, but it does imply that functional explanation applies poorly to supervised machine learning. This is somewhat surprising, as section 4.2.1 describes how supervised models appear a natural fit for functional explanation as clearly defined systems, with a clearly defined – even measurable – goal in the form of a loss function.

A hint at the reason may come from a division of supervised model interna into the ‘given’ and the ‘learned’. Where they are given, they are not relevant to questions aiming to understand the machine’s intelligence; where they are learned, they (at least their values, not strictly their presence) are explained by the data and the loss function, which is slightly more informative, but still does not depend meaningfully on any information about the functioning of the system. In short, it appears attempting to explain supervised learners through a strictly functional lens begs questions that are one layer of abstraction removed from explaining model decisions, namely questions about human (or AutoML) choices that were made when setting the model architecture. As such its relevance for models of good explanations of decisions may be low.

A final observation is that functional explanation’s key requirement, that what is being explained is an element of a system, was perhaps applied too hastily in this work. In applying theories of functional explanation to supervised learning, we implicitly ask and positively answer the question “are models systems?”. Humans tend not to perceive models as systems, though, but as largely monolithic agents (source). Beyond explaining the failure of functional explanation, this may also act as an argument for why measures of model transparency that incorporate the model’s decomposability (eg. Lipton, 2016: 9) are not particularly well-directed.

(iii) XAI as an Argument for Understanding-Based Pragmatism

Next, XAI makes a strong case for understanding-based pragmatism as proposed by Wilkenfeld (2014) and van Camp (2010). The core element of this case is the rescuing of surrogate-based methods, and to some extent of feature importance methods, from condemnation as non-explanations by the model. These two cases exposed that the model of explanation may be too rigorous in some of its demands, as it is completely at odds with insights that the social sciences provide, which generally do support these methods as generating some form of understanding. This is exactly the line of argument that Wilkenfeld and van Camp take in their general rejection of over-defining the requirements models of explanations set. In the case of XAI, they appear to be right.

6.2 Limitations & Extensions

This work has a number of methodological limitations, though most of them were scoping choices, and many are more opportunity than challenge in that they lend themselves nicely to work building on this model.

Firstly, the methodology was entirely focused on theoretical work and literature synthesis. This is reflected in the literature review constituting the lion’s share of this paper. There was no experimental confirmation of the conclusions that were come to, and the three-stage relation between model, explanation, and user was only examined in its first two stages. The scoping section of this paper justified this focus; one further meta-literal argument for this approach is that it fills a blind spot in the literature. No author that was found in the literature review approached the field of XAI strictly with the purpose of evaluating it against an independently formulated philosophical model of explanation.

Relatedly, the theoretical works surveyed all belonged to the philosophy of explanation, omitting other angles from which XAI is examined. This was again justified in the scoping, but implies some interesting areas of further research. As the above discussion highlighted, this work uncovered some conflicts between what traditional models of explanation and more experimental methods in the social and cognitive sciences find, especially with regards to explanation by analogy. It would thus be interesting and likely fruitful for further research to evaluate how adherence to this work's model of explanation relates to findings from more practical studies of explanation and understanding. Without wanting to presuppose conclusions, the findings of this work would imply that traditional models of explanations would not fare particularly well in such a comparison.

Next, what was compared against the model of explanation were families of methods, not individual methods. This was a pure scoping choice, but of course limited the nuance with which they could be evaluated. A natural extension of this work would be to analyze individual methods, or more constrained 'leaves' of the method taxonomy, against the model. The nuance of the evaluation of each family was also limited by the choice to remain scoped across the entire field of XAI, rather than focusing on one strand of method. However I briefly reiterate the justification of this approach, which was that the research aimed to characterize the *whole* of XAI through the lens of the philosophy of science, and it would have been a daunting and perhaps unjustifiable task to conclude about the entire field from analysis of one area, however detailed. As such, the broader scope allowed this work to come to stronger conclusions.

Finally, one limitation is that the strict directedness of the modelling process did not allow for direct evaluation of XAI methods against specific accounts of explanation. Having funneled the many competing models into canonical requirements, the model families were evaluated only against these requirements, not against the any of the models they originated from. From a scoping perspective, again, I argue this was highly justified. However, especially in light of the insights discussed in section 6.1.2, it may be fruitful to evaluate methods directly against some of the more 'interesting' models introduced, such as the understanding-based pragmatic approach of Wilkenfeld (2014), or the more recent iterations of counterfactual approaches such as Strevens' (2008) kairetic account.

7 Conclusion

In this work, a comprehensive review of XAI and a comprehensive review of philosophical models of explanation was conducted. The former yielded a taxonomy of XAI methods, within which six evaluable families of methods were identified. The latter yielded a model of explanation which, to my knowledge, is the first rigorously deducted model of explanation as applied to XAI. It has four core requirements of methods: answering why-questions, explaining contextually, creating causal links, and using accurate information. Additionally, it has three ‘optional’ requirements that extend the core requirements for specific accounts of explanation: goal-directedness, the implementation of statistical laws, and the generation of understanding. None of these seven requirements is particularly surprising, but – or rather because – each of them is derived from and justified by a rigorous synergy of models of scientific explanation.

The characterizing contributions of this work fall into two categories. The first are the insights that were generated by the process of deriving the model: firstly, a rigorous and flexibly applicable definition of the concepts of explanandum and explanans in XAI; secondly, a justified expansion of the scope of causality in XAI beyond data and prediction to the interaction of data and model; thirdly, the model itself, which was shown in the discussion to be a potential building block in the bridge between philosophy of explanation and XAI evaluation.

The second category of contribution are the insights generated by evaluation of XAI methods against the model: that local approaches fared better than global ones; that voices calling for the use of transparent models rather than explained black boxes are all but unfounded; that especially post-hoc explanations tend to fail to answer ‘why’-questions; that there is merit to theories suggesting the parallel use of explanatory methods and/or the reframing of explanation through understanding; and that explanatory monism is not thwarted by XAI.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. *ArXiv:1806.07538 [Cs, Stat]*. http://arxiv.org/abs/1806.07538

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable artificial intelligence (Xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *ArXiv:1910.10045 [Cs]*. http://arxiv.org/abs/1910.10045

Bartha, P. (2019). Analogy and analogical reasoning. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2019/entries/reasoning-analogy/

Bechtel, W., & Richardson, R. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research. Princeton University Press.*

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, *7*, 49–72. https://doi.org/10.1162/tacl_a_00254

Botterill, G. (2010). Two Kinds of Causal Explanation. *Theoria*, *76*(4), 287–313. https://doi.org/10.1111/j.1755-2567.2010.01079.x

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, 245–317. https://doi.org/10.1613/jair.1.12228

Chua, E. Y. S. (2017). *Staying relevant: A partial defense of deductive-nomological explanation*. https://www.academia.edu/30236833/Staying_Relevant_A_Partial_Defense_of_Deductive_Nomological_Explanation

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. The University of Chicago Press.

Cummins, R. (1975). Functional analysis. *Journal of Philosophy*, *72*(November), 741–764. https://doi.org/10.2307/2024640

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (Xai): A survey. *ArXiv:2006.11371 [Cs]*. http://arxiv.org/abs/2006.11371

de Graaf, M., & Malle, B. (2017). *How people explain action (and autonomous intelligent systems should too)*. AAAI Technical Report FS-17-01. https://www.aaai.org/ocs/index.php/FSS/FSS17/paper/view/16009

de Regt, H. (2011). Explanation. *The Continuum Companion to the Philosophy of Science*. https://www.academia.edu/3859317/Explanation

De Regt, H. W., & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, *144*(1), 137–170. https://doi.org/10.1007/s11229-005-5000-4

Díez, J., Khalifa, K., & Leuridan, B. (2013). General theories of explanation: Buyer beware. *Synthese*, *190*(3), 379–396. https://doi.org/10.1007/s11229-011-0020-8

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *ArXiv:1702.08608 [Cs, Stat]*. http://arxiv.org/abs/1702.08608

Egler, M. (2021). Why understanding-why is contrastive. *Synthese*. https://doi.org/10.1007/s11229-021-03059-x

Fauvel, K., Masson, V., & Fromont, É. (2020). A performance-explainability framework to benchmark machine learning methods: Application to multivariate time series classifiers. *ArXiv:2005.14501 [Cs, Stat]*. http://arxiv.org/abs/2005.14501

Fidel, G., Bitton, R., & Shabtai, A. (2019). When explainability meets adversarial learning: Detecting adversarial examples using shap signatures. *ArXiv:1909.03418 [Cs, Stat]*. http://arxiv.org/abs/1909.03418

Friedman, M. (1974). Explanation and scientific understanding. *Journal of Philosophy*, *71*(1), 5–19. https://doi.org/10.2307/2024924

Garikov, P. (2021). *Visualkeras*. https://github.com/paulgavrikov/visualkeras

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2019). Explaining explanations: An overview of interpretability of machine learning. *ArXiv:1806.00069 [Cs, Stat]*. http://arxiv.org/abs/1806.00069

Glennan, S. (2017). *The new mechanical philosophy* (First edition). Oxford University Press.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A survey of methods for explaining black box models. *ArXiv:1802.01933 [Cs]*. http://arxiv.org/abs/1802.01933

Gurumoorthy, K. S., Dhurandhar, A., Cecchi, G., & Aggarwal, C. (2019). Efficient data representation by selecting prototypes with importance weights. *ArXiv:1707.01212 [Cs, Stat]*. http://arxiv.org/abs/1707.01212

Hailesilassie, T. (2016). Rule extraction algorithm for deep neural networks: A review. *ArXiv:1610.05267 [Cs]*. http://arxiv.org/abs/1610.05267

Hempel, C. G. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: The Free Press.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*(2), 135–175. https://doi.org/10.1086/286983

Hochstein, E. (2017). Why one model is never enough: A defense of explanatory holism. *Biology & Philosophy*, *32*(6), 1105–1125. https://doi.org/10.1007/s10539-017-9595-x

Hohman, F., Park, H., Robinson, C., & Chau, D. H. (2019). Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *ArXiv:1904.02323 [Cs]*. http://arxiv.org/abs/1904.02323

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. *ArXiv:1902.10186 [Cs]*. http://arxiv.org/abs/1902.10186

Karim, A., Mishra, A., Newton, M. H., & Sattar, A. (2018). Machine learning interpretability: A science rather than a tool. *ArXiv:1807.06722 [Cs, Stat]*. http://arxiv.org/abs/1807.06722

Khalifa, K. (2013a). Is understanding explanatory or objectual? *Synthese*, *190*(6), 1153–1171. https://doi.org/10.1007/s11229-011-9886-8

Khalifa, K. (2013b). Is understanding explanatory or objectual? *Synthese*, *190*(6), 1153–1171. https://doi.org/10.1007/s11229-011-9886-8

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors(Tcav). *ArXiv:1711.11279 [Stat]*. http://arxiv.org/abs/1711.11279

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific Explanation* (pp. 410–505). Minneapolis: University of Minnesota Press.

Kitcher, P., & Salmon, W. (1987). Van fraassen on explanation. *The Journal of Philosophy*, *84*(6), 315–330. https://doi.org/10.2307/2026782

Koh, P. W., & Liang, P. (2020). Understanding black-box predictions via influence functions. *ArXiv:1703.04730 [Cs, Stat]*. http://arxiv.org/abs/1703.04730

Konig, R., Johansson, U., & Niklasson, L. (2008). G-rex: A versatile framework for evolutionary data mining. *2008 IEEE International Conference on Data Mining Workshops*, 971–974. https://doi.org/10.1109/ICDMW.2008.117

Krishnan, S., & Wu, E. (2017). Palm: Machine learning explanations for iterative debugging. *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, 1–6. https://doi.org/10.1145/3077257.3077271

Lipton, Z. C. (2017). The mythos of model interpretability. *ArXiv:1606.03490 [Cs, Stat]*. http://arxiv.org/abs/1606.03490

Liu, G., Zeng, H., & Gifford, D. K. (2019). Visualizing complex feature interactions and feature sharing in genomic deep neural networks. *BMC Bioinformatics*, *20*(1), 401. https://doi.org/10.1186/s12859-019-2957-4

Liu, X., Wang, X., & Matwin, S. (2018). Interpretable deep convolutional neural networks via meta-learning. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–9. https://doi.org/10.1109/IJCNN.2018.8489172

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *ArXiv:1705.07874 [Cs, Stat]*. http://arxiv.org/abs/1705.07874

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about Mechanisms. *Philosophy of Science*, *67*(1), 1–25. https://doi.org/10.1086/392759

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*(1), 213. https://doi.org/10.1186/1471-2105-10-213

Miller, T. (2017). Explanation in artificial intelligence: Insights from the social sciences. *ArXiv:1706.07269 [Cs]*. http://arxiv.org/abs/1706.07269

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: how i learnt to stop worrying and love the social and behavioural sciences. *ArXiv:1712.00547 [Cs]*. http://arxiv.org/abs/1712.00547

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in ai. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Molnar, C. (2020). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., & Bischl, B. (2021). General pitfalls of model-agnostic interpretation methods for machine learning models. *ArXiv:2007.04131 [Cs, Stat]*. http://arxiv.org/abs/2007.04131

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, *26*(1), 67–82. https://doi.org/10.1093/esr/jcp006

Moraffah, R., Karami, M., Guo, R., Raglin, A., & Liu, H. (2020). Causal interpretability for machine learning—Problems, methods and evaluation. *ArXiv:2003.03934 [Cs, Stat]*. http://arxiv.org/abs/2003.03934

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *ArXiv:1902.01876 [Cs]*. http://arxiv.org/abs/1902.01876

Nagel, E. (1961). *The structure of science: Problems in the logic of scientific explanation*. Harcourt, Brace & World.

Nickel, B. (2010). How general do theories of explanation need to be? *: *theories of explanation*. *Noûs*, *44*(2), 305–328. https://doi.org/10.1111/j.1468-0068.2010.00741.x

Northcott, R. (2008, March). *Weighted explanations in history* [Published Article or Volume]. Philosophy of the Social Sciences. http://philsci-archive.pitt.edu/15408/

Offert, F. (2017). *“I know it when I see it”. Visualization and Intuitive Interpretability*.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence(Xai). *Minds and Machines*, *29*(3), 441–459. https://doi.org/10.1007/s11023-019-09502-w

Potochnik, A. (2015). Causal patterns and adequate explanations. *Philosophical Studies*, *172*(5), 1163–1182. https://doi.org/10.1007/s11098-014-0342-8

Puiutta, E., & Veith, E. M. (2020). Explainable reinforcement learning: A survey. *ArXiv:2005.06247 [Cs, Stat]*. http://arxiv.org/abs/2005.06247

Rancourt, B. T. (2016). *Understanding and its role in inquiry* [University of Massachusetts Amherst]. https://doi.org/10.7275/8316594.0

Reutlinger, A. (2016). Is there a monist theory of causal and non-causal explanations? The counterfactual theory of scientific explanation. *Philosophy of Science*, *83*(5), 733–745. https://doi.org/10.1086/687859

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you? ": Explaining the predictions of any classifier. *ArXiv:1602.04938 [Cs, Stat]*. http://arxiv.org/abs/1602.04938

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *AAAI*.

Rice, C., & Rohwer, Y. (2020). How to reconcile a unified account of explanation with explanatory diversity. *Foundations of Science*. https://doi.org/10.1007/s10699-019-09647-y

Roski, S. (2021). Metaphysical explanations and the counterfactual theory of explanation. *Philosophical Studies*, *178*(6), 1971–1991. https://doi.org/10.1007/s11098-020-01518-8

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *ArXiv:1811.10154 [Cs, Stat]*. http://arxiv.org/abs/1811.10154

Rüping, S. (2006). *Learning interpretable models*. https://doi.org/10.17877/DE290R-8863

Salmon, W. C. (1989). 4 decades of scientific explanation. *Minnesota Studies in the Philosophy of Science*, *13*, 3–219.

Salmon, W. C., Jeffrey, R. C., & Greeno, J. G. (1971). *Statistical explanation & statistical relevance*. University of Pittsburgh Press.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2016). Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, *128*(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *ArXiv:1704.02685 [Cs]*. http://arxiv.org/abs/1704.02685

Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). SmoothGrad: Removing noise by adding noise. *ArXiv:1706.03825 [Cs, Stat]*. http://arxiv.org/abs/1706.03825

Sober, E. (2020). A theory of contrastive causal explanation and its implications concerning the explanatoriness of deterministic and probabilistic hypotheses. *European Journal for Philosophy of Science*, *10*(3), 34. https://doi.org/10.1007/s13194-020-00299-5

Sokol, K., & Flach, P. (2020a). Explainability fact sheets: A framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 56–67. https://doi.org/10.1145/3351095.3372870

Sokol, K., & Flach, P. (2020b). One explanation does not fit all: The promise of interactive explanations for machine learning transparency. *KI - Künstliche Intelligenz*, *34*(2), 235–250. https://doi.org/10.1007/s13218-020-00637-y

Stepin, I., Alonso, J. M., Catala, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, *9*, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515. https://doi.org/10.1016/j.shpsa.2012.12.005

Tan, S., Caruana, R., Hooker, G., & Lou, Y. (2018). Distill-and-compare: Auditing black-box models using transparent model distillation. *Proceedings of the 2018*

AAAI/ACM Conference on AI, Ethics, and Society, 303–310. https://doi.org/10.1145/3278721.3278725

TensorFlow Developers. (2021). *Tensorflow* (v2.4.3) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.4724125

Thiagarajan, J. J., Kailkhura, B., Sattigeri, P., & Ramamurthy, K. N. (2016). Treeview: Peeking into deep neural networks via feature-space partitioning. *ArXiv:1611.07429 [Cs, Stat]*. http://arxiv.org/abs/1611.07429

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Camp, W. (2014). Explaining understanding (Or understanding explanation). *European Journal for Philosophy of Science*, *4*(1), 95–114. https://doi.org/10.1007/s13194-013-0077-y

Van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404. https://doi.org/10.1016/j.artint.2020.103404

Van Fraassen, B. C. (1980). *The scientific image*. Clarendon Press ; Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *ArXiv:1706.03762 [Cs]*. http://arxiv.org/abs/1706.03762

Verreault-Julien, P. (2019). Understanding does not depend on (Causal) explanation. *European Journal for Philosophy of Science*, *9*(2), 18. https://doi.org/10.1007/s13194-018-0240-6

Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. *ArXiv:2006.00093 [Cs]*. http://arxiv.org/abs/2006.00093

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *ArXiv:1711.00399 [Cs]*. http://arxiv.org/abs/1711.00399

What should we expect of a theory of explanation? (1980). *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1980*(1), 319–328. https://doi.org/10.1086/psaprocbienmeetp.1980.1.192575

Wilkenfeld, D. A. (2014). Functional explaining: A new approach to the philosophy of explanation. *Synthese*, *191*(14), 3367–3391. https://doi.org/10.1007/s11229-014-0452-z

Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.

Woodward, J., & Ross, L. (2021). Scientific explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends® in Information Retrieval*, *14*(1), 1–101. https://doi.org/10.1561/1500000066