What Does Explainable AI Explain?

Timo Freiesleben

Graduate School of
Systemic Neurosciences

LMU Munich

Dissertation der
Graduate School of Systemic Neurosciences der
Ludwig–Maximilians–Universität München

Munich, 6th December 2022

Supervisor: Prof. Dr. Stephan Hartmann
Chair of Philosophy of Science
Faculty of Philosophy, Philosophy of Science and the Study of Religion
Ludwig–Maximilians–Universität München

Second Supervisor: Dr. Alvaro Tejero-Cantero
Group Leader
Cluster of Excellence – Machine Learning for Science
Eberhard Karls Universität Tübingen

Third Supervisor: Prof. Dr. Paul Taylor
Principal Investigator
Munich Center for Neurosciences
Ludwig–Maximilians–Universität München

First Reviewer: Prof. Dr. Stephan Hartmann
Second Reviewer: Prof. Dr. Stephan Sellmaier
Third Reviewer: Prof. Dr. Jan-Willem Romeijn

Date of Submission: 6th December 2022
Date of Defense: 30th May 2023

Summary

Machine Learning (ML) models are increasingly used in industry, as well as in scientific research and social contexts. Unfortunately, ML models provide only partial solutions to real-world problems, focusing on predictive performance in static environments. Problem aspects beyond prediction, such as robustness in employment, knowledge generation in science, or providing recourse recommendations to end-users, cannot be directly tackled with ML models.

Explainable Artificial Intelligence (XAI) aims to solve, or at least highlight, problem aspects beyond predictive performance through explanations. However, the field is still in its infancy, as fundamental questions such as "What are explanations?", "What constitutes a good explanation?", or "How relate explanation and understanding?" remain open. In this dissertation, I combine philosophical conceptual analysis and mathematical formalization to clarify a prerequisite of these difficult questions, namely *what XAI explains*: I point out that XAI explanations are either associative or causal and either aim to explain the ML model or the modeled phenomenon. The thesis is a collection of five individual research papers that all aim to clarify how different problems in XAI are related to these different "whats".

In Paper I, my co-authors and I illustrate how to construct XAI methods for inferring associational phenomenon relationships. Paper II directly relates to the first; we formally show how to quantify uncertainty of such scientific inferences for two XAI methods – partial dependence plots (PDP) and permutation feature importance (PFI). Paper III discusses the relationship between counterfactual explanations and adversarial examples; I argue that adversarial examples can be described as counterfactual explanations that alter the prediction but not the underlying target variable. In Paper IV, my co-authors and I argue that algorithmic recourse recommendations should help data-subjects improve their qualification rather than to game the predictor. In Paper V, we address general problems with model agnostic XAI methods and identify possible solutions.

Contents

Summary iii

1 General Introduction 1

2 Paper I: Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena 19

3 Paper II: Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process 55

4 Paper III: The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples 99

5 Paper IV: Improvement-focused Causal Recourse (ICR) 135

6 Paper V: General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models 161

7 Discussion 193

Bibliography 207

List of Abbreviations 217

List of Publications 225

Acknowledgements 227

Eidesstattliche Versicherung / Affidavit **229**

Author Contributions **231**

Chapter 1

General Introduction

"A problem well stated is a problem half solved."

— Charles Franklin Kettering

Machine learning (ML) has broadened the range of problems in science and elsewhere we can approach algorithmically: problems that scientists tried to solve for many years, such as protein structure prediction, have recently seen scientific breakthroughs using ML (Senior et al., 2020); sensitive tasks that once required human judgment, such as medical diagnosis, are increasingly supported and preprocessed by ML prediction models (Fatima et al., 2017). As with all technologies, their adoption bears opportunities and, at the same time, great risks. How do ML models arrive at their "decisions"? What kind of errors are such ML systems prone to? Does science remain on epistemically safe grounds when relying upon ML algorithms in research? How can patients react to unfavorable algorithmic decisions?

One may argue that nothing is really new about ML with respect to these problems: humans are also opaque in their decision making; all decision makers (including humans) are prone to errors; scientists already use computationally complex models in their research and adapt well to these developments; physicians already use algorithmic decision support systems and still provide patients with valuable action recommendations.

However, there is a fundamental difference between current ML models and other models – ML models are *opaque* (Sullivan, 2020; Boge, 2021; Creel, 2020). Even when all model information is available, it is very hard for humans to understand the reasoning behind the model's predictions; the "decision making" cannot be easily mapped to humanly accessible concepts. This opacity roots in the process by which ML models are generated. Instead of

providing explicit reasoning rules or encoding tons of background knowledge as in the days of good old fashioned artificial intelligence (GOFAI), ML models can flexibly adapt to a range of different phenomena and learn directly from examples as in (un-)supervised learning or from trial and error as in the case of reinforcement learning (Russell, 2010; Hastie et al., 2009; Smith, 2019).

The standard narrative in the young field explainable artificial intelligence (XAI) is therefore that ML opacity must be reduced (Zednik, 2021). Once ML models are made transparent, they again fall into the class of models we know how to deal with. XAI researchers have developed a number of methods (Molnar, 2020) that help to explain, interpret or simply better understand ML models (see Section 1.3 for an overview). It should be noted here that the field of XAI is often also referred to as *interpretable machine learning* (IML). In this dissertation, the names XAI and IML are used as synonyms, even though philosophers have highlighted conceptual differences between explanation and interpretation (Erasmus et al., 2021).

This doctoral thesis focuses on what is explained in XAI – is it the ML model or the underlying phenomenon? Is the explanation associative or causal? Many controversies in the XAI community, as well as the usefulness of XAI in scientific and social applications, depend on these questions. Two aspects characterize my work: I use conceptual analysis to clarify the problem under consideration; and, I employ mathematical tools to provide solutions to the identified problem.

The thesis is a collection of five individual research paper, all having to do with the different *"whats"* that XAI seeks to explain. However, before we come to the papers, we first provide the reader with some background information on the overall topic. Section 1 introduces the problems XAI aims to solve, the methods it uses, and the criticisms it currently faces. Section 2 introduces philosophical and scientific research on explanation in general, discusses the relationship between philosophical and scientific findings, and illustrates the role of the explanandum in explanations. Section 3 introduces four possible explananda that XAI may aim to explain and discusses how my research relates to them.

1 Explainable AI

Before we can talk about XAI, we need to understand the problem that XAI tries to solve. Therefore, we must conceptually understand what ML is about and how it works.

The ML apparatus is designed for problems that can be squeezed into a tight corset; it requires that the problem focuses on association and prediction rather than causality and control, that there are specified input-output spaces and corresponding data in those spaces, and that there is a clearly specified and one-dimensional performance metric that captures success. One class of problems for which ML has proven particularly successful, and which is also the focus of this dissertation, is supervised learning tasks.

1.1 The Supervised ML Paradigm

In *supervised learning tasks* we want to predict a target $\boldsymbol{Y}$ from variables $\boldsymbol{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. Supervised ML is therefore concerned with finding an ML model $\hat{m}$ that maps inputs from space $\mathcal{X}$ to an output space $\mathcal{Y}$ and has low (or ideally even minimal) expected prediction error (EPE). The EPE describes the expected error the ML model makes when predicting $\boldsymbol{Y}$ from $\boldsymbol{X}$, where the error is measured via a so-called loss function on space $\mathcal{Y}$.

Finding such a suitable ML model in practice requires the user to specify the *model class*, the *input* $\mathcal{X}$ and *output space* $\mathcal{Y}$, a *labeled dataset* $\mathcal{D} := ((x^{(1)}, y^{(1)}), \ldots, (x^{(k)}, y^{(k)}))$, and a *loss function*. With all these information specified, one can run (given hyperparameters) a so-called *learning algorithm* that searches for a *model* $\hat{m}$ in the model class that minimizes the *empirical risk* [1]. If the learning process is successful, the model $\hat{m}$ we obtain makes accurate prediction also for data it has not seen i.e. $\hat{m}$ generalizes well beyond the dataset $\mathcal{D}$.

The most successful classes of ML models in the last decade have been artificial neural networks (ANN) and random forests (RF). Both classes of models are highly *expressive* i.e. they can describe a large variety of functions and therefore capture even complex phenomena. The reason for their expressive power is the large number of free parameters and the architecture of these models, which is composed of stacked simple units (e.g. neurons in ANNs and decision trees in RFs). Both the architecture and the parameter values are selected to achieve *high-predictive performance*, rather than maintain

[1] The empirical risk is an estimation of the EPE with data.

interpretability: model architecture is often selected by trial-and-error and rule-of-thumb heuristics, rather than by theoretical guarantees or knowledge of the application domain; parameter values are even automatically determined through an iterative learning process that fits the parameters to the data. While this focus on predictive performance and automatic learning from data has the advantage that trained ML models show great capacities in practice, it makes them opaque to human understanding and interpretation. For this reason, ML models obtained from ML learning algorithms have often been called *black-boxes* (Rudin, 2019; Zednik, 2021; von Eschenbach, 2021).

1.2 XAI and Incomplete Problem Formulation

XAI is usually motivated on the grounds that trained models are black boxes and thus need to be analyzed to become more transparent/fairer/interpretable/etc. However, I prefer a more problem-centered framing to XAI, called *incomplete problem formulation* (Doshi-Velez & Kim, 2017):

1. We transform our actual problem into a prediction problem that falls within the ML paradigm.
2. Unfortunately, many aspects of the actual problem relevant to us are not captured by the prediction problem.
3. Because we like the solution of the prediction problem (i.e., the ML model) at least in the aspect it solved (i.e., the prediction), we do not dump the ML model, but try to solve post hoc the parts of the actual problem that the prediction problem did not capture.

In this picture, the XAI is the remedy for the initial misformulation or deformation of the problem. An example will illustrate the idea:

Example. Stroke [2] is a serious disease responsible for approximately 11% of deaths worldwide (Lallukka et al., 2017). Predicting stroke risk is therefore of critical interest to human society. Fortunately, we have more and more data available from medical records or smartphone health tracking, which makes

[2] In a stroke, blood flow in the brain is reduced or blocked (called ischemic), or even open bleeding occurs (called hemorrhagic), both of which lead to cell death. Scientists know of a number of risk factors for stroke (e.g., high blood pressure, high cholesterol, smoking, or obesity), early symptoms (e.g., facial drooping, arm weakness, or speech difficulties), and immediate symptoms (e.g., loss of consciousness, headache, or vomiting) (Boehme et al., 2017; Mackay et al., 2011).

stroke prediction an ideal application for ML. Different stakeholder in different contexts will have different goals with ML stroke prediction, four of them are:

- **Reliability.** A physician uses an ML model as a decision support to provide better stroke risk predictions to patients. While accuracy is important, she is most interested in what features the model focused on in its prediction when she disagrees with it.
- **Robustness.** A private health insurance company commissions a robust stroke prediction model from a technology company to optimize health insurance premiums. Robustness has importance beyond accuracy here to ensure that the system cannot be easily gamed.
- **Knowledge Generation.** A scientist may be interested to learn yet unknown valuable predictors (causes, symptoms, or risk factors) of stroke. A highly accurate prediction model addresses only secondary aspects but not the primary problem of learning predictive factors.
- **Recourse.** Patients may want to use online tools to check their stroke risk and lower it if it is high. While a highly accurate prediction model is desirable, patients would also like to receive recommendations for action to reduce their personal stroke risk.

If we had a human predictor instead of our ML model, we would ask her: in the reliability scenario, what features she paid attention to in making her decision; in the robustness scenario, how she would decide in a set of difficult cases; in the knowledge generation scenario, what features she considers predictive; in the recourse scenario, which actions she would recommend to the patient to reduce their stroke risk. In all scenarios, what is demanded could be called an explanation. However, they are different types of explanations, for different stakeholders, and with different purposes. XAI has been called to the rescue for all these different situations to provide a post hoc fix of an incompletely formulated problem.

1.3 XAI Methods

The core of current XAI research is defined by a common set of methods rather than a common problem. Therefore, to better understand current XAI, we provide a brief overview of the standard methods and taxonomies established within the discipline (Molnar, 2020).

Three taxonomies are widely accepted:

- **Model Interpretability:** the model structure and model elements of *inherently interpretable models* have an intuitive interpretation; the model structure and model elements of *opaque models* do not have an intuitive interpretation.
- **Method Scope:** *model-agnostic* XAI techniques are applicable to any functional mapping; *model-specific* XAI techniques are tailored to models with specific properties (e.g. continuity) or from a certain class (e.g. neural nets)
- **Explanation Scope:** *global* XAI techniques depict properties of the overall model; *local* XAI techniques explain individual predictions.

Model Interpretability. The following models are often referred to as inherently interpretable: (generalized) linear models, decision trees, rule based models or Bayesian networks; one can assign meaning to the individual model elements and also cognitively understand the "decision making process" of the model. However, the assumption that these models are necessarily inherently interpretable has also been challenged (Lipton, 2018; Breiman et al., 2001). It has been argued that for complex processes, models either remain interpretable but not adequately capture the process, i.e. they trade off interpretability against accuracy, making them less useful, or they better capture the complex process by incorporating many features, including interaction effects, but do so at the expense of interpretability (think of linear models with thousand of features and higher order interaction terms).[3]

Models that are generally considered to fall into the class of opaque models include deep neural networks (DNNs) in all forms, ensemble-based methods such as random forests (RFs), or support vector machines (SVMs). However, there are attempts to configure the training process for opaque models to render them inherently interpretable: DNNs can be divided into two tasks, concept learning and classification (Koh et al., 2020); tree ensembles can be constructed to allow decomposition into effects of different orders (Hiabu et al., 2020); SVMs can incorporate additional interpretability desiderata such as sparsity (Martin-Barragan et al., 2014).

[3]The existence of such a trade-off between accuracy and interpretability has been questioned by Rudin (2019), who argues that there are always inherently interpretable models that are as accurate as black-box models.

Method and Explanation Scope. Well-known global model-agnostic XAI techniques include: partial dependence plot (PDP), which describes the expected effect of a feature change on prediction (Friedman et al., 1991); permutation feature importance (PFI), which describes the decrease in model performance when the information of a specific feature is removed (Breiman, 2001; Fisher et al., 2019); and shapley additive global importance (SAGE), which describes the average conditional PFI of a feature over all possible trajectories of how that feature can be added (Covert et al., 2020).

Well-known local model-agnostic XAI techniques include: local interpretable model-agnostic explanations (LIME), which approximates the ML model locally by a simpler model (e.g. linear model or decision tree) (Ribeiro et al., 2016); Shapley Values that describe, for a given instance, the fair contribution of individual features to the predicted value (Štrumbelj & Kononenko, 2014); and counterfactual explanations (CEs) (Wachter et al., 2017) that explain individual predictions by alternative but maximally similar scenarios.

Well-known global model-specific XAI techniques include: activation maximization, which analyzes individual neurons in an Artificial Neural Network (ANN) and searches for inputs that maximally trigger that neuron (Mahendran & Vedaldi, 2016; Olah et al., 2017); network dissection, which describes a method for testing which elements in an ANN are associated with prespecified humanly intelligible concepts (Bau et al., 2017); and testing with concept activation vectors (TCAV), which measures how much a given concept (defined via a concept dataset) affects predictions (Kim et al., 2018).

Well-known local model-specific XAI techniques include: gradient-based feature attribution techniques, which describe the gradient of the prediction with respect to individual inputs (e.g. pixels) (Simonyan et al., 2013; Alqaraawi et al., 2020); path-attribution techniques, which compare a given input to a certain reference input and integrate over the path (Qi et al., 2019).

Counterfactual Explanations and the Partial Dependence Plot. We now depict two interpretation techniques in more detail, namely CE and PDP, since they are used in several of the papers. Both are conceptually simple and provide a good first intuition for (model agnostic) XAI methods. In both, we assume a trained ML model $\hat{m} : \mathcal{X} \to \mathcal{Y}$ that predicts whether someone will get a stroke $\boldsymbol{Y}$ from a set of features $\boldsymbol{X}$ including age, blood-pressure, body-mass-index (BMI), etc.

Counterfactual Explanation. Suppose a patient is predicted to have high stroke risk ($\hat{\boldsymbol{Y}} = 1$) and wants to know why she did not receive a low stroke

risk prediction ($\hat{\boldsymbol{Y}} = 0$). For example, a counterfactual explanation would be the following:

> If the patient had a BMI of 27 instead of 33, her stroke prediction would have been 0 instead of 1.

There could also be many alternative explanations of a similar form, representing alternative scenarios with one or more alternated features. Such alternative scenarios may provide understanding of predictions, action recommendations, or opportunities for contesting adverse algorithmic decisions (Wachter et al., 2017).

More formally, a counterfactual for a particular algorithmic decision $(x, \hat{m}(x)) \in \mathcal{X} \times \mathcal{Y}$ is given by an alternative input $x^c \in \mathcal{X}$ with minimal distance[4] d to x that leads to desired classification $y_{des} \in \mathcal{Y}$ rather than $\hat{m}(x)$ i.e.

$$x^c \in \underset{x' \in \mathcal{X}}{\arg\min}\ d_{\mathcal{X}}(x, x') + d_{\mathcal{Y}}(\hat{m}(x'), y_{des}).$$

Such x^c can for example be found using gradient based techniques (for differentiable models) (Wachter et al., 2017) or by genetic algorithms (Dandl et al., 2020). A counterfactual explanation is the depiction of scenario x^c in contrast to x.

Partial Dependence Plot. Suppose an ML engineer wants to know how the ML model uses age in its predictions. One can think of it as looking at how changing age to a certain value would change the predicted risk of stroke on average. This is exactly the idea behind PDPs (Friedman et al., 1991). The PDPs for age and BMI are presented in Figure 1.1.

Formally, the PDP of a specific feature X_p is defined as

$$PDP_p(z) := \mathbb{E}_{X_{-p}}[\hat{m}(X_{-p}, z)].$$

Since we generally do not have access to $\mathbb{P}(X_{-p})$, this term can be approximated and efficiently computed for finite data $\mathcal{D}$ via

$$\widehat{PDP}_p(z) := \frac{1}{k} \sum_{i=1}^{k} \hat{m}(x_{-p}^{(i)}, z).$$

[4] d is a distance function on space $\mathcal{X}$ that maps any two instances $x_1, x_2 \in \mathcal{X}$ to a real value $d(x_1, x_2) \geq 0$.

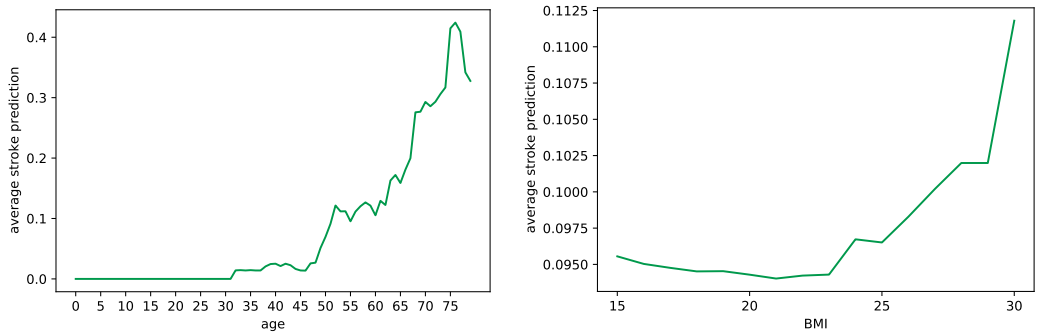


Figure 1.1: **Left plot:** PDP for the feature age. We see that the prediction model (RF model) strongly relies on age; the average predicted stroke increases for increasing age. **Right plot:** PDP for the feature BMI. the average predicted stroke increases with BMI. Due to the reliance on artificially generated datapoints, interpreting PDPs is generally difficult as we discuss in Paper I, Paper II, and Paper V.

1.4 Current Criticisms of XAI

XAI has often been attacked at its very foundations. Researchers have questioned whether XAI is, can be, or even should be a research field of interest. It has been argued that:

1. XAI conflates several goals (Lipton, 2018; Páez, 2019) or has no real goal (Krishnan, 2020).

2. XAI lacks proper evaluation metrics (Doshi-Velez & Kim, 2017; Mohseni et al., 2021) and should orient at social sciences with large scale human experiments (Miller, 2019; Liao & Varshney, 2021).

3. XAI explanations are misleading (Rudin, 2019; Mittelstadt et al., 2019), non-robust (Seuß, 2021), and do not incorporate uncertainties (Watson, 2022).

We will return to these criticisms and discuss them in the context of this dissertation in Chapter 7.

2 Explanation, Explanans, and Explanandum

In (probably too) general terms, the goal of XAI is the explanation of algorithmic decisions and the interpretation of algorithmic behavior. The conceptual nature of explanation and how humans explain in their daily life has been extensively researched by philosophers of science and social scientists respectively. The following brief overviews of this research follow the authoritative introductions by Woodward & Ross (2021) and Miller (2019). For a mini-glossary of relevant terms in the context of explanation see Figure 1.2.

Mini Glossary on Explanation Terms
Explanandum (pl. explananda): what is explained.
Explanans (pl. explanantia): how something is explained.
Explainer (pl. explainers): person who explains.
Explainee (pl. explainees): person who receives explanation.

Figure 1.2: Relevant terms in the context of explanation.

2.1 Explicating *Explanation* in Philosophy of Science

In philosophy of science, explanation has been largely described as a two-place relationship:

> Every explanation consists of an *explanandum* and an *explanans*.

Philosophers are concerned with discussing and defining how explanandum and explanans must be related in order to establish a proper explanation.

Hempels deductive-nomological (DN) account of explanation was one of the first and most famous competitors on the market (Hempel & Oppenheim, 1948; Hempel et al., 1965). Hempel proposal has a logical and a nomological[5] aspect: the logical aspect is that the explanandum must follow logically from the explanans and that the explanans must be true; the nomological aspect is that the explanans must contain physical laws and that these laws must be an

[5] lawlike

essential part for establishing the explanandum from the explanans. DN has been criticized, for example. for focusing on logical entailment, the unclear notion of a physical law, and the allowance for irrelevant explanantia (Salmon, 1979).

Salmon reacts to these criticisms and present the statistical relevance (SR) account that explicates explanation in terms of statistical associations (Salmon, 1971). Instead of logical entailment, SR focuses on whether the state of the explanans makes a difference for the explanandum statistically. SR has also been criticized, for example, for its focus on association instead of causation, treating low probability explanantia par with highly like ones, and the problem that the same explanans can explain logically exclusive exaplananda (Cartwright, 1979).

These criticisms led to various causal accounts of explanation that are now prominent (Woodward, 1989, 2005; Halpern & Pearl, 2020; Salmon, 1984). Instead of asking if observing different explanantia has impact on the explanandum, we ask if intervening on these explanantia would affect or would have affected the explanandum. However, also these accounts face problems such as the specifying a physical process, distinguishing between different explanations, and accounting for non-causal explanations (Hitchcock, 1995; Kitcher, 1989; Reutlinger & Saatsi, 2018).

There is also diverse research about how we can explain with scientific models (Bailer-Jones, 2003a,b; Bokulich, 2017; Jebeile & Kennedy, 2015; Strevens, 2011). In model-based accounts of explanation, the explanandum is represented in a model: explanantia are the factors that impact the explanandum within that model i.e. explanations are model-relative. Model-based accounts, however, are highly demanding because they require that it be specified in what sense the explanandum is represented (Frigg, 2002) and how complete and faithful the model must be in order to explain not only the model itself but really the modeled phenomenon (Giere, 2004).

Finally, pragmatic accounts of explanation, such as that of Van Fraassen et al. (1980); Achinstein (1983); De Regt & Dieks (2005), claim that attempts to explicate *explanation* as a two-place relationship failed because they did not take into account contextual information i.e. information about the explainer and the explainee. According to this school of thought, explanations describe answers to why questions that are inherently pragmatic entities and only interpretable relative to a given context. Pragmatic theories have been criticized, depending on how they fill the details, for providing a too liberal or deflationary notion of explanation (Kitcher, 1989; Kitcher & Salmon, 1987; Woodward & Ross, 2021).

2.2 Researching *Human Explanations* in the Social Sciences

Similar to pragmatic accounts in philosophy, social science research sees explanation as a three-place relationship that involves contextual information:

> An *explainer* explains a *phenomenon* to an *explainee*.

Instead of asking about the formal relationship between explanans and explanandum, the social sciences ask about the factors that make a good or successful explanation for the explainer and the explainee. Investigating human factors, such as human reasoning, cognitive biases, or explanation goals is therefore crucial in this research.

Malle's process model gives an example of an account focused on the explainer (Malle, 2011). His model introduces and relates two important distinctions: it separates the *explainer* from the *explanatory tools* she uses; and it distinguishes *information processes* from *impression management processes*. The explainer's information access refines her information processes and her goals guide her impression management processes. The information available to the explainer refines the explanatory tools she can use and the impression management process selects the right tools.

Since explanation is a cooperative act, Hilton (1990) proposed to apply Grice (1975) four maxims for cooperative conversation to explanation.[6] The four maxims applied to explanation are: *quality* i.e. explanations should be (approximately) correct; *quantity* i.e. explanations should provide just the right amount of detail and be informative to the explainee; *relation* i.e. explanations should be relevant in the conversation context; *manner* i.e. explanations should be provided in a respectful and helpful way. This normative account of explanation as conversation, including the four maxims was empirically supported (Slugoski et al., 1993; Tetlock & Boettger, 1989).

Psychological research has investigated how people actually explain in everyday interaction. They found that: people incorporate intentions, beliefs, and desires when explaining (Kashima et al., 1998); why questions are usually understood as contrastive (Lipton, 1990; Chin-Parker & Cantelon, 2017); people use cognitive biases when selecting explanations (Hesslow, 1988; Lombrozo, 2010); reference to causes is perceived more explanatory than reference to probabilistic associations (Hilton, 1996; Josephson & Josephson, 1996; McClure, 2002); reference to abnormal causes is perceived as more explanatory than to expected causes (Kahneman & Tversky, 1981).

[6] Antaki & Leudar (1992) extended the conversational model from simple dialogues to more complex arguments.

2.3 The Role of the Explanandum in Explanation Selection

It seems as if the social science perspective on explanation as a three-place relationship generalizes the philosophical two-place view: Philosophers wanted to determine what counts as an explanation, while social scientists ultimately want to figure out what good explanations are. However, we cannot have one without the other – establishing what counts as an explanation is a prerequisite for finding good explanations.

In my opinion, the philosophical and social science view must be seen as two complementary steps in a process that can be explained with Figure 1.3: the larger eclipse on the left can be seen as all possible explanantia that explain explanandum E, the definition of criteria for what belongs to this set is a philosophical question; the smaller green circle describes the good explanantia in a context C (i.e. for a specific explainer and explainee), finding this highly relevant subset is what social scientists are striving for.[7] From the philosophical point of view, we can consider the explainer and the explainee as fine-tuning parameters that help select good explanantia; what they do not affect is what counts as an explanation.[8]

Solving the three-place problem is both interesting and difficult; however, current XAI already fails at the preceding problem of specifying the explanandum properly.

The explanandum plays a major role in the search for good explanations, as also shown in Figure 1.3. The two black eclipses describe the explanantia for two different explananda, namely E and E'. It could be that the two eclipses overlap and describe explanantia that explain both E and E', but this need not be the case in general. Suppose we think we want to explain E, but if we were clear about the goal, we would see that we want to explain E': then choosing a good explanans for E (green circle) might not be an explanans at all for E', which is even worse than choosing a bad (but still a) explanans for E'.

[7] This does not mean that philosophers should not help establish criteria for good explanations in a specific context, or that social scientist's cannot help establish formal criteria; it just highlights the primary questions in the respective fields.

[8] Our view conflicts with pragmatic theories of explanation presented above that assume we cannot explicate explanation without accounting for context (Van Fraassen et al., 1980; Kitcher & Salmon, 1987).

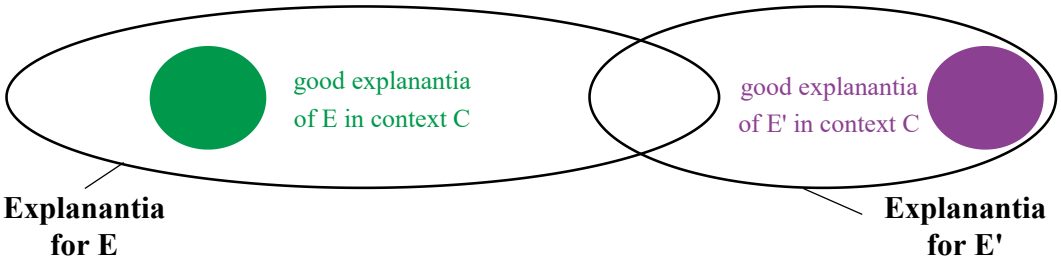


Figure 1.3: Philosophers are concerned with defining criteria for the membership of explanantia to one of the eclipses. Psychologists aim to find good explanantia that are helpful in a context (here C) that belong to the green or purple circles. Different explananda (here E and E') usually have partially different explanantia.

3 The Explanandum of XAI and My Research

What is the explanandum in XAI? There could be many, but the problem that supervised ML addresses already strongly narrows down the possibilities. If we abstract away from the specific application context of the ML model, we find four "whats" that XAI may be interested to explain. The four whats can be structured by two questions: Are we concerned with the phenomenon or the model? Do we want to gain causal or associative knowledge?

3.1 Four "Whats"

The four explananda can be found in Figure 1.4 and relate to the different philosophical accounts from Section 2.1, namely, the difference between associative and causal explanations and between explaining models and explaining phenomena with models. The first explanandum concerns how the ML model associates input features $\boldsymbol{X}$ with predictions $\hat{\boldsymbol{Y}}$ (associative model level). The second explanandum concerns how intervening on input feature $\boldsymbol{X}$ affect prediction $\hat{\boldsymbol{Y}}$ (causal model level). The third explanandum concerns how input variable $\boldsymbol{X}$ is associated with target variable $\boldsymbol{Y}$ (associative phenomenon level). The fourth explanandum concerns how intervening on input variable $\boldsymbol{X}$ affects target variable $\boldsymbol{Y}$ (causal phenomenon level).

It is important to keep these four explananda separate. While it is possible for an explanans to explain several different explananda, this is generally not the case. Different XAI methods and particularly their specifications (e.g. in their sampling (Janzing et al., 2020; Chen et al., 2020)) are suited for different

Level	Explanandum	Application
Associative Model	How is X associated with $\widehat{Y}$?	Reliability
Causal Model	How does intervening on X affect $\widehat{Y}$?	Robustness
Associative Phenomenon	How is X associated with Y?	Knowledge Generation
Causal Phenomenon	How does intervening on X affect Y?	Recourse

Figure 1.4: Four levels of what one may want to explain with XAI.

explananda. If this is not made explicit, XAI methods might be used in the wrong context. We may therefore end up in a situation like above, where our explanans is well suited to explain E in context C, but does not explain E', which we originally wanted to explain, at all. Separating these four explananda not only prevents misuse of XAI methods, but also legitimizes the usage of XAI methods across applications as long as the abstract explanandum is identical.

3.2 My Work and the Four "Whats"

All papers in this dissertation are about differentiating these different explananda. The papers are either concerned with showing the conflation of different explananda or with providing arguments for why particular goals require addressing a particular explanandum.

In total, the dissertation includes five research papers, which I wrote together with other researchers (Papers I,II,IV,V) and alone (Paper III):

I Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena

II Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process

III The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples

IV Improvement-focused Causal Recourse (ICR)

V General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models

Paper I. The theme of the different explananda is most evident in this paper, which concerns the question how to perform scientific inference using XAI. The paper discusses how we can use XAI to address questions that scientists are concerned with, particularly questions about the associative phenomenon level. We describe how this problem has been tackled with the traditional (statistical) approach by making assumptions about the phenomenon and analyzing individual model elements, but argue that ML requires a more holistic approach due to lacking representationality. We present a five step guide that shows how XAI methods must be constructed to make them useful for scientists who want to generate knowledge about the associative phenomenon level.

Paper II. This paper goes in a similar direction as paper I, but instead of a philosophical analysis of the problem of scientific inference, it shows how a theoretical ground truth of two XAI methods, namely the PDP and the PFI, can be defined. The ground truth is constituted by the XAI method applied to the underlying data generating mechanism. In the paper, we show how the error of our ML model explanation (PDP/PFI) compared to the "true" explanation (DGP-PDP/DGP-PFI) can be decomposed into three components: the bias and the variance induced by the learning process, and the error due to Monte Carlo integration. We provide a formal analysis of these errors and show how to define confidence intervals under the assumption of learner unbiasedness.

Paper III. I argue that counterfactual explanations and adversarial examples (instances that are misclassified by the ML model) are formally distinct, although they can be generated by solving the same optimization problem. The distinction again lies in the explanandum: while counterfactuals in their general form remain model explanations, adversarials point out model errors; to establish that a prediction is erroneous, it must be related to the associative or causal phenomenon level to compare it with an underlying ground-truth. As I show in the paper, an adversarial example can be seen as a counterfactual explanation that is misclassified i.e. in which $\hat{\boldsymbol{Y}}$ differs from $\boldsymbol{Y}$ even though we input the same value for x^c.

Paper IV. This paper has been inspired by Paper III. Rather than on general purpose counterfactual explanations it focuses on situations, where we want to provide data-subjects of ML models with action recommendation to revert unfavorable algorithmic decisions. For this situation, an explanandum at the causal model level has been proposed by Karimi et al. (2020) in which we search for the minimal interventions on $\boldsymbol{X}$ that would result in a change in prediction $\hat{\boldsymbol{Y}}$. We argue that this approach leads to the problem that data-subjects receive recommendations for how to game the predictor. Instead, we propose to focus on action recommendations on $\boldsymbol{X}$ that causally change $\boldsymbol{Y}$ i.e. explanations on the causal phenomenon level. We show that such recourse recommendations not only (most likely) change the prediction in the desired way, but also give robust and helpful action guidance to data-subjects.

Paper V. The paper discusses eight pitfalls of model-agnostic XAI techniques: 1. assuming one fits all interpretability, 2. bad model generalization, 3. unnecessary use of complex models, 4. ignoring feature dependence, 5. misleading interpretations due to feature interactions, 6. ignoring model and approximation uncertainty, 7. failure to scale to high-dimensional settings, and 8. unjustified causal interpretation. Pitfalls 1 and 8 directly concern the conflation or lacking clarity about the explanandum. Pitfalls 2 and 4 are closely related to the different levels of explananda, e.g. bad model generalization and feature dependencies are not a problem if our explanandum remains on the causal model level. For all pitfalls we provide a short description, a solution, and point out open problems.

Structure. The next five chapters consist of the five papers in the aforementioned order. Each of them has its own introduction, discussion, and bibliography. Chapter 7 consist of a short overall discussion; it discusses the extent to which the papers have helped solving the problem of the different explananda and whether they addressed the more general criticisms of XAI. The thesis ends with a personal perspective and outlook on the field XAI.

Chapter 2

Paper I: Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena

Freiesleben, T., König, G., Molnar, C. and Tejero-Cantero, A. (unpublished). Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena. *Under review at 'The British Journal for Philosophy of Science'*.

Author contributions:

T.F. wrote large parts of the paper and developed the initial idea. A.TC., G.K., and C.M. added valuable new ideas, proofread and helped revise the paper. A.TC. helped in the design of Figures 1,3,4, and 7. G.K. wrote large parts of the section on causal learning and A.TC. contributed a paragraph on mechanistic models.

Scientific Inference With Interpretable Machine Learning

Analyzing Models to Learn About Real-World Phenomena

Timo Freiesleben
Munich Center for Mathematical Philosophy
& Graduate School of Systemic Neurosciences
LMU Munich

Gunnar König
Department of Statistics
LMU Munich & University of Vienna

Christoph Molnar
Independent Researcher
Munich

Alvaro Tejero-Cantero
Cluster of Excellence Machine Learning
New Perspectives for Science
University of Tübingen

Abstract

Interpretable machine learning (IML) is concerned with the behavior and the properties of machine learning models. Scientists, however, are only interested in models as a gateway to understanding phenomena. Our work aligns these two perspectives and shows how to design *IML property descriptors*. These descriptors are IML methods that provide insight not just into the model, but also into the properties of the phenomenon the model is designed to represent. We argue that IML is necessary for scientific inference with machine learning (ML) models because their elements do not individually represent phenomenon properties; instead, the model in its entirety does. However, current IML research often conflates two goals of model analysis — model audit and scientific inference – making it unclear which model interpretations can be used to learn about phenomena. Building on statistical decision theory, we show that IML property descriptors applied on a model provide access to relevant aspects of the joint probability distribution of the data. We identify what questions such descriptors can address, provide a guide to building appropriate descriptors and quantify their epistemic uncertainty.

Keywords: Scientific Modeling, Interpretable Machine Learning, Scientific Representation, Inference, XAI

1 Introduction

Scientists increasingly use machine learning (ML) in their daily work. This development is not limited to natural sciences like the geosciences (Reichstein et al. 2019) or material science (Schmidt et al. 2019), but also extends to social sciences such as education science (Luan and Tsai 2021) and archaeology (Bickler 2021).

When building predictive models for problems with complex data structures, ML outcompetes classical statistical models in both performance and convenience. Impressive recent examples of successful prediction models in science include the automated particle tracking at CERN (Farrell et al. 2018), or DeepMind's AlphaFold, which has essentially solved the protein structure prediction challenge CASP (Senior et al. 2020). In such examples, some see a paradigm shift towards theory-free science that "lets the data speak" (Kitchin 2014, Anderson 2008, Mayer-Schönberger and Cukier 2013, Spinney 2022). Indeed, prediction is one of the core aims of science (Luk 2017, Douglas 2009), but so are, as philosophers of science and statisticians emphasize, explanation and knowledge generation (Salmon 1979, Longino 2018, Shmueli et al. 2010). Focusing exclusively on prediction may therefore represent a historical step back (Toulmin 1961, Pearl 2018).

What hinders scientists from using ML models to gain real-world insights is model complexity and an unclear connection between model and phenomenon — the so-called *opacity problem* (Boge 2022, Sullivan 2020). Interpretable machine learning (IML, also called XAI, for eXplainable artificial intelligence) aims to solve the opacity problem by analyzing individual model elements or inspecting specific model properties (Molnar 2020). Different stakeholders with different goals hold diverse expectations of IML (Zednik 2021), including scientists (Roscher et al. 2020), ML engineers (Bhatt et al. 2020), regulatory bodies (Wachter et al. 2017), and laypeople (Arrieta et al. 2020). Due to this plurality, IML has been criticized for lacking a proper definition (Lipton 2018).

Nevertheless, scientists increasingly use IML for inferring which features are predictive of e.g. crop yield (Shahhosseini et al. 2020, Zhang et al. 2019), personality traits (Stachl et al. 2020), or seasonal precipitation (Gibson et al. 2021). Although researchers are aware that their IML analyses remain just model descriptions, it is often implied that the explanations, associations, or effects found also extend to the corresponding real-world properties. Unfortunately, drawing inferences with IML can currently be epistemically problematic because the interpretation methods are not designed for that purpose (Molnar et al. 2022). In particular, the difference between model-only versus phenomenon explanations is often unclear (Chen et al. 2020, Hooker et al. 2021), and a theory to quantify the uncertainty of interpretations is lacking (Molnar et al. 2020a, Watson 2022).

Contributions. In this paper, we present an account of scientific inference with IML inspired by ideas from philosophy of science and statistical inference. We focus on supervised learning on identically and independently distributed (i.i.d.) data; we discuss other learning scenarios in Section 5.3. Our key contributions are: 1. We argue that ML cannot profit from the traditional approach to scientific inference via model elements because its parameters do not represent phenomenon properties (Section 3). While current IML methods aim to restore representationality of the model as a whole, they conflate the model audit and scientific inference goals of interpretation. 2. We identify the properties that

IML methods need to fulfill to provide access to aspects of the conditional probability distribution $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$, where $\boldsymbol{X}$ describes predictor variables and $\boldsymbol{Y}$ the target (Section 4). We call methods that are suitable for inference *IML property descriptors*. We provide a guide to build such descriptors starting with a phenomenon question about $\boldsymbol{X}$ and $\boldsymbol{Y}$ and evaluating whether it can be addressed, followed by an answer to this question with ML models and finite data, and conclude with the quantification of epistemic uncertainty. We illustrate our approach using conditional partial dependence plots (cPDP) as an example IML descriptor.

Terminology. For the purposes of our discussion below, a *phenomenon* is a real-world process whose aspects of interest can be described by random variables. Observations of the phenomenon are drawn from the unknown joint distribution induced by the random variables and form the dataset or just *data*. A *ML model* is a mathematical model optimized with the aid of a learning algorithm applied on the collected data in order to accurately predict unknown or withheld phenomenon observations, i.e. to generalize beyond the initial data. Here we focus on the supervised learning setting. Finally, *scientific inference* is the process of rationally deriving conclusions about a phenomenon from data (via ML, or other types of models). We employ *inference* to imply investigating unobserved variables and parameters similar to statistical inference, i.e. in a more general sense than is common in some of the ML literature, where it is used exclusively as a synonym for prediction. The knowledge gained by scientific inference can build the basis of *scientific explanations*. These brief conceptual remarks are meant to reduce ambiguity in our usage: we lay no claim as to their universality.

2 Related Work

Whether and how ML models, and specifically IML, can help obtain knowledge about the world is a debated topic among philosophers of science, statisticians, and also the IML community.

Philosophy of Science. It has been argued that ML models are only suitable for prediction because their parameters are instrumental and lack meaning (Bailer-Jones and Bailer-Jones 2002, Bokulich 2011). On the other hand, Sullivan (2020) argues that nothing prevents us from gaining real-world knowledge with ML models as long as the *link uncertainty* — the connection between the phenomenon and the model — can be assessed. Cichy and Kaiser (2019) and Zednik and Boelsen (2022) claim that IML can help in learning about the real world, but they remain vague about how model and phenomenon are connected. Like Watson (2022), we explain that IML methods relying on conditional sampling are faithful to the phenomenon. However, while he assigns IML inferences to the causal phenomenon level, we clarify that, without additional assumptions, such inferences only reveal associational relationships (Räz 2022). Our work makes precise that ML models can be described as epistemic representations of a certain phenomenon that allow us to perform valid inferences (Contessa 2007) via interpretations.

Statistical Modeling and Machine Learning. Breiman et al. (2001) describes ML (algorithmic modeling) and statistics (data modeling) as two approaches for reaching conclusions from data. On a medical example he shows that post-hoc analysis of ML models can allow more correct inferences about the underlying phenomenon than standard, inherently interpretable data models. Our paper gives an epistemic foundation for such post-hoc analyses. Shmueli et al. (2010) distinguishes statistics and ML by their goals — prediction (ML) and explanation (statistics). Like Hooker and Mentch (2021), we argue against such a clear distinction and offer steps to integrate the two fields.

This paper builds on ideas from Molnar et al. (2021), where they introduce ground-truth and confidence intervals for partial dependence plots (PDP) and permutation feature importance (PFI) of arbitrary ML models. Our work generalizes these ideas to arbitrary IML methods and draws the connection to the underlying phenomenon.

Interpretable Machine Learning. IML as a field has been widely criticized for being ill-defined, mixing different goals (e.g. transparency and causality), conflating several notions (e.g. simulatability and decomposability), and lacking a proper measure of success (Doshi-Velez and Kim 2017, Lipton 2018). Some even argued against the central IML leitmotif of analyzing trained ML models post hoc in order to explain them (Rudin 2019). In this paper, we show that, if we focus on interpretations for scientific inference, a clear foundation including a proper theory of success can be provided and these criticisms can be partially addressed.

3 Scientific Inference and Elementwise Representationality

The goal of this paper is to analyze and describe how we can conduct scientific inference on ML models using IML methods. This section explains why inference with ML models cannot be done as in traditional scientific models and why current IML methods do not generally address the problem. The next section describes our solution and illustrates it with a complete example from question formulation to uncertainty quantification.

3.1 ML Models are not Elementwise Representational

In scientific modeling, there is a paradigm that many models implicitly follow — we call it the paradigm of *elementwise representationality*.

Definition. *A model is* elementwise representational *(ER) if all model elements (variables, relations, and parameters) represent an element in the phenomenon (components, dependencies, properties).*

Figure 1 depicts the relationship between ER models and the phenomenon:[1] variables describe phenomenon components; mathematical relations between variables describe structural, causal or associational dependencies between components; parameters specify the mathematical relations and describe properties of the component dependencies. The upward arrows describe the *encoding* i.e. the translation of a phenomenon observation to a model configuration;

[1] See Appendix A for the philosophical origins of our perspective.

The downward arrows describe the *decoding* i.e. the translation of knowledge about the model into knowledge about the phenomenon. ER is obtained through model construction; ER models are usually "hand-crafted" based on back-

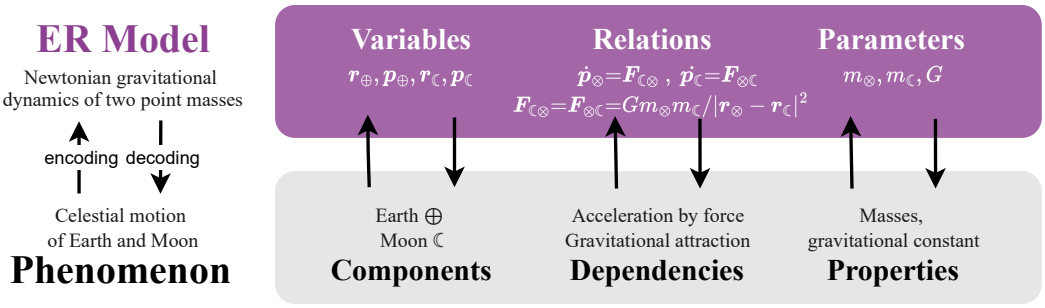


Figure 1: **Model and phenomenon sustain an encoding-decoding relationship**. The main elements of a traditional, ER model, are shown in encoding-decoding correspondence to the phenomenon elements they represent (Stachowiak 1973). Phenomenon and model elements are illustrated with a simple example of two bodies in gravitational interaction and its classical, Newtonian mechanistic description.

ground knowledge and an underlying scientific theory. Variables are selected carefully and sparsely during model construction, and the relations are constrained to a relation class with few free parameters. When ER models need to account for an additional phenomenon aspect, they are gradually extended so that large parts of the "old" model are preserved in the more expressive "new" model. ER even eases this model extension process because model interventions are intelligible on the level of model elements. Usually, ER is explicitly enforced in modeling: if there is a phenomenon element devoid of meaning, researchers either try to interpret it or exclude it from the model.

ER is so remarkable because it gives models capabilities that go beyond prediction. ER simplifies the step of decoding i.e. translating model knowledge into phenomenon knowledge. Scientists can analyze model elements and draw immediate conclusions about the represented phenomenon element (Frigg and Nguyen 2021). However, only those aspects of the phenomenon that have a model counterpart can be analyzed with this approach. Fortunately, as described above, ER models can be extended to account for further relevant aspects identified by the scientist.

Running example:[2] *Linear Model.* Suppose a researcher, we call her Laura, wants to study what attributes influence students' grades in mathematics. Specifically, she wants to research how language skills and math skills are associated. She uses a dataset from Cortez and Silva (2008), who collected data[3], encompassing 32 student attributes in Portuguese schools including math/Portuguese grades, age, parents' education, etc.

Laura starts with a classical ER model — a linear model with one predictor and one target variable. She selects the student grade in Portuguese $\boldsymbol{X}_p$ and in mathematics $\boldsymbol{Y}$ as her proxy variables for students language and math skills respectively.[4] Based on her background knowledge, she assumes that the true relationship can be described as $\boldsymbol{Y} = \beta_0 + \beta_1 \boldsymbol{X}_p + \epsilon$ with $\beta_0, \beta_1 \in \mathbb{R}$ and an error $\epsilon \sim \mathcal{N}(\mu, \sigma^2)$. Laura centers $\boldsymbol{X}_p$ by the average student grade in

[2]Since the physical model from Figure 1 is a mechanistic causal model (Schölkopf et al. 2021), we switch henceforth to an illustrative associational model from the social sciences that compares more fairly with current associational ML models. We strongly simplify things in this example and do not claim that it reflects social science methodology or that ML is even required.

[3]see Appendix B for more details on the dataset

[4]In the Portuguese grading scheme, the range is 0-20, where 0 is the worst and 20 the best grade.

Portuguese and obtains the prediction model

$$\hat{m}_{\text{LIN}}(x_p) = 10.46 + 0.77x_p$$

that minimizes the mean-squared-error (MSE). Laura's model is ER: she can interpret $\hat{\beta}_0 = 10.46$ as the predicted math grade for an average Portuguese student (if $x_p = 12.55$)[5] and $\hat{\beta}_1 = 0.77$ as the strength of association between the Portuguese grade and the math grade.

Laura can analyze the model to draw scientific inferences about the underlying phenomenon, for example, with 95% confidence intervals[6] for her estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ with $[10.05; 10.88]$ and $[0.63; 0.91]$ respectively. The inference she draws is on these parameters; Laura can only draw conclusions about the phenomenon given ER and high predictive model capacity. Laura may conclude from $\hat{\beta}_1$ that language skills and math skills are strongly and positively related. To reach a more expressive and predictively accurate model, Laura can also extend the model to include additional features, relations, or interaction terms. As long as she preserves ER, she can directly draw scientific inferences from analyzing model elements. Indeed, these inferences are only as valid as the modeling assumptions (e.g. target normality, homoscedasticity, or linearity).

ML Models are generally not ER. ER makes model elements interpretable and allows to reason about the effects of model or even real-world interventions; as such, ER models suit our image of science as an endeavor aimed at understanding. However, as mentioned above, we usually require background knowledge on which components are relevant, and we need to severely restrict the class of relations that can be considered for the given phenomenon. These difficulties might lead scientists to either limit their investigations to phenomena that are already well studied or, as Breiman et al. (2001) argued, to develop overly simple models for complex phenomena and possibly draw wrong conclusions.

ML models excel in modeling complex problems with an unbounded number of components that display ambiguous and entangled relationships i.e. ML models are highly expressive (Gühring et al. 2020). ML models are subject-domain independent (Bailer-Jones and Bailer-Jones 2002), this means that we do not necessarily need subject-domain background knowledge in modeling. Instead, ML modeling only requires specifying a broad model class and a set of hyperparameters. The choice of these hyperparameters is data-domain specific i.e. they reflect inductive biases that allow for efficient learning.

The gain in generality and convenience with ML comes at a price — ML models are generally not ER. As also argued in Boge (2022), Bokulich (2011), Bailer-Jones and Bailer-Jones (2002), ML models (e.g. artificial neural networks) contain model elements such as weights, activation functions, or network structure that have no corresponding phenomenon counterpart.

[5] Centering features is common in linear regression to make the intercept term interpretable.

[6] i.e., intervals $[a, b]$ such that, if the model assumptions hold, a 'true' parameter β is found inside 95% of all observational samples, $\mathbb{P}(a<\beta<b) = 0.95$.

Running example: Artificial Neural Network (ANN). Suppose Laura is dissatisfied with her linear model and fits a dense three-layer neural network to predict math grades using all available features.[7] She reduces the test MSE from 16.0 in the linear-one-variable model case to 8.9. A formal description of the model is given by:

$$\hat{m}_{\text{ANN}}(\boldsymbol{x}) = \sigma_3(W_3\sigma_2(W_2\sigma_1(W_1\boldsymbol{x} + b_1) + b_2) + b_3)$$

where model elements are the values of the weight matrices W_1, W_2, W_3 and bias vectors b_1, b_2, b_3, and the activation functions $\sigma_1, \sigma_2, \sigma_3$. Unlike in the linear model above, it is highly unclear what these parameters correspond to in our data or phenomenon. While the vector x is still representational, the weights, activation functions, or three-layer architecture are very hard or even impossible to interpret: A high value of weight $W_1^{(3,2)}$ might have a positive, neutral, or negative effect on the target, dependent on all other model elements; the activation function only reflects the currently popular heuristics in model training; and the particular three-layer architecture is a result of model selection based on predictive performance and rules of thumb, but with little phenomenon-based rationale.

3.2 Scientific Inference in Light of Current IML

We have argued so far that:

i) If models are ER, they allow for scientific inference.

ii) ML models are generally not ER.

How can we still do scientific inference with ML models? We discuss two strategies to enable scientific inference with ML: We argue that the first strategy, namely restoring ER, fails because ML models are designed to represent in a distributed manner; the second strategy, embracing holistic representationality, is highly promising but current attempts conflate different goals of model analysis. This discussion sets the basis for the next section, where we show how a holistic account of representationality can enable scientific inference.

Restore ER. One strategy towards scientific inference with ML is to challenge Proposition ii) and show that ML models are ER too. Researchers in this camp argue that individual elements in ML have a natural phenomenon counterpart, but this counterpart only becomes evident when these model elements are extensively scrutinized.[8] This would be surprising: ER is not enforced in state-of-the-art techniques and, even worse, some methods such as training with dropout purposefully discourage ER in order to gain robustness (Srivastava et al. 2014); ML models like ANNs are designed for *distributed representation* (Buckner and Garson 2019, McClelland et al. 1987).

It has been claimed that model elements represent high-level constructs constituted from low-level phenomenon

[7] We chose a neural net to make our argument. For training the neural network, Laura splits data into training and test, uses ReLu activation functions and minimizes the MSE loss via gradient descent with an adaptive learning rate.

[8] The underlying epistemological reasoning is that human representations are near-optimal and will be eventually rediscovered by ML algorithms.

components that are often called *concepts* (Buckner 2018, Olah et al. 2020).[9] If this is the case, model elements or aggregates of such elements can be reconnected to the phenomenon; ER would be restored by the representations of coarse-grained phenomenon components. Research on neural networks supports that some model elements are associated with concepts (Mu and Andreas 2020, Voss et al. 2021, Kim et al. 2018, Olah et al. 2017), however, often these elements are neither the only associated elements nor exclusively associated with one concept as shown in Figure 2 (Donnelly and Roegiest 2019, Bau et al. 2017, Olah et al. 2020). Problematically, intervening on these model elements generally does not have the expected effect on the prediction — the elements do not share the causal role of the "represented" concepts, even in prediction (Gale et al. 2020, Donnelly and Roegiest 2019). It is therefore questionable in what sense they still represent.[10] Moreover, this line of research predominantly focuses on images, where nested concepts are arguably easier to identify for humans.

Figure 2: **ML models are generally not ER**. Three input images that independently trigger a single model element (unit 55 in layer *mixed4e*, Olah et al. (2017; 2020)). A single unit in a neural net may respond to very different "concepts", e.g. heads of cats (left image), car bodies (center), or bees (right), suggesting that units generally do not represent disentangled concepts (Mu and Andreas 2020, Nguyen et al. 2016).

Research on the representational correlates of model elements seems indeed fascinating. However, current ML models that do not enforce ER will rely on distributed representations and cannot be reduced to logical concept machines. The associative connection between model elements and phenomenon concepts should not be confused with their equivalence. Analyzing single model elements will therefore be a hopeless enterprise.

Embrace Holistic Representationality. An alternative route to scientific inference is to accept that ER is well-suited for scientific inference and that ML models are not ER but reject that ER is the only approach for scientific inference. To choose this route, one must offer an alternative path for drawing scientific inference with ML models that goes beyond the analysis of model elements.

Our approach is to regard the model as representational of phenomenon aspects *only as a whole* — we call this *holistic representationality* (HR). HR implicitly underlies large parts of the current research program in IML: Model-agnostic methods, in particular, analyze the entire ML model simply as an input-output mapping (Scholbeck et al.

[9]The idea is that similar to the hierarchical structure of components in nature, where lower level components such as atoms combine to form higher level entities such as molecules, cells, and organisms; in deep nets, hierarchies evolve from pixels to shapes to objects.

[10]Though note generative adversarial networks as an exception; here, interventions on model elements have been linked to interventions on concepts in the generated images (Bau et al. 2018).

2019); In the same spirit, many model-specific IML methods like gradient or path-based feature attribution treat ML models as mappings with additional useful properties such as differentiability (Alqaraawi et al. 2020).

Model-agnostic and model-specific methods share the idea that relevant model properties such as the effects or importances of variables can be derived by analyzing the model just as a functional mapping. Initial definitions of, for example, global feature effects (Friedman et al. 1991) and feature importance (Breiman 2001) or local feature contribution (Štrumbelj and Kononenko 2014) and model behavior (Ribeiro et al. 2016) have been presented. However, many researchers have pointed out that these methods lead to counterintuitive results for dependent or interacting features and offered alternative definitions (Apley and Zhu 2020, Strobl et al. 2008, Molnar et al. 2020b, Goldstein et al. 2015, König et al. 2021b, Janzing et al. 2020, Slack et al. 2020, Alqaraawi et al. 2020).

We believe that these controversies stem from a lack of clarity about the goal of model analysis. Are we interested in model properties to learn about the model (model audit) or do we want to use these model properties as a gateway to learn about the underlying phenomenon (scientific inference)? These two goals must not be conflated.

The auditor examines model properties e.g. for debugging, to check if the model satisfies legal or ethical norms, or to improve her understanding of the model by intervening on it (Raji et al. 2020). Auditors even take interest in model properties that have no corresponding phenomenon counterpart such as single model elements or the model behavior for unrealistic feature combinations. The scientist who wants to draw inferences, on the other hand, wants to learn about model properties that can be interpreted in terms of the phenomenon.

Scientific inference and model audit should be viewed as two different but interacting goals. In each of them, we take different stances toward the ML model: The auditor adopts a skeptical attitude of the model, she has ground-truth information or normative standards to check the model against; the scientist adopts a trusting attitude, she wants to learn from the model. Both cases describe a knowledge asymmetry (Gobet 2018, Rosser et al. 2008) but in opposite directions. Auditing the model is an indispensable step for scientists to gain enough trust in it. Only after several rounds of auditing and improvement should the researcher rely on the model to draw scientific conclusions.

4 Scientific Inference with IML Property Descriptors

We just argued that ML models are generally not ER and therefore do not allow for scientific inference in the standard way. HR offers a viable alternative, but currently different goals of model analysis are conflated. In this section, we show that a HR perspective enables scientific inference using IML methods. Particularly, we show that certain IML methods — we call them *IML property descriptors* — can represent phenomenon properties. Figure 3 describes our conceptual move: instead of matching phenomenon properties with model parameters as in ER models, we match them with external descriptions of the whole model.

Idea. Instead of first thinking about the model and its properties (the model audit approach), we propose to start with the phenomenon and a scientific question about it. IML methods for inference should answer, or at least help answer, a scientific question concerning the phenomenon. The crucial step in our framework is to establish a link between the

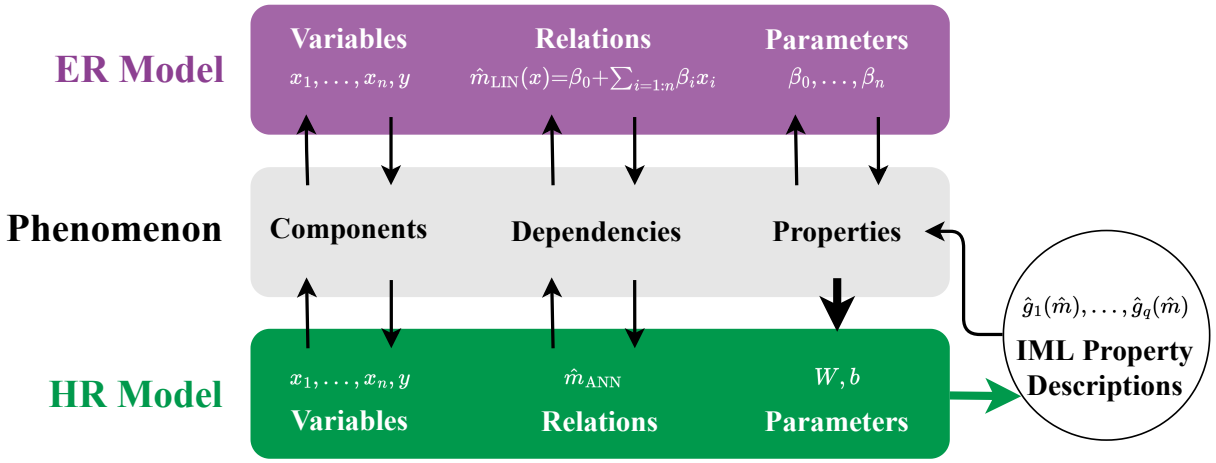


Figure 3: **IML property descriptions distill phenomenon properties from HR models**. Instead of explicitly encoding phenomenon properties as parameters like for ER models, HR models (e.g. ML models) encode phenomenon properties in the whole model. We propose that these encoded properties can be read out with IML property descriptions external to the model. In this way, IML offers an indirect route to scientific inferences through model analysis.

phenomenon and the model; we propose to draw this link with statistical decision theory, which shows what optimal ML models can holistically represent. Clearly, an approximate ML model will not provide an answer if even the optimal model cannot. However, if a question can in principle be answered with the optimal model, an ML model trained on data can approximate the answer in practice. The problem then becomes to quantify the approximation error.

4.1 ML Representationality and Optimal Predictors

Which aspects of a phenomenon ML models can represent even under ideal circumstances depends on the data, the learning paradigm, and the loss function. For identically and independently distributed (i.i.d.) data used for supervised learning, the theory of optimal predictors from statistical decision theory provides an answer (Hastie et al. 2009, p18-22). Besides the advanced theory available in this setting, supervised learning on i.i.d. data is the most popular ML setup in practical applications. We briefly discuss representationality and scientific inference in the case of unsupervised and causal learning in Section 5.3.

Basic Notation. We assume that the random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ and $\boldsymbol{Y}$ fully characterize the phenomenon. We write the joint feature vector as $\boldsymbol{X} := (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ with $\mathcal{X} := \text{Range}(\boldsymbol{X})$ and $\mathcal{Y} := \text{Range}(\boldsymbol{Y})$. $\boldsymbol{X}$ and $\boldsymbol{Y}$ jointly describe the phenomenon.

Optimal Predictors. An optimal predictor m can predict realizations of the target $\boldsymbol{Y}$ from realizations of $\boldsymbol{X}$ with minimal expected prediction error i.e. $m = \arg\min_{\hat{m} \in \mathcal{M}} \text{EPE}_{\boldsymbol{Y}|\boldsymbol{X}}(\hat{m})$, with $\text{EPE}_{\boldsymbol{Y}|\boldsymbol{X}}(\hat{m}) := \int_{\boldsymbol{Y}} L(\boldsymbol{Y}, \hat{m}(\boldsymbol{X}))\, \mathbb{P}_{\boldsymbol{Y}|\boldsymbol{X}}(y|x)\, \mathrm{d}y$, where

L describes a loss function $L(\boldsymbol{Y}, m(\boldsymbol{X})) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^+$ and $\hat{m}$ a model in the set $\mathcal{M}$ of mappings from $\mathcal{X}$ to $\mathcal{Y}$. Table 1 shows the optimal predictors for standard loss functions.

Problem	**Loss**	$L(\boldsymbol{Y}, \hat{m}(\boldsymbol{X}))$	**Optimal Predictor** m
Regression	Mean Squared Error	$(\boldsymbol{Y} - \hat{m}(\boldsymbol{X}))^2$	$\mathbb{E}_{\boldsymbol{Y}\mid\boldsymbol{X}}[\boldsymbol{Y} \mid \boldsymbol{X}]$
	Mean Absolute Error	$\lvert\boldsymbol{Y} - \hat{m}(\boldsymbol{X})\rvert$	Median$(\boldsymbol{Y} \mid \boldsymbol{X})$
Classification	0-1 Loss	0 if $\hat{m}(\boldsymbol{X}) = \boldsymbol{Y}$, else 1	$\underset{y\in\mathcal{Y}}{\arg\max}\ \mathbb{P}(\boldsymbol{Y}{=}y \mid \boldsymbol{X})$
	KL divergence	$\sum_{r\in\mathcal{Y}} \mathbb{P}_{\boldsymbol{Y}}(r) \log\Big(\frac{\mathbb{P}_{\boldsymbol{Y}}(r)}{\mathbb{P}_{\hat{m}(\boldsymbol{X})}(r)}\Big)$[11]	$\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$

Table 1: **The optimal predictors for standard loss functions can be derived from** $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$.

Supervised learning. Supervised learning seeks to find an optimal predictor m by using a learning algorithm[12] $I{:}\Delta \to \mathcal{M}$ that selects a a model $\hat{m}$ from a set $\mathcal{M}$ with the aid of a dataset $\mathcal{D} \coloneqq ((x^{(1)}, y^{(1)}), \ldots, (x^{(k)}, y^{(k)}))$ with $\mathcal{D}$ in the set of datasets Δ drawn i.i.d. from the joint distribution, i.e. $(x^{(i)}, y^{(i)}) \sim (\boldsymbol{X}, \boldsymbol{Y})$. Instead of the EPE itself, the learning algorithm minimizes the empirical risk on the *test data* (i.e. on data not used to train $\hat{m}$), which is an estimator of the EPE that can be computed from finite data.

4.2 IML Property Descriptors

We have just argued that ML models, when considered as a whole, approximate phenomenon aspects that can be derived from the conditional distribution $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$. *IML property descriptors* can help to investigate these aspects by describing their relevant properties.

Five Steps Towards IML Methods for Inference. Our proposal consists of the five steps in Figure 4, which we now discuss in detail. For each of the five steps, we provide an inference example based on the prediction of student grades in mathematics. In what follows, we assume that we have a supervised learning ML model $\hat{m}$ that approximates a phenomenon aspect described by the optimal predictor m.

Step 1) Formalize Scientific Question. Science starts by formulating a question. To address it with ML, this question has to be formalized. Exemplary questions that can already be addressed with IML methods following the scheme below are discussed in Section 4.3. Note that IML for scientific inference only helps answer questions that concern the association between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

[11] This describes the forward KL divergence $\mathrm{KL}(\boldsymbol{Y} \| \hat{m}(\boldsymbol{X}))$ for discrete $\boldsymbol{Y}$ and $\hat{m}(\boldsymbol{X})$, which differs from the backward KL divergence $\mathrm{KL}(\hat{m}(\boldsymbol{X}) \| \boldsymbol{Y})$.

[12] The domain of I is only completely specified when the parameters that define the learning procedure and the search space of the algorithm (called hyperparameters in the context of $\hat{m}$) are fixed. For our discussion, the reader may assume hyperparameters to have been set a priori by a human or an automated ML algorithm (Hutter et al. 2019).

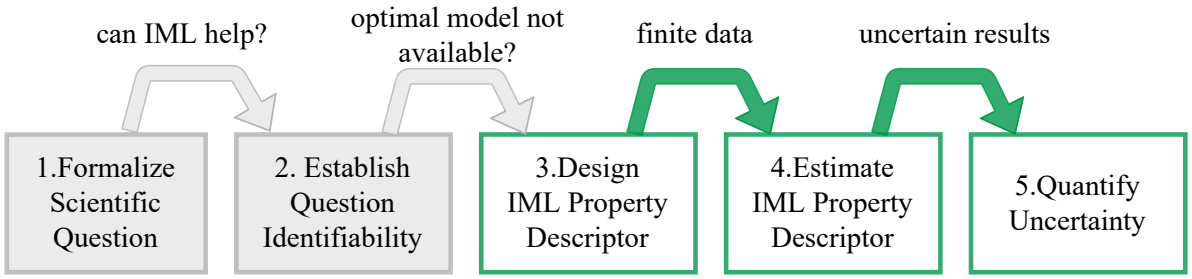


Figure 4: **An epistemic foundation for scientific inference with IML**. Steps 1 and 2 clarify *what* kinds of scientific inferences we can draw with IML. Steps 3, 4, and 5 show *how* to draw such inferences and provide estimates of their precision.

Formally. We denote the formalized question by Q.

Example. Suppose Laura wants to find out how students' language skills are related to their math skills. She approaches this problem by asking how students' expected math grades are related to their Portuguese grades. Laura formalizes this question as the conditional expectation i.e. $Q = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}_p}[\boldsymbol{Y} \mid \boldsymbol{X}_p]$, where $\boldsymbol{Y}, \boldsymbol{X}_p$ respectively stand for the math and Portuguese grade variable.[13]

Step 2) Establish Question Identifiability. Many scientific questions cannot be addressed using an ML model. ML models can only help answer questions that could theoretically be addressed with the optimal predictor. We call a question that the optimal predictor can help answer together with additional probabilistic knowledge (e.g. aspects of $\mathbb{P}(\boldsymbol{X}, \boldsymbol{Y})$) *identifiable*. A constructive strategy to establish identifiability relative to some probabilistic knowledge is to think of the transformations, using solely the probabilistic knowledge, that take the optimal predictor into the question Q. Of course, it is desirable to keep the amount of probabilistic knowledge required to identify the question to a minimum.

Formally. The optimal predictor is denoted by m. We say that a question is *identifiable* relative to probabilistic knowledge K if we can compute Q from m and K.

Example. Assume that Laura has trained her neural network, which, unlike the simple linear model presented above, takes into account *all* available features $\boldsymbol{X}$, to minimize the MSE loss, i.e. $m(\boldsymbol{X}) = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\boldsymbol{Y} \mid \boldsymbol{X}]$. Is Laura's question identifiable? For specific values of $\boldsymbol{X}$, the optimal predictor allows to compute the expected value of $\boldsymbol{Y}$ i.e. $m(x) = \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\boldsymbol{Y} \mid \boldsymbol{X}{=}x]$. The only difference to Q is that m takes into account features besides the Portuguese grade, that we denote $\boldsymbol{X}_{-p}$. If we have access to the conditional distribution $\mathbb{P}(\boldsymbol{X}_{-p} \mid \boldsymbol{X}_p)$ (required[14] probabilistic knowledge

[13] This conditional expectation is the best possible point estimate of the math grade under the MSE loss, given just the Portuguese grades.

[14] Usually we do not have access to probabilistic knowledge K. We discuss this in more detail in Step 4.

K), we can integrate these other features out by taking the expected value

$$\begin{aligned} Q &:= \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}_p}[\boldsymbol{Y} \mid \boldsymbol{X}_p] \\ &= \mathbb{E}_{\boldsymbol{X}_{-p}|\boldsymbol{X}_p}[\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\boldsymbol{Y} \mid \boldsymbol{X}] \mid \boldsymbol{X}_p] \qquad \text{(by the } \textit{tower rule}\text{, see App. C)} \\ &= \mathbb{E}_{\boldsymbol{X}_{-p}|\boldsymbol{X}_p}[m(\boldsymbol{X}) \mid \boldsymbol{X}_p]. \end{aligned}$$

Thus, Q is identifiable via m given $K = \mathbb{P}(\boldsymbol{X}_p \mid \boldsymbol{X}_{-p})$.

Step 3) Design IML Property Descriptor. It is not enough to identify a question. We need a way to estimate an answer for ML models — we need *IML property descriptors*. An IML property descriptor describes a continuous function that applies the transformation from the question identification step above to a given ML model and outputs an element of the space $\mathcal{Q}$. Thus, given the optimal predictor, an IML property descriptor outputs an answer to Q. Continuity guarantees that if our ML model is close to the optimal model, our answer is approximately correct. We call the application of a property descriptor to a specific ML model, $g_K(\hat{m})$, a *model property description*.

Formally. An *IML property descriptor* is a continuous function g_K (w.r.t. metrics $d_{\mathcal{M}}$ and $d_{\mathcal{Q}}$)[15] that identifies Q using probabilistic knowledge K:

$$g_K : \mathcal{M} \to \mathcal{Q} \quad \text{with} \quad g_K(m) = Q.$$

The output space $\mathcal{Q}$ remains unspecified to account for the variety of scientific questions; $\mathcal{Q}$ could denote a set of real numbers, vectors, functions, probability distributions, etc.

Example. The property descriptor describes the transformations that identify Q, i.e.

$$g_K(\hat{m})(x_p) := \mathbb{E}_{\boldsymbol{X}_{-p}|\boldsymbol{X}_p}[\hat{m}(\boldsymbol{X}) \mid \boldsymbol{X}_p{=}x_p]. \tag{4.1}$$

This is indeed a property descriptor because conditional expectation is continuous on $\mathcal{M}$, and Q is identifiable given $K = \mathbb{P}(\boldsymbol{X}_{-p} \mid \boldsymbol{X}_p)$. Note that Equation (4.1) describes the well-known conditional partial dependence plot, or cPDP, also known as M-plot (Molnar 2020, Apley and Zhu 2020).

Step 4) Estimate IML Property Descriptor. Often we lack access to relevant probabilistic knowledge K. Instead, we have a finite amount of data on which we can evaluate our ML mapping, which we call the *evaluation data*. It may bundle up our training and test data $\mathcal{D}$ (see Section 4.1), as well as additionally available (unlabeled) data, and artificially generated data. The *IML property description estimator* describes a way to estimate property descriptions with access only to the ML model plus the evaluation data.

[15] $d_{\mathcal{M}}$ is a metric on the function space $\mathcal{M}$, $d_{\mathcal{M}}(m_1, m_2) := \int_{\boldsymbol{X}} L(m_1(x), m_2(x))\, \mathbb{P}_{\boldsymbol{X}}(x)\, \mathrm{d}x$ for $m_1, m_2 \in \mathcal{M}$. $d_{\mathcal{Q}}$ describes a metric on space $\mathcal{Q}$.

Formally. We denote the *evaluation dataset* by $\mathcal{D}^*$ and the random process that generates it by $\boldsymbol{D}^*$. We call $\hat{g}_{\mathcal{D}^*} : \mathcal{M} \to Q$ the *IML property description estimator* if it is an unbiased estimator of g_K i.e.

$$\mathbb{E}_{\boldsymbol{D}^*}[\hat{g}_{\boldsymbol{D}^*}(\hat{m})] = g_K(\hat{m}) \quad \text{for all } \hat{m} \in \mathcal{M}.$$

Example. Laura's evaluation dataset $\mathcal{D}^*$ is her initial training and test dataset $\mathcal{D}$ augmented by artificial instances created by the following manipulation: Laura makes six copies of the data, and jitters the Portuguese grade by $1, -1, 2, -2, 3$ or -3 respectively. This augmentation strategy reflects how Laura understands the Portuguese grade as noisy based on her background knowledge (student performance varies daily and teachers may grade inconsistently). Let the students with (jittered) Portuguese grade i be $\mathcal{D}^*_{|x_p=i} := (x \in \mathcal{D}^* \mid x_p = i)$, then, we can define the IML property description estimator[16] at i as:

$$\hat{g}_{\mathcal{D}^*}(\hat{m})(i) := \frac{1}{|\mathcal{D}^*_{|x_p=i}|} \sum_{x \in \mathcal{D}^*_{|x_p=i}} \hat{m}(x) \tag{4.2}$$

The estimated answer to Laura's question is plotted in Figure 5. The plot on the left suggests that math grade is only strongly dependent on Portuguese grades in the interval $8 - 17$. However, as we show in the next step, we must also take into account that we have very sparse data in some regions (e.g. very few students scored below 8) before confirming this first impression.

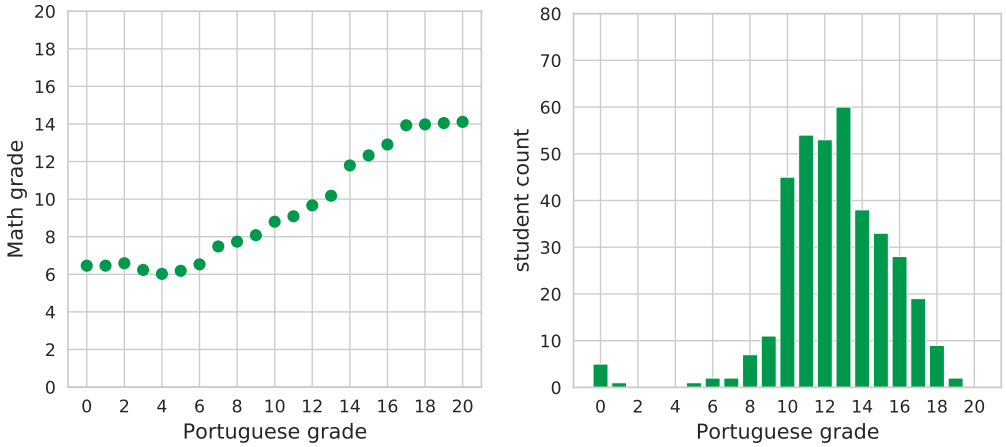


Figure 5: **Left Plot:** Estimate of $\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}_p}[\boldsymbol{Y} \mid \boldsymbol{X}_p]$ via Equation (4.2). **Right Plot:** Histogram of grades in Portuguese.

Step 5) Quantify Uncertainty. We have shown how we can estimate Q using an approximate ML model paired with a suitable evaluation dataset. But how good is our estimate? Two steps involve approximations:

[16] the conditional mean is an unbiased estimator of the conditional expectation

1. Applying the IML property descriptor to the ML model $\hat{m}$ instead of the optimal model m; we call the resulting error *model error*

$$\text{ME}[\hat{m}] = d_Q\Big(g_K(m), g_K(\hat{m})\Big).$$

The model error depends on the ML model $\hat{m}$ we obtained from the learning algorithm trained on the given dataset.

2. Applying the IML property description estimator on our evaluation dataset $\mathcal{D}^*$ instead of computing the true model property description based on K directly; we call this error *estimation error*

$$\text{EE}[\mathcal{D}^*] = d_Q\Big(g_K(\hat{m}), \hat{g}_{\mathcal{D}^*}(\hat{m})\Big).$$

The estimation error depends on the evaluation dataset $\mathcal{D}^*$.

In theory, the model error and the estimation error can be separated. In practice, however, they are statistically dependent because the training and the evaluation data overlap. Generally, neither the model error nor the estimation error can be computed perfectly; this would require access to the optimal model m and infinitely many data instances. Nevertheless, we can quantify in expectation how large the two errors are.

An intuitive approach to quantifying the expected errors is to decompose them into bias and variance contributions. The two decompositions below quantify the range in which the true phenomenon property descriptions are most likely to lie.

Formally. For the bias-variance decomposition, we assume the metric d_Q to be the squared error.[17] Considering the dataset that we entered into the learning algorithm as a random variable $\boldsymbol{D}$, we can decompose the expected $\text{ME}[\hat{m}]$ error as follows

$$\mathbb{E}_{\boldsymbol{D}}[\text{ME}[\hat{m}]] = \underbrace{\left(g_K(m) - \mathbb{E}_{\boldsymbol{D}}[g_K(\hat{m})]\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{V}_{\boldsymbol{D}}[g_K(\hat{m})]}_{\text{Variance}}$$

where $\hat{m} \coloneqq I(\mathcal{D})$ is the output of a machine learning algorithm I for dataset $\mathcal{D}$ (Section 4.1). Considering the evaluation data as a random variable $\boldsymbol{D}^*$, we can decompose the expected EE_{D^*} error as follows

$$\mathbb{E}_{\boldsymbol{D}^*}[\text{EE}[D^*]] = \underbrace{\left(g_K(\hat{m}) - \mathbb{E}_{\boldsymbol{D}^*}[\hat{g}_{\boldsymbol{D}^*}(\hat{m})]\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{V}_{\boldsymbol{D}^*}[\hat{g}_{\boldsymbol{D}^*}(\hat{m})]}_{\text{Variance}} = \mathbb{V}_{\boldsymbol{D}^*}[\hat{g}_{\boldsymbol{D}^*}(\hat{m})].$$

The bias term vanishes because the property description estimator is by definition unbiased w.r.t. the IML property descriptor.

Example. Laura obtains different cPDPs (Figure 5) for different models with similar performance, as well as for different selections of evaluation data, how much can she then rely on these cPDPs?

The estimates of the variances of the cPDP by Molnar et al. (2021) allow to calculate pointwise confidence intervals

[17] A bias-variance decomposition is also possible for other loss functions, including the 0-1 loss (Domingos 2000).

(Figure 6). We can define a confidence interval that only incorporates the estimation uncertainty by

$$\mathrm{CI}_{\mathrm{EE}[\boldsymbol{D}^*]} := \left[\hat{g}_{\mathcal{D}^*}(\hat{m})(i) \pm t_{1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}_{\boldsymbol{D}^*}[\hat{g}_{\boldsymbol{D}^*}(\hat{m})(i)]}\right]$$

and a confidence interval that incorporates both model and estimation uncertainty by

$$\mathrm{CI}_{\mathrm{ME}[\hat{m}]\wedge\mathrm{EE}[\boldsymbol{D}^*]} := \left[\hat{g}_{\mathcal{D}^*}(\hat{m})(i) \pm t_{1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}_{\boldsymbol{D},\boldsymbol{D}^*}[\hat{g}_{\boldsymbol{D}^*}(\hat{m})(i)]}\right].$$

For the combined confidence interval we require a strong and unfortunately not testable assumption to be satisfied — unbiasedness of the ML algorithm. Unbiasedness implies that, in expectation over training sets, the ML algorithm learns the optimal model, i.e. $m = \mathbb{E}_{\boldsymbol{D}}[\hat{m}]$.[18]

Figure 6 shows that for students with Portuguese grades between 8 and 17, Laura can be very confident in her model and the relationship it identifies between math and Portuguese grade.[19] However, both for Portuguese grades below 8 or above 17, the true value might be far off from our estimated value using a given model, as we can see from the width of the confidence intervals. For these grade ranges, gathering more data may reduce Laura's uncertainty.

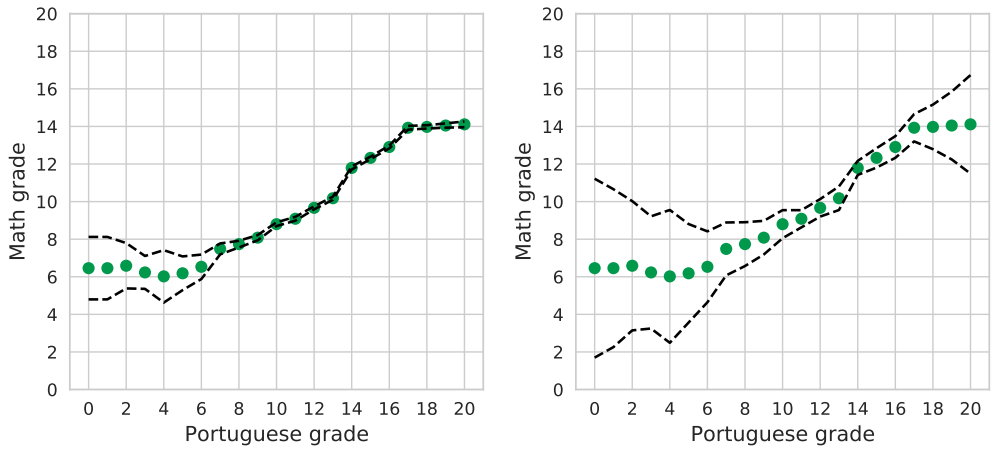


Figure 6: **Uncertainty evaluation of an IML property description**. **Left:** cPDP and its estimation error due to Monte-Carlo integration. **Right:** cPDP with *both* estimation and model error. Confidence bands in dashed lines cover the true expected math grade in 95% of all cases. These plots jointly suggest that most of the uncertainty is due to the model error.

Summary. Figure 7 gives an overview of all functions and spaces involved in IML for scientific inference. We started from a phenomenon and formalized a scientific question Q about it. Using a learning algorithm I on dataset $\mathcal{D}$ from the phenomenon, we learned an ML model $\hat{m}$ that approximates the optimal model m. We then set out to answer Q from $\hat{m}$. We defined a property descriptor g_K, that is, a function that allows to compute Q from m given K, respectively approximates Q from $\hat{m}$ given K. Because g_K requires probabilistic knowledge about $\mathbb{P}(\boldsymbol{X}, \boldsymbol{Y})$, we

[18] Since unbiasedness is tied to a specific context, there is no conflict with the no-free-lunch theorems (Sterkenburg and Grünwald 2021).

[19] We used resampling techniques to estimate the two variances. In real-data settings it is generally not possible to always sample new data for the model training and the evaluation. Although resampling may result in an underestimation of the variance, our goal here is simply to illustrate the process of quantifying uncertainty for a concrete IML method.

introduced a property description estimator $\hat{g}_{\mathcal{D}^*}$, a function estimating Q solely from finite data, the evaluation set $\mathcal{D}^*$. Finally, we showed how the expected error of our estimation steps can be quantified with confidence intervals $\text{CI}_{\text{ME}[\boldsymbol{D}]}$ and $\text{CI}_{\text{EE}[\boldsymbol{D}^*]}$.

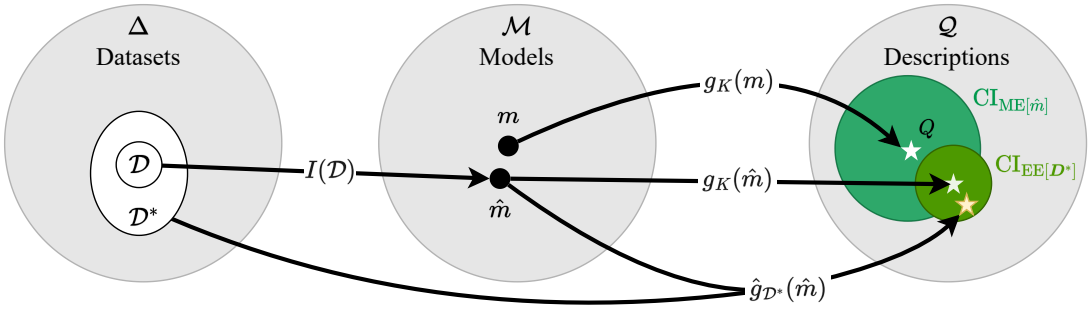


Figure 7: **From datasets to inferences via ML models**. Mappings are represented by arrows and sets are represented by filled circles, with confidence regions in green shades. Practical IML descriptions $\hat{g}_{\mathcal{D}^*}(\hat{m})$ are approximate, uncertainty-aware answers to a question Q that are built from a model $\hat{m}$ fit on $\mathcal{D}$ and an evaluation dataset $\mathcal{D}^*$.

4.3 Property Descriptors and Current IML Methods

Many questions that can be answered based on the conditional probability distribution $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$ are widely relevant. The goal of practical IML research for inference should be to define relevant descriptors and provide accessible implementations of these descriptors, including quantification of uncertainty. To find out which specific questions are relevant to scientists, and therefore what descriptors are necessary, IML researchers, statisticians and scientists must closely interact.

In Table 2 we present a few examples of elementary inference questions that can in principle be addressed by existing IML methods i.e. these methods can operate as property descriptors already. We distinguish between global and local phenomenon questions: global questions concern general associations, local questions concern associations for a specific instance. The last column highlights current IML methods that provide approximate answers, albeit often without uncertainty quantification. Note how we ultimately require conditional versions of existing marginal IML methods, which suggests that marginal sampling, which generates unrealistic instances, is inadequate in scientific inference.

5 Discussion

ER models enable straightforward scientific inference because their elements represent something about the underlying phenomenon. While ML models are generally not ER, IML can offer an indirect route to scientific inference, provided

[20] Only defined on phenomenon if $\mathbb{P}(X_p = p, X_{-p} = x_{-p}) > 0$.
[21] Only with the right similarity metric that accounts for the realistic constraint.

Global

Question	Formalization	IML method
Effect: What is the best estimate of $\boldsymbol{Y}$ if we only know $\boldsymbol{X}_p$?	$m_{\boldsymbol{X}_p}(\boldsymbol{X}_p)$	cPDP (Apley and Zhu 2020)
Conditional Contribution: How much worse can $\boldsymbol{Y}$ be predicted from $\boldsymbol{X}$ if we hadn't known $\boldsymbol{X}_p$?	$\text{EPE}_{\boldsymbol{X},\boldsymbol{Y}}(m_{\boldsymbol{X}}(\boldsymbol{X})) - \text{EPE}_{\boldsymbol{X}_{-p},\boldsymbol{Y}}(m_{\boldsymbol{X}_{-p}}(\boldsymbol{X}_{-p}))$	cPFI (Fisher et al. 2019)
Fair Contribution: What is the fair share of feature $\boldsymbol{X}_p$ in the prediction of $\boldsymbol{Y}$?	$\frac{1}{n}\sum_{S\subseteq\{1,\dots,n\}\setminus j}\binom{n-1}{\lvert S\rvert}^{-1}\big(\text{EPE}_{\boldsymbol{X}_{S\cup\{j\}},\boldsymbol{Y}}(m_{\boldsymbol{X}_{S\cup\{j\}}}(\boldsymbol{X}_{S\cup\{j\}})) - \text{EPE}_{\boldsymbol{X}_S,\boldsymbol{Y}}(m_{\boldsymbol{X}_S}(\boldsymbol{X}_S))\big)$	SAGE (Covert et al. 2020)
Relevant Value: Under which realistic conditions can we expect to observe relevant value y_{rel}?	$\underset{x\in\mathcal{X}\wedge\mathbb{P}(\boldsymbol{X}=x)>0}{\arg\min}\; d_{\mathcal{Y}}(m_{\boldsymbol{X}}(x), y_{rel})$	no method yet

Local

Effect: How does the best estimate of $\boldsymbol{Y}$ change relative to $\boldsymbol{X}_p$, knowing that $\boldsymbol{X}_{-p} = x_{-p}$?	$m_{\boldsymbol{X}}(\boldsymbol{X}_p, x_{-p})$	ICE-curve [20] (Goldstein et al. 2015)
Conditional Contribution: How much worse can $\boldsymbol{Y}$ be predicted from $\boldsymbol{X} = x$ if we hadn't known $\boldsymbol{X}_p$?	$L(y, m_{\boldsymbol{X}}(x)) - L(y, m_{\boldsymbol{X}_{-p}}(x_{-p}))$	no method yet
Fair Contribution: What is the fair share of feature $\boldsymbol{X}_p$ in the prediction of $\boldsymbol{Y}$ if $\boldsymbol{X} = x$?	$\frac{1}{n}\sum_{S\subseteq\{1,\dots,n\}\setminus j}\binom{n-1}{\lvert S\rvert}^{-1}\big(m_{\boldsymbol{X}_{S\cup\{j\}}}(x_S, x_j) - m_{\boldsymbol{X}_S}(x_S)\big)$	conditional Shapley Values (Aas et al. 2021)
Relevant Value: Under which realistic conditions similar to $\boldsymbol{X} = x$ can we expect to observe relevant value y_{rel}?	$\underset{x'\in\mathcal{X}\wedge\mathbb{P}(\boldsymbol{X}=x')>0}{\arg\min}\; d_{\mathcal{Y}}(m_{\boldsymbol{X}}(x'), y_{\text{rel}}) + \lambda\, d_{\mathcal{X}}(x, x')$	Counterfactuals [21] (Dandl et al. 2020)

Table 2: **Global and local formalized questions and matching IML property descriptors**. Note that questions are relative to a specific loss function L; for a set $S \subseteq \{1, \dots, n\}$, the term $m_{\boldsymbol{X}_S}$ describes the optimal predictor of $\boldsymbol{Y}$ w.r.t. loss function L and random variable(s) $\boldsymbol{X}_i$ with $i \in S$. $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$ describe suitable metrics on $\mathcal{X}$ and $\mathcal{Y}$ respectively.

model properties have a corresponding phenomenon counterpart. We have shown how phenomenon representation can be achieved through optimal predictors and described how to practically construct IML property descriptors following five-steps: the first two steps clarify what questions we can address with IML, step three and four show how to answer them with ML models and finite data, and step five allows to evaluate how certain the answers are. We pointed out that some current IML methods can already be seen as IML property descriptors.

Is the lack of elementwise representationality specific to ML models? No, ML shows only an extreme case. In fact, there is a continuum between fully ER models and HR-only models: Some scientific models contain elements that are difficult or impossible to interpret e.g. the wave function in physics (Callender 2015); complex classical statistical models like generalized additive models also contain elements that are difficult to interpret. Our main message is: the five-step approach can be used to extend inference to any non ER model (whether ML or not).

One could argue that science should only rely on ER models (Rudin 2019). Indeed, it would be great if we could always build models from simple to complex and keep ER from beginning to end. However, more and more problems

seem to be very difficult to tackle with this approach (Nearing et al. 2021); Interpretable but inaccurate models (w.r.t. to the phenomenon) are not a solution (Breiman et al. 2001). In situations where we cannot construct accurate ER models because we lack background knowledge or the phenomenon is very complex, scientific inference with ML models may thus be the only viable alternative.

5.1 Implications

Adopting a phenomenon-centric perspective on IML allows us to answer a variety of questions that were puzzling from a model-centric perspective:

Which questions can be addressed with IML property descriptors? IML property descriptors can help retrieve relevant phenomenon properties i.e. properties derived from the conditional distribution $\mathbb{P}(\boldsymbol{Y} \mid \boldsymbol{X})$. Which phenomenon properties are relevant is context-specific and up to researchers to identify. While formulating questions, researchers must be aware that supervised ML models are only representational of associative structure and not the underlying causal mechanism (see Section 5.3).

Why use (I)ML for inference? Supervised ML can help draw scientific inference when sampling from $\boldsymbol{X}$ is easy but sampling from $\boldsymbol{Y}$ is difficult, e.g. when $\boldsymbol{Y}$ is hard to measure or determined only in the future. In such situations, analyzing both the model and the data with IML methods can allow for better conclusions than analyzing just the data — the ML model fills the gaps by interpolation. Extrapolation to out-of-distribution data is generally not a strength of ML and can lead to incorrect conclusions; such extrapolations should only be trusted if the learning algorithm incorporated a powerful and suitable inductive bias.

When sampling from $\boldsymbol{X}$ is difficult or the property of interest can be computed more reliably by other means, we advice against using IML for inference.

How important is model performance in inference? If the model is a poor approximation or representation of the modeled phenomenon, the conclusions we draw from that model are unreliable (Cox 2006, Good and Hardin 2012). Thus, a good fit is vital for gaining reliable knowledge.

Note that even for the optimal model, there remains the so-called Bayes error rate, an irreducible error arising from the fact that $\boldsymbol{X}$ does not completely determine $\boldsymbol{Y}$ (Hastie et al. 2009). Thus, high error does not necessarily flag a low-quality model, but rather may indicate that $\boldsymbol{X}$ provides insufficient information about $\boldsymbol{Y}$.

What kind of data should be used for IML? Many IML methods (e.g. Shapley Values, LIME, etc.) rely on probing the ML model on permuted data (Scholbeck et al. 2019). These artificial "data" may never occur in the real world. This may be useful to audit the model, but if we want to learn about the world, artificial data is supposed to credibly supplement observations. Our analysis therefore substantiates the criticism of Hooker and Mentch (2021), Hooker et al. (2021), Mentch and Hooker (2016) concerning the permutation of features irrespective of the dependency structure in the data.

5.2 Open Problems

There are several open issues that we have not addressed:

What about non-tabular data? For some data types, such as images, audio, or video data, it is extremely difficult to formulate scientific questions only in terms of low-level features such as pixels or audio frequencies. To follow our approach, we need a translation of high-level concepts (e.g. objects in images or words in audio) that scientists can use to formulate their questions into low-level features (e.g. pixels or audio frequencies) that the model works with. Such translations are notoriously difficult to find; deep learning may help here (Jia et al. 2013, Zaeem and Komeili 2021, Zhou et al. 2018, Koh et al. 2020).

How to assess if data is realistic? In IML, we often need to augment our data. However, using unrealistic data is highly problematic for scientific inference, as mentioned earlier. Reasonable permutations of features such as Laura's grade jitter strategy (see Section 4.2), can supply realistic data. However, this requires expert knowledge about what permutations make sense. Conditional density estimation techniques or generative models (e.g. generative adversarial networks, normalizing flows, variational autoencoders, etc.) may provide additional paths to obtain realistic data. However, modeling the conditional density can be computationally intensive and more difficult than the original prediction problem, or may even be epistemically problematic since it only approximates sampling real data.

To what extent does a property determine the true model? Sometimes, we know that the model answer to a scientific question is correct. How strongly does this confirm the correctness of the model? Property descriptions narrow down the potential models and sufficiently many property descriptions can even completely determine the model, e.g. for the FANOVA decomposition (Apley and Zhu 2020, Hooker 2004). Model property descriptions may eventually be used to incorporate background knowledge in training. Both directions, extracting knowledge from ML models, and using background knowledge to build more adequate ML models, are elementary for scientific progress (Dwivedi et al. 2021, Nearing et al. 2021, Razavi 2021).

5.3 Other Forms of Scientific Inference With ML

In this paper, we focused exclusively on scientific inference with supervised learning ML models on i.i.d. data. For this setting, there is sufficient theory in both statistical decision theory and IML research to provide secure epistemic foundations for scientific inference. We have explained what we can learn about the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. We can even learn that $\boldsymbol{X}$ contains little information about $\boldsymbol{Y}$ to predict it, which is scientifically interesting (Taleb 2005, Shmueli et al. 2010). However, many questions that scientists regularly face are of a different nature and go beyond conditional distributions.

Unsupervised Learning. Unsupervised learning is concerned with estimating aspects of the joint distribution $\mathbb{P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$. Unsupervised learning is hard as it typically targets a high-dimensional joint distribution and, often,

lacks a clear measure of success (Hastie et al. 2009, p486). In principle, our five-step guide is also applicable to unsupervised learning, however, we lack a theoretical counterpart to optimal predictors.

Causal Learning. The observational joint probability distribution is interesting, but it remains on rung one of Judea Pearl's ladder of causation — the associational level (Pearl and Mackenzie 2018). What scientists are often much more interested in is answering causal questions such as average treatment effects (rung 2) or counterfactual questions (rung 3) (Salmon 1998, Woodward and Ross 2021). Laura may be interested not only whether students' language and math skills are associated (rung 1), but also in whether the provision of tutoring in Portuguese affects students' math skills (rung 2) or whether a specific student (who is not a native Portuguese speaker) would have done better in math if she had received Portuguese tutoring at a young age (rung 3).

Supervised ML models only represent aspects of the observational distribution (rung 1) and therefore generally do not allow answering causal questions. As a consequence, the IML descriptors of the models also generally do not allow causal insight into the data. Many IML papers that discuss causality (Schwab and Karlen 2019, Janzing et al. 2020, Wang et al. 2021, Heskes et al. 2020) are only concerned with causal effects on the model's prediction, which do not necessarily translate into a causal insight into the phenomenon.

In order to answer causal questions, causal models should be used instead.[22] To learn a causal model, we must gather interventional data and/or make strong, untestable assumptions. Causal inference constitutes thus a challenging problem and remains an active area of research (Heinze-Deml et al. 2018, Kalisch and Bühlmann 2014, Constantinou et al. 2021, Peters et al. 2017).

In certain situations, ML models can nevertheless be useful for causal inference. Firstly, if all predictor variables are causally independent and the prediction target is caused by the features, the causal model interpretation implies the causal data interpretation. Secondly, associative models in combination with IML can help estimate causal effects even in the absence of causal independence if they are in principle identifiable by observation. For example, the partial dependence plot coincides with the so-called adjustment formula and therefore identifies a causal effect if the backdoor criterion is met (and the model optimally predicts the conditional expectation) (Zhao and Hastie 2021). Thirdly, when there is access to observational and interventional data during training, training ML models with invariant risk minimization yields models that predict accurately in interventional environments (Peters et al. 2016, Pfister et al. 2021, Arjovsky et al. 2019). For such intervention-stable models, IML methods that provide insight into the effect of interventions on the prediction also describe causal effects on the underlying real-world components (König et al. 2021a).

Another way in which ML supports causal inference is by facilitating practical scientific inference relying on complex mechanistical models, frequently implemented as numerical simulators. Indeed simulators can represent complex, causal, dynamics in an ER fashion, but often at the price of an intractable likelihood and thus expensive inference. A variety of new ML methods for likelihood-free inference on simulators (Cranmer et al. 2020) allows to

[22]Given a causal graph, observational data can allow to identify average causal effects (rung 2), e.g. with the so-called backdoor criterion (Pearl 2009). For estimating counterfactuals (rung 3), assumptions beyond a causal graph and observational data must be met (Holland 1986, Peters et al. 2017).

estimate a full posterior distribution over ER parameters for increasingly complex models.

While supervised learning learns from a fixed dataset, reinforcement learning (RL) systems are designed to act and can therefore assess the effect of interventions. As such, RL models can be designed to provide causal interpretations (Bareinboim et al. 2015, Zhang and Bareinboim 2017, Gasse et al. 2021).

6 Conclusion

Traditional scientific models were designed to satisfy elementwise representationality. This allowed scientists to directly inspect model elements to learn about Nature. Although ML models do not satisfy elementwise representationality, we have showed that it is still possible to learn about the phenomenon using them. All we need to do is to interrogate the model with suitable IML property descriptors. We have shown how such descriptors must be designed to enable scientific inference.

Funding

This work was supported by the Graduate School of Systemic Neuroscience (GSN) of the LMU Munich and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A, and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

Acknowledgements

We are very significantly indebted to Tom Sterkenburg, Seth Axen, Thomas Grote, Alexandra Gessner, Nacho Molina and Christian Scholbeck for their comments on the manuscript and their hints to related literature. Moreover, we thank the workshop participants of the Tübingen workshop "Philosophy of Science Meets Machine Learning" for the insightful comments and discussions after presenting some of the ideas described in the paper.

References

Aas, K., Jullum, M., and Løland, A. (2021). Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502.

Achinstein, P. (1974). Concepts of science. *Philosophy*, 49(187).

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., and Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285.

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):16–07.

Apley, D. W. and Zhu, J. (2020). Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Bailer-Jones, D. M. (2003a). Models, theories and phenomena. *Proceedings of Logic Methodology and Philosophy of Science*.

Bailer-Jones, D. M. (2003b). When scientific models represent. *International studies in the philosophy of science*, 17(1):59–74.

Bailer-Jones, D. M. and Bailer-Jones, C. A. (2002). Modeling data: Analogies in neural networks, simulated annealing and genetic algorithms. In *Model-Based Reasoning*, pages 147–165. Springer.

Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549.

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., and Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657.

Bickler, S. H. (2021). Machine learning arrives in archaeology. *Advances in Archaeological Practice*, 9(2):186–191.

Boge, F. J. (2022). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, 32(1):43–75.

Bokulich, A. (2011). How scientific models can explain. *Synthese*, 180(1):33–45.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese*, 195(12):5339–5372.

Buckner, C. and Garson, J. (2019). Connectionism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition.

Callender, C. (2015). One world, one beable. *Synthese*, 192(10):3153–3177.

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. (2020). True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

Cichy, R. M. and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4):305–317.

Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K. (2021). Large-scale empirical validation of bayesian network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning*, 131:151–188.

Contessa, G. (2007). Scientific representation, interpretation, and surrogative reasoning. *Philosophy of science*, 74(1):48–68.

Cortez, P. and Silva, A. (2008). Using data mining to predict secondary school student performance. *EUROSIS*.

Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223.

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge university press.

Cranmer, K., Brehmer, J., and Louppe, G. (2020). The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. (2020). Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pages 448–469. Springer.

Domingos, P. (2000). A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, pages 231–238. Morgan Kaufmann Stanford.

Donnelly, J. and Roegiest, A. (2019). On interpretability and feature representations: an analysis of the sentiment neuron. In *European Conference on Information Retrieval*, pages 795–802. Springer.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Douglas, H. E. (2009). Reintroducing prediction to explanation. *Philosophy of Science*, 76(4):444–463.

Dwivedi, D., Nearing, G., Gupta, H., Sampson, A. K., Condon, L., Ruddell, B., Klotz, D., Ehret, U., Read, L., Kumar, P., et al. (2021). Knowledge-guided machine learning (kgml) platform to predict integrated water cycle and associated extremes. Technical report, Artificial Intelligence for Earth System Predictability.

Farrell, S., Calafiura, P., Mudigonda, M., Anderson, D., Vlimant, J.-R., Zheng, S., Bendavid, J., Spiropulu, M., Cerati, G., Gray, L., et al. (2018). Novel deep learning methods for track reconstruction. *arXiv preprint arXiv:1810.06111*.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81.

Friedman, J. H. et al. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Frigg, R. and Nguyen, J. (2021). Scientific Representation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Gale, E. M., Martin, N., Blything, R., Nguyen, A., and Bowers, J. S. (2020). Are there any 'object detectors' in the hidden layers of cnns trained to identify objects or scenes? *Vision Research*, 176:60–71.

Gasse, M., Grasset, D., Gaudron, G., and Oudeyer, P.-Y. (2021). Causal reinforcement learning using observational and interventional data. *arXiv preprint arXiv:2106.14421*.

Gibson, P. B., Chapman, W. E., Altinok, A., Delle Monache, L., DeFlorio, M. J., and Waliser, D. E. (2021). Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Communications Earth & Environment*, 2(1):1–13.

Gobet, F. (2018). Three views on expertise: Philosophical implications for rationality, knowledge, intuition and education. *Education and Expertise*, pages 58–74.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65.

Good, P. I. and Hardin, J. W. (2012). *Common errors in statistics (and how to avoid them)*. John Wiley & Sons.

Gühring, I., Raslan, M., and Kutyniok, G. (2020). Expressivity of deep neural networks. *arXiv preprint arXiv:2007.04759*.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Heinze-Deml, C., Maathuis, M. H., and Meinshausen, N. (2018). Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33:4778–4789.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Hooker, G. (2004). Discovering additive structure in black box functions. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 575–580.

Hooker, G. and Mentch, L. (2021). Bridging breiman's brook: From algorithmic modeling to statistical learning. *Observational Studies*, 7(1):107–125.

Hooker, G., Mentch, L., and Zhou, S. (2021). Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31(6):1–16.

Hutter, F., Kotthoff, L., and Vanschoren, J. (2019). *Automated machine learning: methods, systems, challenges*. Springer Nature.

Janzing, D., Minorics, L., and Blöbaum, P. (2020). Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pages 2907–2916. PMLR.

Jia, Y., Abbott, J. T., Austerweil, J. L., Griffiths, T., and Darrell, T. (2013). Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. *Advances in Neural Information Processing Systems*, 26.

Kalisch, M. and Bühlmann, P. (2014). Causal structure learning and inference: a selective review. *Quality Technology & Quantitative Management*, 11(1):3–21.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big data & society*, 1(1):2053951714528481.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR.

König, G., Freiesleben, T., and Grosse-Wentrup, M. (2021a). A causal perspective on meaningful and robust algorithmic recourse. *arXiv preprint arXiv:2107.07853*.

König, G., Molnar, C., Bischl, B., and Grosse-Wentrup, M. (2021b). Relative feature importance. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9318–9325. IEEE.

Levy, A. (2012). Models, fictions, and realism: Two packages. *Philosophy of Science*, 79(5):738–748.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Longino, H. E. (2018). *The fate of knowledge*. Princeton University Press.

Luan, H. and Tsai, C.-C. (2021). A review of using machine learning approaches for precision education. *Educational Technology & Society*, 24(1):250–266.

Luk, R. W. (2017). A theory of scientific study. *Foundations of Science*, 22(1):11–38.

Mayer-Schönberger, V. and Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1987). *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, volume 2. MIT press.

Mentch, L. and Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Molnar, C., Casalicchio, G., and Bischl, B. (2020a). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer.

Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M. N., and Bischl, B. (2021). Relating the partial dependence plot and permutation feature importance to the data generating process. *arXiv preprint arXiv:2109.01433*.

Molnar, C., König, G., Bischl, B., and Casalicchio, G. (2020b). Model-agnostic feature importance and effects with dependent features–a conditional subgroup approach. *arXiv preprint arXiv:2006.04628*.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B. (2022). General pitfalls of model-agnostic interpretation methods for machine learning models. In Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W., editors, *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 39–68, Cham. Springer International Publishing.

Mu, J. and Andreas, J. (2020). Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning? *Water Resources Research*, 57(3):e2020WR028091.

Nguyen, A., Yosinski, J., and Clune, J. (2016). Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *arXiv preprint arXiv:1602.03616*.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Olah, C., Mordvintsev, A., and Schubert, L. (2017). Feature visualization. *Distill*, 2(11):e7.

Pearl, J. (2009). *Causality*. Cambridge university press.

Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic books.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016). Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P. (2021). Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3):1220–1246.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., and Barnes, P. (2020). Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 33–44.

Räz, T. (2022). Understanding deep learning with statistical relevance. *Philosophy of Science*, 89(1):20–41.

Razavi, S. (2021). Deep learning, explained: Fundamentals, explainability, and bridgeability to process-based modelling. *Environmental Modelling & Software*, 144:105159.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM.

Ritchey, T. (2012). Outline for a morphology of modelling methods. *Acta Morphologica Generalis AMG Vol*, 1(1):1012.

Roscher, R., Bohn, B., Duarte, M. F., and Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216.

Rosser, J. S. J. B. et al. (2008). A nobel prize for asymmetric information: the economic contributions of george akerlof, michael spence and joseph stiglitz. In *Leading Contemporary Economists*, pages 162–181. Routledge.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Salmon, W. C. (1979). Why ask,'why?'? an inquiry concerning scientific explanation. In *Hans Reichenbach: logical empiricist*, pages 403–425. Springer.

Salmon, W. C. (1998). *Causality and explanation*. Oxford University Press.

Schmidt, J., Marques, M. R., Botti, S., and Marques, M. A. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5(1):1–36.

Scholbeck, C. A., Molnar, C., Heumann, C., Bischl, B., and Casalicchio, G. (2019). Sampling, intervention, prediction, aggregation: a generalized framework for model-agnostic interpretations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 205–216. Springer.

Schwab, P. and Karlen, W. (2019). Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems*, 32.

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Towards causal representation learning.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710.

Shahhosseini, M., Hu, G., and Archontoulis, S. V. (2020). Forecasting corn yield with machine learning ensembles. *arXiv preprint arXiv:2001.09055*.

Shmueli, G. et al. (2010). To explain or to predict? *Statistical science*, 25(3):289–310.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

Spinney, L. (2022). Are we witnessing the dawn of post-theory science? *The Guardian*.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Stachl, C., Au, Q., Schoedel, R., Gosling, S. D., Harari, G. M., Buschek, D., Völkel, S. T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., and Bühner, M. (2020). Predicting personality from patterns of behavior collected with smartphones. *Proceedings of the National Academy of Sciences*, 117(30):17680–17687.

Stachowiak, H. (1973). *Allgemeine modelltheorie*. Springer.

Sterkenburg, T. F. and Grünwald, P. D. (2021). The no-free-lunch theorems of supervised learning. *Synthese*, 199(3):9979–10015.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC bioinformatics*, 9(1):1–11.

Štrumbelj, E. and Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

Sullivan, E. (2020). Understanding from machine learning models. *The British Journal for the Philosophy of Science*.

Suppes, P. (1966). Models of data. In *Studies in logic and the foundations of mathematics*, volume 44, pages 252–261. Elsevier.

Taleb, N. (2005). The black swan: Why don't we learn that we don't learn. *NY: Random House*.

Toulmin, S. E. (1961). *Foresight and understanding: An enquiry into the aims of science*. Greenwood Press.

Voss, C., Cammarata, N., Goh, G., Petrov, M., Schubert, L., Egan, B., Lim, S. K., and Olah, C. (2021). Visualizing weights. *Distill*, 6(2):e00024–007.

Wachter, S., Mittelstadt, B., and Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99.

Wang, J., Wiens, J., and Lundberg, S. (2021). Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR.

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, 200(1):1–33.

Woodward, J. and Ross, L. (2021). Scientific Explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

Zaeem, M. N. and Komeili, M. (2021). Cause and effect: Concept-based explanation of neural networks. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2730–2736. IEEE.

Zednik, C. (2021). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2):265–288.

Zednik, C. and Boelsen, H. (2022). Scientific exploration and explainable artificial intelligence. *Minds and Machines*, pages 1–21.

Zhang, J. and Bareinboim, E. (2017). Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780.

Zhang, Z., Jin, Y., Chen, B., and Brown, P. (2019). California almond yield prediction at the orchard level with a machine learning approach. *Frontiers in Plant Science*, 10:809.

Zhao, Q. and Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281.

Zhou, B., Sun, Y., Bau, D., and Torralba, A. (2018). Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134.

Appendix A Background on Models and Phenomena

We follow Bailer-Jones ([2003b], p61) and others (Achinstein 1974, Levy 2012, Contessa 2007) in seeing models as "an interpretative description of a phenomenon that facilitates perceptual as well as intellectual access to that phenomenon", where a phenomenon describes a fact or event in nature that is subject to be researched (Bailer-Jones 2003a). Phenomenon and scientific models have been described as a continuous hierarchy with data living close to the phenomenon and the model close to theory (Suppes 1966). Models represent only some phenomenon aspects but not others (Ritchey 2012, Bailer-Jones 2003b, Frigg and Nguyen 2021); a good model is true to the aspects that are relevant to the model user (Bailer-Jones 2003b, Stachowiak 1973).

Appendix B Dataset

Figure 8 gives a descriptions of the different features and is copied from Cortez and Silva (2008). In our trained models, we only used the final G3 student grades. The data was collected during 2005 and 2006 from two public schools, from the Alentejo region in Portugal. The database is collected from a variety of sources from both school reports and questionnaires. Cortez and Silva (2008) integrated the information into a mathematics dataset (with 395 examples) and a Portuguese language dataset (649 records).

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

Figure 8: **Attributes in the Cortez and Silva (2008) dataset**.

Appendix C Tower Rule for Expectations

For arbitrary random variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ holds that

$$\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\boldsymbol{Y} \mid \boldsymbol{X}] = \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X}}\Big[\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}[\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z}] \mid \boldsymbol{X}\Big].$$

This is also known as the rule of total expectation. Intuitively it says that it doesn't matter if we directly take the expectation of $\boldsymbol{Y}$ on $\boldsymbol{X}$ or if we first take the expectation of $\boldsymbol{Y}$ conditioned on a set of random variables $\boldsymbol{X}, \boldsymbol{Z}$ that includes $\boldsymbol{X}$ and then, "integrate $\boldsymbol{Z}$ out".

Chapter 3

Paper II: Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process

Freiesleben, T., Molnar, C., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., Wright, M. and Bischl, B. (unpublished). Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. *Under review at 'Machine Learning', status: major revision, this is the revised version sent to the journal.*

Author contributions:

T.F., C.M, and G.K. contributed equally to this work. C.M. developed the initial idea and wrote the initial draft. **T.F.** generalized the initial definitions and theorems to arbitrary sampling procedures and contributed the corresponding paragraphs concerning the sampling and extrapolation problem in Section 2 and Section 5. Moreover, **T.F.** and G.K. made conceptual contributions (particularly on Section 2), contributed some of the proofs (particularly Theorem 3), restructured the manuscript, and contributed the motivating example. Also, **T.F.** carefully revised and proofread all proofs and formal definitions. G.K. contributed the application section. C.M, M.W., J.H., and T.R. implemented and run the simulation study. **All authors** added valuable new discussion points and helped revise the text.

Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process

Timo Freiesleben[1,4,5*†], Christoph Molnar[3,1†], Gunnar König[1,2†], Julia Herbinger[1], Tim Reisinger[1], Giuseppe Casalicchio[1], Marvin N. Wright[3,6,7] and Bernd Bischl[1]

[1]Department of Statistics, LMU Munich, Munich, Germany.
[2]University of Vienna, Vienna, Austria.
[3]Leibniz Institute for Prevention Research and Epidemiology – BIPS, Bremen, Germany.
[4]Graduate School of Systemic Neurosciences, LMU Munich, Munich, Germany.
[5] Cluster of Excellence Machine Learning, Tübingen, Germany.
[6]University of Bremen, Bremen, Germany.
[7]University of Copenhagen, Copenhagen, Denmark.

*Corresponding author(s). E-mail(s): timo.freiesleben@campus.lmu.de;
†These authors contributed equally to this work.

Abstract

Scientists and practitioners increasingly rely on machine learning to model data and draw conclusions. Compared to statistical modeling approaches, machine learning makes fewer explicit assumptions about data structures, such as linearity. Consequently, the parameters of machine learning models usually cannot be easily related to the data generating process. To learn about the modeled relationships, partial dependence (PD) plots and permutation feature importance (PFI) are often used as interpretation methods. However, PD and PFI lack a theory that relates them to the data generating process. We formalize PD and PFI as statistical estimators of ground truth estimands rooted in the data generating process. We show that PD and PFI estimates deviate from this ground truth due to statistical biases, learner variance, and Monte Carlo

approximation errors. To account for learner variance in PD and PFI estimation, we propose the learner-PD and the learner-PFI based on model refits and propose corrected variance and confidence interval estimators.

Keywords: Interpretable Machine Learning, Permutation Feature Importance, Partial Dependence Plot, Statistical Inference, Uncertainty Quantification

1 Introduction

Statistical models such as linear or logistic regression models are frequently used to learn about relationships in data. Assuming that a statistical model reflects the data generating process (DGP) well, we may interpret the model coefficients in place of the DGP and draw conclusions about the data. An important part of interpreting the coefficients is the quantification of their uncertainty via standard errors, which allows separation of random noise (non-significant coefficients) from real effects. Statistical biases and violation of assumptions are well-studied for many model classes, such as heterogeneous residuals, deviations from normality, and non-additivity for linear models [1].

Increasingly, machine learning approaches – such as gradient-boosted trees, random forests or neural networks – are being used instead of or in addition to statistical models. Compared to statistical models that are driven by considerations of the DGP, machine learning approaches often lack a mapping between model parameters and properties of the DGP. Due to the ability of many machine learning models to address highly non-linear relationships and interactions, they often outperform more restrictive statistical models.

Scientific applications of machine learning are widespread and range from modeling volunteer labor supply [2], mapping fish biomass [3], analyzing urban reservoirs [4], identifying disease-associated genetic variants [5], to inferring behavior from smartphone use [6]. In these scientific applications, the model is only the means to an end: a better understanding of the DGP, in particular to learn what features are predictive of the target variable.

Model-agnostic interpretation methods [7] are a (partial) remedy to the lack of interpretable parameters of more complex models. Model-agnostic methods follow a general procedure of 1) sampling data, 2) manipulating this data, 3) predicting and finally 4) aggregating the predictions [8]. Since none of these steps depends on specific model properties, model-agnostic interpretation techniques allow us to study the behavior of arbitrary models. Partial dependence (PD) plots [9] and permutation feature importance (PFI) [10, 11] are popular model-agnostic methods for describing the relationship between input features and model outcome on a global level. PD plots visualize the average effect that features have on the prediction, and PFI estimates how much each feature contributes to the model performance and therefore how relevant a feature is.

Scientists who want to use PD and PFI to draw conclusions about the DGP face a problem as these methods have been designed to describe the prediction function, but lack a theory linking them to the DGP. In particular, the uncertainty of PD and PFI with respect to the DGP is not quantified, making it hard for scientists to assess the extent to which they are justified to draw conclusions based on the PD and PFI applied to a single ML model. Treating PD and PFI as statistical estimators (like coefficients in a regression model) would allow us to quantify this epistemic uncertainty and remedy these concerns. However, this requires a theoretical counterpart of PD and PFI in the DGP: a ground truth estimand that these interpretation methods are intended to retrieve.

Contributions

We are the first to treat PD and PFI as statistical estimators of ground truth properties in the DGP. We introduce two notions, model-PD/PFI and learner-PD/PFI, which allow to analyze the uncertainty due to Monte-Carlo integration and uncertainty due to the training data/process, respectively. We propose bias-variance decompositions, theorems of PD/PFI unbiasedness, standard estimators, and confidence intervals for both PD and PFI. In addition, we leverage a variance correction approach from model performance estimation [12] to improve the variance estimation. We demonstrate the quality of our proposed confidence intervals in simulations and their usefulness on a medical example.

Structure

We start with a motivating example (Section 1.1) and a discussion of related work (Section 1.2). In the methods section (Section 2), we introduce PD and PFI formally, relate them to the DGP, and provide bias-variance decompositions, variance estimators and confidence intervals. In the simulation study in Section 3, we test our proposed methods in various settings and compare them to alternative approaches. In the application in Section 4, we demonstrate how our confidence intervals for PD/PFI may help scientists to draw more justified conclusions about the DGP. Finally, we discuss the limitations of our work in Section 5.

1.1 Motivating Example

Imagine a researcher who wants to study chronic heart disease. She has data available from the UCI machine learning repository [13] (Cleveland data, $n = 296$) containing sociological and medical indicators such as age, blood pressure and maximum heart rate. She wants to use machine learning methods to predict heart disease, and also to learn about the most predictive features. She compares the performance w.r.t. the predicted probabilities of a logistic regression model, a decision tree (CART) [14], and a random forest classifier [10] using 5-fold cross validation measured by the Brier score on the dataset; the mean losses for the different models are: 0.199 (logistic regression), 0.250

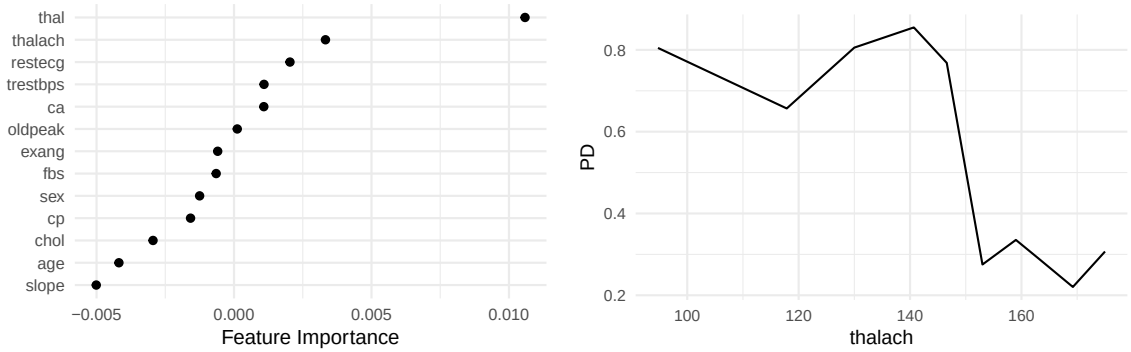


Figure 1 Left: Conditional Feature Importance. Right: Conditional Partial Dependence Plot for the maximum heart rate (`thalach`).

(tree), and 0.130 (random forest). Since the random forest outperforms the linear model and decision tree, she uses a random forest for further analysis; she fits a random forest on 60 percent of the data and uses the remaining 40 percent as test set.[1]

In order to learn about how predictive the features are in the data generating process (DGP), she applies the PD and PFI to her trained model. The features, such as age and blood pressure, are strongly correlated. To get interpretations that are true to the data and that avoid extrapolation, she employs conditional sampling based versions of PD and PFI (for a discussion of marginal versus conditional sampling, we refer to the literature [15, 16], Section 2.1, and Section 2.2). The conditional PD corresponds to the expected prediction and therefore indicates how the probability of having heart disease varies with the feature of interest [16]. Conditional feature importance quantifies the surplus contribution of each feature over the remaining features (and can be linked to conditional dependence with the prediction target [17, 18]).

Conditional interpretation methods require sampling from conditional distributions. She samples categorical variables using a log-loss optimal classifier; and samples continuous variables by predicting the conditional mean and resampling residuals (thereby assuming homoscedasticity). For all sampling tasks, she fits a random forests once on the dataset. In order to model multivariate mixed distributions, she employs a sequential design [19, 20].

The results (Figure 1) match with the researcher's intuition. Many conditional PFI values are small, indicating that the features are dependent and could be replaced with the remaining variables. The most important features are thalassemia (`thal`), the maximum heart rate (`thalach`), and resting state ECG (`restecg`). In order to further understand the association with the maximum heart rate, she inspects the corresponding conditional PD plot. She observes that the probability of having chronic heart disease drops when faster maximum heart rates were observed.

Although the researcher finds the results plausible, she is unsure whether her conclusions extend to properties of the data generating process. Various uncertainties could influence her result: The feature importance and conditional

[1] All code is publicly available as part of the supplementary material.

PD results vary when they are recomputed — even for the same model; and the random forest fit itself is a random variable with especially high variance on small datasets.

In order to assess whether the results extend to the DGP she would need to quantify the involved uncertainties. Over the course of this paper we propose confidence intervals for partial dependence and feature importance values that take the uncertainties from the estimation of the interpretability method and the model fitting into account. We will return to this example in Section 4, where we show how our approach can help the researcher to evaluate the uncertainty in her estimates.

1.2 Related Work

For models with inherent variance estimators it is possible to construct model-based confidence intervals – for example for Bayesian additive regression trees [21]. Moosbauer et al. [22] introduced a variance estimator for PD which is applicable to all probabilistic models that provide information on posterior (co)variance, such as Gaussian Processes (GPs). Furthermore, various applied articles contain computations of PD confidence bands [2–4, 23–25]. These approaches either quantify only the error due to Monte Carlo approximation or do not account for underestimation of the variance when covering learner variance. This demonstrates the need for a theoretical underpinning of this inferential tool for practical research. For PFI and related approaches, multiple suggestions for confidence intervals and variance estimation are available. Since PFI has first been introduced for random forests [10] (the PFI is also known as Random Forest Feature Importance), several contributions are specific to the random forest PFI [26–28], for which Altmann et al. [29] propose a test for null importance.

Model-agnostic PFI confidence intervals that are similar to ours have been proposed [18, 30, 31]. Our approach additionally corrects for variance underestimation arising from resampling [12] and relate the estimators to the proposed ground truth PFI. An alternative approach for providing bounds on PFI is proposed by Fisher et al. [11] via Rashomon sets, which are sets of models with similar near-optimal prediction accuracy. Our approach differs since we consider bounds for a fixed model or a fixed learning process, while Rashomon sets are defined by a model class and an error bound. Furthermore, alternative approaches of "model-free" inference have been introduced [32–34], which aim to infer properties of the data without an intermediary machine learning model.

2 Methods

In this Section, we present our formal framework: We introduce notation and background on PD and PFI (Section 2.1); formulate PD and PFI as estimators of (proposed) ground truth estimands in the DGP (Section 2.2); apply bias and variance decompositions and separate different sources of uncertainty (Section

2.3); and propose variance estimators and confidence intervals for the model-PD/PFI (which only takes the variance from Monte-Carlo integration into account, see Section 2.4) and the learner-PD/PFI (which also takes learner variance into account, see Section 2.5).

2.1 Background and Notation

We denote the joint distribution induced by the data generating process as $\mathbb{P}_{XY}$, where X is a p-dimensional random variable and Y a 1-dimensional random variable. We consider the case where we aim to describe the true mapping from X to the target Y with $f(X) = E[Y \mid X = x]$.[2] We denote a single random draw from the DGP with $x^{(i)}$ and $y^{(i)}$. A dataset consisting of multiple draws from $\mathbb{P}_{XY}$ will be called $\mathcal{D}_n = \{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$, where n is the number of samples and with each $(x^{(i)}, y^{(i)}) \sim \mathbb{P}_{XY}$, $i \in \{1, \ldots, n\}$.

A machine learning model $\hat{f}$ is a function ($\hat{f} : \mathcal{X} \to \mathcal{Y}$) that maps a vector from the feature space $\mathcal{X} \in \mathbb{R}^p$ to a prediction (e.g. $\mathcal{Y} = \mathbb{R}$ for regression). The model $\hat{f}$ is induced based on a dataset $\mathcal{D}_n$, using a loss function $L : \mathcal{Y} \times \mathbb{R}^p \to \mathbb{R}_0^+$. As the true function f is unknown, the model $\hat{f}$ is interpreted instead of f – for example, with PD plots and PFI. The model $\hat{f}$ is induced by the learner algorithm $I : \mathcal{D} \to \mathcal{H}$ that maps from the space of datasets to the function hypothesis space $\mathcal{H}$. The learning process contains an essential source of randomness, namely the training data as a random sample from $\mathbb{P}_{XY}$. Since the model $\hat{f}$ is induced by the learner fed with data, it can be seen as a realization of a random variable F with distribution $\mathbb{P}_F$. We assume that the model is evaluated with a risk function $\mathcal{R}(\hat{f}) = \mathbb{E}_{XY}[L(Y, \hat{f}(X))] = \int L(y, \hat{f}(x)) d\mathbb{P}_{XY}$, based on a loss function L. To obtain unbiased estimates of the risk, model training and evaluation use different datasets. The dataset $\mathcal{D}_n$ is split into $\mathcal{D}_{n_1}$ for model training and $\mathcal{D}_{n_2}$ for evaluation. The empirical risk is estimated with $\hat{\mathcal{R}}(\hat{f}_{\mathcal{D}_{n_2},\lambda}) := \frac{1}{n_2} \sum_{i=1}^{n_2} L\left(y^{(i)}, \hat{f}_{\mathcal{D}_{n_2},\lambda}(x^{(i)})\right)$.

Let V and W be two random variables, we define a sampler as a function ϕ that maps an input $v \in \mathcal{V}$ to a density function on a space $\mathcal{W}$ i.e. $\phi : \mathcal{V} \to \{\psi \mid \psi \text{ density on } \mathcal{W}\}$. The two most common samplers in the context of PD and PFI are the marginal and the conditional sampler: the marginal sampler ϕ_{marg} maps every input $v \in \mathcal{V}$ to the density of W i.e. for all $v \in \mathcal{V}$: $\phi_{marg}(v) = \psi_W$; the conditional sampler ϕ_{cond} maps every input $v \in \mathcal{V}$ with $\psi_V(v) > 0$ to the conditional density of W i.e. for all $v \in \mathcal{V}$: $\phi_{cond}(v) = \psi_{W|V=v} = \frac{\psi_{W,V=v}}{\psi_{V=v}}$. As such, samples from $\phi_{marg}(v)$ follow $P(W)$, and samples from $\phi_{cond}(v)$ follow $P(W \mid V = v)$.

Simulation vs Real World Scenario

We distinguish between the "simulation" and the "real world" scenario [36]. In the simulation scenario, we can generate a quasi-infinite number of datasets, which allows us to refit the model multiple times using fresh data each time.

[2] This choice for f is motivated by the fact that the conditional expectation is the Bayes-optimal predictor for the L2 loss and for the log-loss optimal predictor in binary classification [35].

In the real world setting, we assume that a single dataset of size n is available. To fit multiple models (of the same class) and to obtain multiple estimates of the risk, resampling techniques such as bootstrapping, cross-validation and repeated subsampling must be used. We denote by B_d the set of indices for the training data in the d-th split repetition and with B_{-d} the corresponding test data indices, where $B_d \cup B_{-d} = \{1, \ldots, n\}$, $b \in \{1, \ldots, m\}$, and m is the number of models trained with different data.

The Role of Samplers

Like all model-agnostic interpretation techniques, both PD and PFI are based on sampling data and evaluating the model on these data [8]. Dependent on how we sample, we obtain different versions of PD and PFI and their results must be interpreted in a different way [11, 18, 37–39]. The two most common theoretical samplers in PD and PFI research are the marginal and the conditional sampler. The choice of the sampler should depend on the modeler's objective and the structure of the data. Under certain conditions, the marginal sampler allows to estimate causal effects [40]. However, for correlated input features the marginal sampler may create unrealistic data outside the training distribution, which is problematic if the goal is to draw inference about the DGP; under such conditions, the conditional sampler may be a better choice [16]. Samplers, especially conditional samplers, are generally not readily available, but must be learned with techniques such as conditional subgroups [38] or conditional density estimators [41–46]. The learning process of the sampler may introduce another source of uncertainty that we do not consider in this work; we discuss this limitation in Section 5.

Difference Between Fixed Model and Random Variable

We distinguish between the interpretation of a single model and the distribution of models produced by a learner. Frequently, a fixed trained model $\hat{f}$ is the subject of interpretation. Any interpretation of a fixed model neglects the variance originating from the learning process. We are often interested in extending the interpretation to the distribution of models produced by a learner. For example, the importance of a feature in a decision tree might be zero because it was never selected for a split. However, if we were to train the tree on a slightly different sample from the same distribution, it might obtain a non-zero importance. A similar distinction between model and learner can be made for performance estimation, where model performance is estimated with a test set, but learner performance requires averaging performance over m repetitions and thus m model refits.

2.1.1 Partial Dependence Plot

The PD of a feature set X_S, $S \subseteq \{1, \ldots, p\}$ (usually $\mid S \mid = 1$) for a given $x \in X_S$, a model $\hat{f}$ and a sampler $\phi : \mathcal{X}_S \to \{\psi \mid \psi \text{ density on } \mathcal{X}_C\}$ is:

$$PD_{S,\hat{f},\phi}(x) := \mathbb{E}_{\tilde{X}_C \sim \phi(x)}[\hat{f}(x, \tilde{X}_C)] = \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \phi(x)(\tilde{x}_c)\hat{f}(x, \tilde{x}_c)\, d\tilde{x}_c, \tag{1}$$

where $\tilde{X}_C$ is a random variable distributed with density $\phi(x)$, and C denote the indices of the remaining features so that $S \cup C = \{1, \ldots, p\}$ and $S \cap C = \emptyset$. To estimate the PD for a specific function $\hat{f}$ using Monte Carlo integration, we draw $r \in \mathbb{N}$ samples for every $x \in \mathcal{X}_S$ from $\phi(x)$ and denote the corresponding dataset by $B_{\phi(x)} = (\tilde{x}_C^{(i,x)})_{i=1,\ldots,r}$. The estimation is given by:

$$\widehat{PD}_{S,\hat{f},\phi}(x) = \frac{1}{r} \sum_{i=1}^{r} \hat{f}(x, \tilde{x}_C^{(i,x)}). \tag{2}$$

By partial dependence plot (PDP) we denote the graph that visualizes the PDP. The PDP consists of a line connecting the points $\{(x^{(g)}, \widehat{PD}_{S,\hat{f},\phi}(x^{(g)})\}_{g=1}^{G}$, with G grid points that are usually equidistant or quantiles of $\mathbb{P}_{X_S}$. See Figure 1 for an example of a PDP.
For the marginal sampler, the PDP of a model $\hat{f}$ visualizes the expected effect of a feature after marginalizing out the effects of all other features [9]; for the marginal version we do not necessarily need a sampler, since we can just use the training and test data as a sample from X_C for every $x \in \mathcal{X}_S$. For the conditional sampler, the PDP is also called M-plot and visualizes the expected effect of a feature on the prediction, taking into account its associational dependencies with all other features [9, 39].

2.1.2 Permutation Feature Importance

The PFI of a feature set X_S (usually just one feature) for a model $\hat{f}$ and a sampler $\phi : \mathcal{X}_C \to \{\psi \mid \psi \text{ density on } \mathcal{X}_S\}$ is defined by:

$$PFI_{S,\hat{f},\phi} := \mathbb{E}_{X_C,Y}[\mathbb{E}_{\tilde{X}_S \sim \phi(X_C)}[L(Y, \hat{f}(\tilde{X}_S, X_C))]] - \mathbb{E}_{XY}[L(Y, \hat{f}(X))], \tag{3}$$

where $\tilde{X}_S$ is a random variable distributed with density $\phi(X_C)$, $\tilde{X}_S \perp\!\!\!\perp Y \mid X_C$ and X_C are the remaining features so that $S \cup C = \{1, \ldots, p\}$ and $S \cap C = \emptyset$. To estimate the PFI for a specific function $\hat{f}$ and a sampler ϕ using Monte Carlo integration, we draw $r \in \mathbb{N}$ samples for every datapoint $x_C^{(i)} \in \mathcal{X}_C$ ($x_C^{(i)}$ describes the feature values in C of the i-th instance in the evaluation[3] dataset D_{n_2}) from $\phi(x_C^{(i)})$ and denote the corresponding datasets by $B_{\phi(x_C^{(i)})} =$

[3]The estimation of $\widehat{PFI}$ requires unseen data, so that the loss estimates deliver unbiased results [47, 48].

$(\tilde{x}_S^{(k,i)})_{k=1,\ldots,r}$. The estimation is given by:

$$\widehat{PFI}_{S,\hat{f},\phi} = \frac{1}{n_2} \sum_{i=1}^{n_2} \left(\frac{1}{r} \sum_{k=1}^{r} L(y^{(i)}, \hat{f}(\tilde{x}_S^{(k,i)}, x_C^{(i)})) - L(y^{(i)}, \hat{f}(x^{(i)})) \right). \quad (4)$$

We assume that the loss used for PFI can be computed per instance, which excludes losses such as the area under the receiver operating characteristic curve (AUC). See Figure 1 for a PFI example.
For the marginal sampler, the PFI of a model $\hat{f}$ describes the change in loss if the feature values in X_S are randomly sampled from X_S i.e. the possible dependence to X_C and Y is broken (extrapolation); for the marginal version, we do not even need a sampler because we can simply take the permuted feature values in X_S in the evaluation dataset D_{n_2} [10, 11]. For the conditional sampler, PFI is also called the conditional PFI and may be interpreted as the *additional* importance of a feature *given that we already know the other feature values* [18, 38, 49, 50].

Indices

To avoid indices overhead and because PDP/PFI and their respective estimations are always relative to a fixed feature set S and sampler ϕ, we will abbreviate $PD_{S,\hat{f},\phi}, \widehat{PD}_{S,\hat{f},\phi}, PFI_{S,\hat{f},\phi}, \widehat{PFI}_{S,\hat{f},\phi}$ with $PD_{\hat{f}}, \widehat{PD}_{\hat{f}}, PFI_{\hat{f}}, \widehat{PFI}_{\hat{f}}$ respectively.

2.2 Relating the Model to the Data Generating Process

The goal of statistical inference is to gain knowledge about the DGP. Therefore, the modeler aims to establish relationships between properties of the model and the DGP. For example, under certain assumptions, the coefficients of a generalized linear model (i.e. model properties) can be related to parameters of the respective conditional distribution defined by the DGP, such as conditional mean and covariance structure (i.e. DGP properties). Machine learning models such as random forests or neural networks lack such a mapping between learned model parameters and properties of the DGP. This lack of counterparts in the DGP make it difficult to interpret complex machine learning models and to draw conclusions about the real world. Interpretation methods such as PD and PFI provide **external descriptors** of how features affect the model predictions. However, PD and PFI are estimators that lack a counterpart estimand in the DGP. We propose an inference approach for these external descriptors. We define a ground truth version of PD and PFI through the DGP, namely the DGP-PD and the DGP-PFI. The DGP-PD and the DGP-PFI are defined as the PD and PFI, but applied to the true function f instead of the trained model $\hat{f}$. Consequently, the DGP-PD becomes the feature effect of features X_S on the underlying function f:

Definition 1 (DGP-PD) The DGP-PD is the PD applied to function $f : \mathcal{X} \mapsto \mathcal{Y}$ of the DGP with sampler $\phi : \mathcal{X}_S \to \{\psi \mid \psi \text{ density on } \mathcal{X}_C\}$.

$$\text{DGP-PD}(x) := PD_f(x)$$

Definition 2 (DGP-PFI) The DGP-PFI is the PFI applied to function $f : \mathcal{X} \mapsto \mathcal{Y}$ of the DGP with sampler $\phi : \mathcal{X}_C \to \{\psi \mid \psi \text{ density on } \mathcal{X}_S\}$.

$$\text{DGP-PFI} := PFI_f$$

Note that the DGP-PD and DGP-PFI may not be well-defined for all possible samplers. The DGP $f(x) = \mathbb{E}[Y \mid X = x]$ is undefined for $x \in \mathcal{X}$ with zero density ($\psi_X(x) = 0$). For the marginal sampler, for instance, DGP-PD and DGP-PFI might not be defined if the input features show strong correlations [49]. Conditional samplers, on the other side, do not face this threat as they preserve dependencies between features and therefore do not create unrealistic inputs [11, 18, 38, 39].[4] However, under certain conditions, it can still be useful to also use other samplers than the conditional samplers to gain insight into the DGP. For example, under certain conditions, the marginal PDP allows to estimate causal effects [40] or recover relevant properties of linear DGPs [51].

Clearly, the function f is unknown in most applications, which makes it impossible to know the DGP-PD and DGP-PFI for these cases. However, Definitions 1 and 2 enable, at least in theory, to compare the PD/PFI of a model with the PD/PFI of the DGP **in simulation studies** and to research statistical biases. More importantly, the ground truth definitions of DGP-PD and DGP-PFI allow us to treat PD and PFI as statistical estimators of properties of the DGP.

In this work, we study PD and PFI as statistical estimators of the ground truth DPG-PD and DGP-PFI – including bias and variance decompositions – as well as confidence interval estimators. DGP-PD and DGP-PFI describe interesting properties of the DGP concerning the associational dependencies between the predictors and the target [16]; however, practitioners must decide whether these properties are relevant to answer their question or if different tools of model-analysis provide more interesting estimands.

2.3 Bias-Variance Decomposition

The definition of DGP-PD and DGP-PFI gives us a ground truth to which the PD and PFI of a model can be compared – at least in theory and simulation. The error of the estimation (mean squared error between estimator and estimand) can be decomposed into the systematic deviation from the true estimand (statistical bias) and the learner variance. PD and PFI are both expectations over the (usually unknown) joint distribution of the data. The expectations are therefore typically estimated from data using Monte Carlo

[4] To illustrate the idea of unrealistic data points, think of two strongly correlated features such as the weight and height of a person. Not every combination of feature values is possible – a person with a weight of 4kg and a height of 2m is from a biological perspective inconceivable.

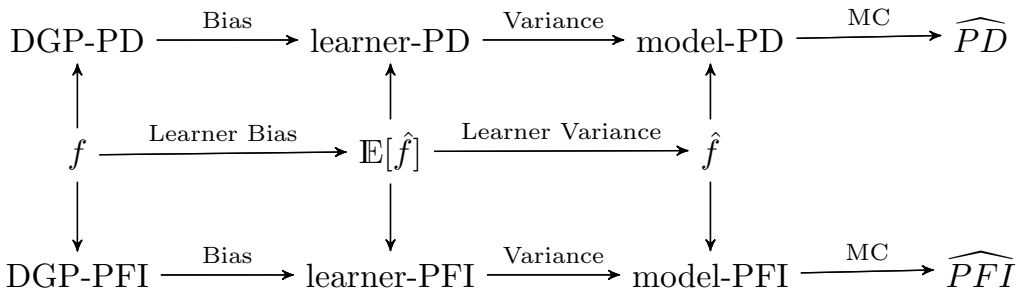


Figure 2 A model $\hat{f}$ deviates from f due to learner bias and variance. Similarly, $\widehat{PD}$ and $\widehat{PFI}$ estimates deviate from their ground truth versions DGP-PD and DGP-PFI due to bias, variance, and Monte Carlo integration (MC).

integration, which adds another source of variation to the PFI and PD estimates. Figure 2 visualizes the chain of errors that stand between the estimand (DGP-PD, DGP-PFI) and the estimates ($\widehat{PD}$, $\widehat{PFI}$).

For the PD, we compare the mean squared error (MSE) between the true DGP-PD (PD_f as defined in Equation 1) with the theoretical PD of a model instance $\hat{f}$ ($PD_{\hat{f}}$) at position x.

$$\mathbb{E}_F[(PD_f(x) - PD_{\hat{f}}(x))^2] = \underbrace{(PD_f(x) - \mathbb{E}_F[PD_{\hat{f}}(x)])^2}_{Bias^2} + \underbrace{\mathbb{V}_F[PD_{\hat{f}}(x)]}_{Variance}$$

Here, F is the distribution of the trained models, which can be treated as a random variable. The bias-variance decomposition of the MSE of estimators is a well-known result [52]. For completeness, we provide a proof in Appendix A. Figure 3 visualizes bias and variance of a PD curve, and the variance due to Monte Carlo integration.

Similarly, the MSE of the theoretical PFI of a model (Equation 3) can be decomposed into squared bias and variance. The proof can be found in Appendix B.

$$\mathbb{E}_F[(PFI_{\hat{f}} - PFI_f)^2] = Bias_F^2[PFI_{\hat{f}}] + \mathbb{V}_F[PFI_{\hat{f}}]$$

The learner variance of PD/PFI stems from variance in the model fit, which depends on the training sample. When constructing confidence intervals, we must take into account the variance of PFI and PDP across model fits, and not just the error due to Monte Carlo integration. As we show in an application (Section 4), whether PD and PFI are based on a single model or are averaged across model refits can impact both the interpretation and especially the certainty of the interpretation. We therefore distinguish between model-PD/PFI and learner-PD/PFI, which are averaged over refitted models. Variance estimators for model-PD/PFI only account for variance due to Monte Carlo integration.

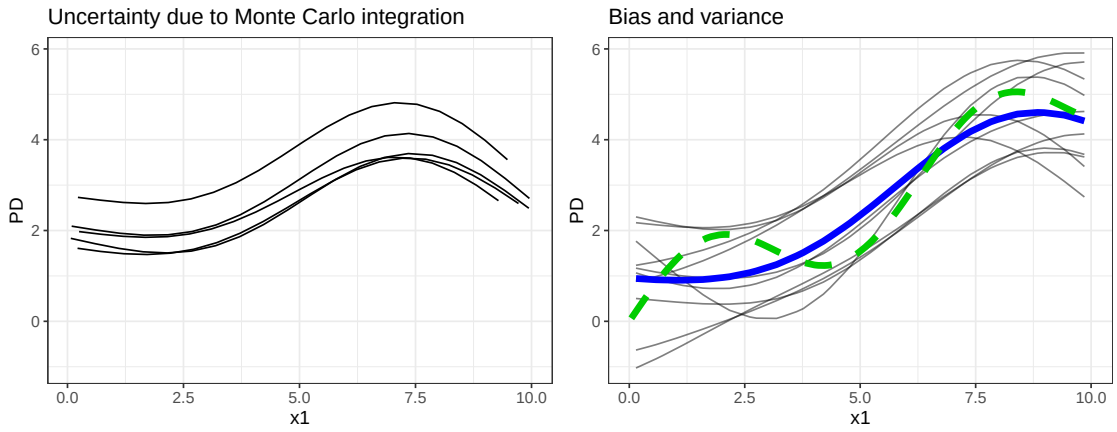


Figure 3 Illustration of bias, variance and Monte Carlo approximation for the PD with marginal sampling. Left: Various PDPs using different data for the Monte Carlo integration, but keeping the model fixed. Right: The green dashed line shows the DGP-PDP of a toy example. Each thin line is the PDP for the model fitted with a different sample, and the thick blue line is the average thereof. Deviations of the DGP-PDP from the expected PDP are due to bias. Deviations of the individual model-PDPs from the expected PDP are due to learner variance.

2.4 Model-PD and Model-PFI

Here, we study the model-PD and the model-PFI, and provide variance and confidence interval estimators. With the terms model-PD and model-PFI, we refer to the original proposals for PD [9] and PFI [10, 11] for fixed models. Conditioning on a given model $\hat{f}$ ignores the learner variance due to the learning process. Only the variance due to Monte Carlo integration can be considered in this case.

The model-PD estimator (Equation (2)) is unbiased regarding the theoretical model-PD (Equation (1)). Similarly, the estimated model-PFI (Equation 4) is unbiased with respect to the theoretical model-PFI (Equation 3). These findings rely on general properties of Monte Carlo integration, which state that Monte Carlo integration converges to the integral due to the law of large numbers. Proofs can be found in Appendix C and E. Moreover, under certain conditions, model-PD and model-PFI are unbiased estimators of the DGP-PD (Theorem 1) and DGP-PFI (Theorem 2), respectively.

To quantify the variance due to Monte Carlo integration and to construct confidence intervals, we calculate the variance across the sample. For the model-PD, the variance can be estimated with:

$$\widehat{\mathbb{V}}(\widehat{PD}_{\hat{f}}(x)) = \frac{1}{r(r-1)} \sum_{i=1}^{r} \left(\hat{f}(x, \tilde{x}_C^{(i,x)}) - \widehat{PD}_{\hat{f}}(x) \right)^2. \tag{5}$$

Similarly for the model-PFI, the variance can be estimated with:

$$\widehat{\mathbb{V}}(\widehat{PFI}_{\hat{f}}) = \frac{1}{n_2(n_2-1)} \sum_{i=1}^{n_2} \left(L^{(i)} - \widehat{PFI}_{\hat{f}} \right)^2,$$

where $L^{(i)} = \frac{1}{r}\sum_{k=1}^{r} L(y^{(i)}, \hat{f}(\tilde{x}_S^{(k,i)}, x_C^{(i)})) - L(y^{(i)}, \hat{f}(x^{(i)}))$.

The model-PD and model-PFI are mean estimates of independent samples with estimated variance. As such, they can be modelled approximately with a t-distribution with $r-1$ and n_2-1 degrees of freedom, respectively. This allows us to construct point-wise confidence bands for the model-PD and confidence intervals for the model-PFI that capture the Monte Carlo integration uncertainty. We define point-wise $1-\alpha$-confidence bands around the estimated model-PD:

$$CI_{\widehat{PD}_{\hat{f}}(x)} = \left[\widehat{PD}_{\hat{f}}(x) \pm t_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathbb{V}}(\widehat{PD}_{\hat{f}}(x))}\right]. \tag{6}$$

where $t_{1-\frac{\alpha}{2}}$ is the $1-\alpha/2$ quantile of the t-distribution with $r-1$ degrees of freedom. We proceed in the same manner for PFI but with n_2-1 degrees of freedom:

$$CI_{\widehat{PFI}_{\hat{f}}} = \left[\widehat{PFI}_{\hat{f}} \pm t_{1-\frac{\alpha}{2}} \sqrt{\widehat{\mathbb{V}}(\widehat{PFI}_{\hat{f}})}\right]. \tag{7}$$

Confidence intervals for model-PD and model-PFI ignore the learner variance. Therefore, the interpretation is limited to variance regarding the Monte Carlo integration, and we cannot generalize results to the DGP. The model-PD/PFI and their confidence bands/intervals are applicable when the focus is a fixed model.

2.5 Learner-PD and Learner-PFI

To account for the learner variance, we propose the learner-PD and the learner-PFI, which average the PD/PFI over m model fits $\hat{f}_d$ with $d \in \{1, \ldots, m\}$. The models are produced by the same learning algorithm, but trained on different data samples, denoted by training sample indices B_d and the remaining test data B_{-d} so that $B_d \cap B_{-d} = \emptyset$ and $B_d \cup B_{-d} = \mathcal{D}_n$. The learner-variants are averages of the model-variants, where for each model-PD/PFI, the model is repeatedly "sampled" from the distribution of models F.

The learner-PD is therefore the expected PD over the distribution of models generated by the learning process, i.e. $\mathbb{E}_F[PD_{\hat{f}}(x)]$. We estimate the learner-PD with:

$$\overline{\widehat{PD}}(x) = \frac{1}{m}\sum_{d=1}^{m}\frac{1}{r}\sum_{i=1}^{r}\hat{f}_d\left(x, x_C^{i,x,d}\right), \tag{8}$$

where $\hat{f}_d$ is trained on sample indices B_d and the PD estimated with data $B_{\phi(x),d}$ using a sampler ϕ m-times.

Following the PD, the learner-PFI is the expected PFI over the distribution of models produced by the learner: $\mathbb{E}_F[PFI_{\hat{f},\phi}]$. We propose the following

estimator for the learner-PFI:

$$\overline{\widehat{PFI}} = \frac{1}{m}\sum_{d=1}^{m}\frac{1}{n_2}\sum_{i=1}^{n_2}\left(\bar{\tilde{L}}_d^{(i)} - L_d^{(i)}\right), \tag{9}$$

where losses $L_d^{(i)} = L(y^{(i)}, \hat{f}_d(x^{(i)}))$ and $\bar{\tilde{L}}_d^{(i)} = \frac{1}{r}\sum_{k=1}^{r} L(y^{(i)}, \hat{f}_d(\tilde{x}_S^{(k,i,d)}, x_C^{(i)}))$ are estimated with data B_{-d} and m-times sampled data $B_{\phi(x),d}$ for a model trained on data B_d. A similar estimator has been proposed by Janitza et al. [28] for random forests.

2.5.1 Bias of the Learner-PD

The learner-PD is an unbiased estimator of the expected PD over the distribution of models F, since

$$\mathbb{E}_F[\overline{\widehat{PD}}(x)] = \mathbb{E}_F\left[\frac{1}{m}\sum_{d=1}^{m}\widehat{PD}_{\hat{f}_d}(x)\right] = \frac{m}{m}\mathbb{E}_F[PD_{\hat{f}_d}(x)] = \mathbb{E}_F[PD_{\hat{f}_d}(x)].$$

The bias of the learner-PD *regarding the DGP-PD* is linked to the bias of the learner. If the learner is unbiased, the PDs are unbiased as well.

Theorem 1 *Learner unbiasedness implies PD unbiasedness:*
$\mathbb{E}_F[\hat{f}(x)] = f(x) \implies \mathbb{E}_F[PD_{\hat{f}}(x)] = PD_f(x)$

Proof Sketch 1 Applying Fubini's Theorem allows us to switch the order of integrals. Further replacing $\mathbb{E}_F[\hat{f}(x)]$ with f proves the unbiasedness. A full proof can be found in Appendix D.

By learner bias, we refer to the expected deviation between the estimated $\hat{f}$ and the true function f. Particularly interesting in this context is the inductive bias (i.e. the preference of one generalization over another) that is needed for learning ML models that generalize [53]. A wrong choice of inductive bias, such as searching models $\hat{f}$ in a linear hypotheses class when f is non-linear, leads to deviations of the expected $\hat{f}$ from f. But there are also other reasons why a bias of $\hat{f}$ from f may occur, for example if using an insufficiently large sample of training data. We discuss the critical assumption of learner unbiasedness further in Section 5.

2.5.2 Bias of the Learner-PFI

The learner-PFI is unbiased regarding the expected learner-PFI over the distribution of models F, since the learner-PFI is a simple mean estimate. However, unlike the learner-PD, learner unbiasedness does not generally imply unbiasedness of the learner-PFI *regarding the DGP-PFI*. This is generally only the case, if we use the conditional sampler.

Theorem 2 *If the learner is unbiased with* $\mathbb{E}_F[\hat{f}] = f$ *and the L2-loss is used, then the conditional model-PFI and conditional learner-PFI are unbiased estimators of the conditional DGP-PFI.*

Proof Sketch 2 Both L and $\tilde{L}$ can be decomposed into bias, variance, and irreducible error. Due to the subtraction, the irreducible error vanishes, and the differences of biases and variances remain. Model unbiasedness sets the bias terms to zero and variance becomes zero due to conditional sampling. The extended proof can be found in Appendix F.

Intuitively, the model-PFI and learner-PFI should tend to have a negative bias and therefore underestimate the DGP-PFI. A model cannot use more information about the target than what is encoded in the DGP. However, as Theorem 3 shows, under specific conditions, the PFI using conditional sampling can be larger than the DGP-PFI.

Theorem 3 *The difference between the conditional model-PFI and the conditional DGP-PFI is given by:*

$$PFI_f - PFI_{\hat{f}} = 2\mathbb{E}_{X_C}\left[\mathbb{V}_{X_S|X_C}[f] - Cov_{X_S|X_C}[f, \hat{f}]\right].$$

Proof Sketch 3 For the L2 loss, the expected loss of a model $\hat{f}$ can be decomposed into the expected loss between $\hat{f}$ and f and the expected variance of Y given X. Due to the subtraction, the latter term vanishes. The remainder can be simplified using that $Y \perp\!\!\!\perp \tilde{X_S} \mid X_C$ and $P(\tilde{X_S}, X_C) = P(X_S, X_C)$ due to the conditonal sampling. The extended proof can be found in Appendix G.

However, for an overestimation of the conditional PFI to occur, the expected conditional variance of $\hat{f}$ must be greater than the one of f. Moreover, $\hat{f}$ and f must have a large expected conditional covariance, meaning that $\hat{f}$ has learned something about f.

2.5.3 Variance Estimation

The learner-PD and learner-PFI vary not only due to learner variance (refitted models), but also due to using different samples each time for the Monte Carlo integration. Therefore, their variance estimates capture the entire modeling process. Consequently, learner-PD/PFI along with their variance estimators bring us closer to the DGP-PD/PFI, and only the systematic bias remains unknown.

We can estimate this point-wise variance of the learner-PD with:

$$\widehat{\mathbb{V}}(\overline{\widehat{PD}}(x)) = \left(\frac{1}{m} + c\right) \cdot \frac{1}{(m-1)} \sum_{d=1}^{m} (\widehat{PD}_{\hat{f}_d}(x) - \overline{\widehat{PD}}(x))^2$$

And equivalently for the learner-PFI:

$$\hat{\mathbb{V}}(\overline{\widehat{PFI}}) = \left(\frac{1}{m} + c\right) \cdot \frac{1}{(m-1)} \sum_{d=1}^{m} (\widehat{PFI}_{\hat{f}_d} - \overline{\widehat{PFI}})^2$$

The correction term c depends on the data setting. In simulation settings that allow us to draw new training and test sets for each model, we can use $c = 0$, yielding the standard variance estimators. In real world settings, we usually have a fixed dataset of size n, and models are refitted using resampling techniques. Consequently, data are shared by model refits, and variance estimators will underestimate the true variance [12]. To correct the variance estimate of the generalization error for bootstrapped or subsampled models, Nadeau and Bengio [12] suggested the correction term $c = \frac{n_2}{n_1}$ (where n_2 and n_1 are sizes of test and training data). However, the correction remains a rough correction, relying on the strongly simplifying assumption that the correlation between model refits depends only on the number of shared observations in the respective training datasets and not on the specific observations that they share. While this assumption is usually wrong, we show in Section 3.1 that the correction term offers a vast improvement for variance estimation – compared to using no correction.

2.5.4 Confidence Bands and Intervals

Since the learner-PD and learner-PFI are means with estimated variance, we can use the t-distribution with $m - 1$ degrees of freedom to construct confidence bands/intervals, where m is the number of model fits. The point-wise confidence band for the learner-PD is:

$$CI_{\overline{\widehat{PD}}(x)} = \left[\overline{\widehat{PD}}(x) \pm t_{1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\overline{\widehat{PD}}(x))}\right],$$

where $t_{1-\frac{\alpha}{2}}$ is the respective $1 - \alpha/2$ quantile of the t-distribution with $m - 1$ degrees of freedom. Equivalently, we propose a confidence interval for the learner-PFI:

$$CI_{\overline{\widehat{PFI}}} = \left[\overline{\widehat{PFI}} \pm t_{1-\frac{\alpha}{2}} \sqrt{\hat{\mathbb{V}}(\overline{\widehat{PFI}})}\right].$$

Taking the learner variance into account can affect the interpretation, as we show in the application in Section 4. An additional advantage of the learner-PD and learner-PFI is that they make better use of the data, since a larger share of the data is employed as test data compared to only using a small holdout set.

Table 1 Coverage Probability of the 95% PDP Confidence Bands. boot = bootstrap, subs = subsampling, * = with adjustment.

dgp	model	n	boot	boot*	subs	subs*	ideal
linear	lm	100	0.41	0.89	0.34	0.82	0.95
linear	lm	1000	0.41	0.89	0.33	0.80	0.95
linear	rf	100	0.39	0.86	0.36	0.83	0.95
linear	rf	1000	0.38	0.87	0.35	0.83	0.95
linear	tree	100	0.54	0.96	0.47	0.92	0.95
linear	tree	1000	0.57	0.96	0.48	0.91	0.95
non-linear	lm	100	0.43	0.90	0.36	0.84	0.95
non-linear	lm	1000	0.41	0.89	0.33	0.81	0.95
non-linear	rf	100	0.39	0.87	0.36	0.84	0.95
non-linear	rf	1000	0.38	0.86	0.36	0.83	0.95
non-linear	tree	100	0.58	0.98	0.51	0.95	0.95
non-linear	tree	1000	0.59	0.97	0.51	0.94	0.95

3 Simulation Studies

In this Section, we study the coverage of the confidence intervals for the learner-PD/PFI on simulated examples (Section 3.1) and compare our proposed refitting-based variance estimation with model-based variance estimators (Section 3.2).

3.1 Confidence Interval Coverage Simulation

In simulations, we compared confidence interval performance between bootstrapping and subsampling, with and without variance correction. We simulated two DGPs: a *linear* DGP was defined as $y = f(x) = x_1 - x_2 + \epsilon$ and a *non-linear* DGP as $y = f(x) = x_1 - \sqrt{1 - x_2} + x_3 \cdot x_4 + (x_4/10)^2 + \epsilon$. All features were uniformly sampled from the unit interval $[0;\ 1]$, and for both DGPs, we set $\epsilon \sim N(0, 1)$. We studied the two settings "simulation" and "real world" as described in Section 2.1. In both settings, we trained linear models (lm), regression trees (tree) and random forests (rf) each 15 times, and computed confidence intervals for the learner-PD and learner-PFI across the 15 refitted models. In the "simulation" setting, we sampled $n \in \{100, 1000\}$ fresh data points for each model refit, where 63.2% of the data were used for training and the remaining 36.8% for PDP and PFI estimation.[5]

In the "real world" setting, we sampled $n \in \{100, 1000\}$ data points **once** per experiment, and generated 15 training data sets using a bootstrap (sample size n with replacement, which yields $0.632 \cdot n$ unique data points in expectation) or subsampling (sample size $0.632 \cdot n$ without replacement). In both settings, the learner-PD and learner-PFI as well as their respective confidence intervals were computed over the 15 retrained models. We repeated the experiment 10,000 times and counted how often the estimated confidence intervals

[5]We choose this training size (63.2%) to match the expected number of unique samples when using bootstrapping, which allows to compare bootstrapping and subsampling.

Table 2 Coverage Probability of the 95% PFI Confidence Intervals. boot = bootstrap, subs = subsampling, * = with adjustment.

dgp	model	n	boot	boot*	subs	subs*	ideal
linear	lm	100	0.27	0.70	0.23	0.63	0.94
linear	lm	1000	0.25	0.68	0.21	0.60	0.95
linear	rf	100	0.44	0.92	0.39	0.88	0.95
linear	rf	1000	0.42	0.90	0.38	0.86	0.95
linear	tree	100	0.52	0.97	0.42	0.90	0.95
linear	tree	1000	0.42	0.90	0.34	0.81	0.95
non-linear	lm	100	0.31	0.81	0.25	0.72	0.94
non-linear	lm	1000	0.25	0.67	0.21	0.59	0.95
non-linear	rf	100	0.47	0.94	0.43	0.91	0.95
non-linear	rf	1000	0.41	0.89	0.38	0.86	0.95
non-linear	tree	100	0.68	0.99	0.56	0.96	0.94
non-linear	tree	1000	0.58	0.97	0.46	0.92	0.95

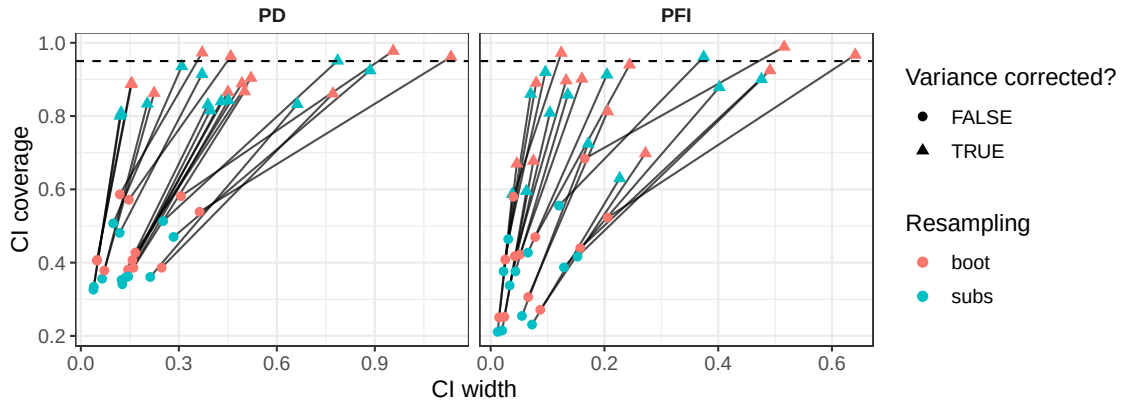


Figure 4 Confidence interval width vs. coverage for bootstrapping (boot) and subsampling (subs), comparing before and after correction. Segments connect identical scenarios.

covered the expected PD or PFI ($\mathbb{E}_F[PD_{\hat{f}}]$ and $\mathbb{E}_F[PFI_{\hat{f}}]$) over the distribution of models F.[6] These expected values were computed using 10,000 separate runs. The coverage estimates were averaged across features per scenario and for PD also across grid points ($\{0.1, 0.3, 0.5, 0.7, 0.9\}$) for all features.

Table 2 and Table 1 show that in the "simulation" setting ("ideal"), we can recover confidence intervals using the standard variance estimation with the desired coverage probability. However, in the "real world" setting, the confidence intervals for both the learner-PD and learner-PFI are too narrow across all scenarios and both resampling strategies when the intervals are based on naive variance estimates. Some coverage probabilities are especially low, such as for linear models with $30\% - 40\%$.

The coverage probabilities drastically improve when the correction term is used (see Figure 4). However, in the simulated scenarios, these probabilities are still somewhat too narrow. For the linear model, the confidence intervals

[6] The coverage does not refer to the DGP-PD/PFI, but rather to the expected learner-PD/PFI, as we studied the choices of resampling and correction for the learner variance.

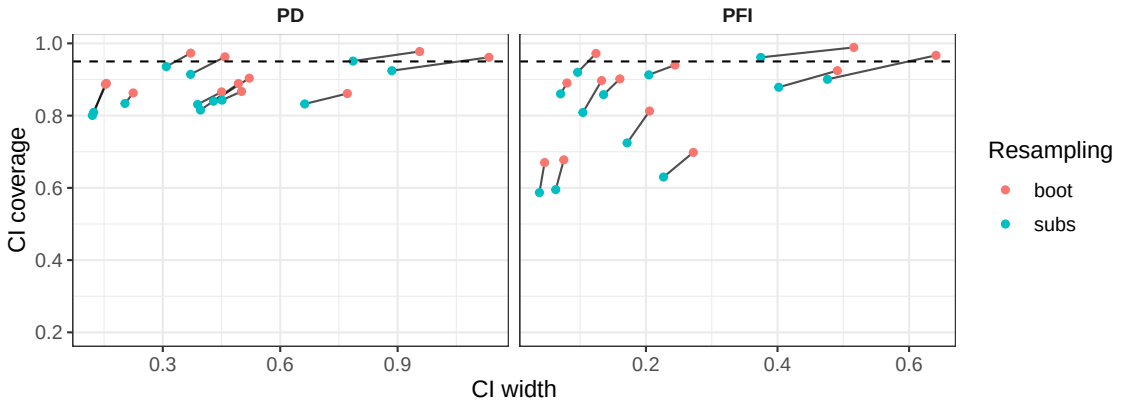


Figure 5 Confidence interval width vs. coverage for bootstrapping (boot) and subsampling (subs), both with correction. Segments connect identical scenarios.

were the narrowest, with coverage probabilities of around $80\% - 90\%$ for PD and $60\% - 80\%$ for PFI across DGPs and sample sizes. The PD confidence bands were not heavily affected by increasing sample size n, but the PFI estimates became slightly narrower in most cases. In the case of decision trees, the adjusted confidence intervals were sometimes too large, especially for the adjusted bootstrap.

Except for trees on the *non-linear* DGP, the bootstrap outperformed subsampling in terms of coverage, i.e. the coverage was closer to the 95% level and rather erred on the side of "caution" with wider confidence intervals (see Figure 5). As recommended by Nadeau and Bengio [12], we used 15 refits. We additionally analyzed how the coverage and interval width changed by increasing refits from 2 to 30 and noticed that the coverage worsened with more refits while the width of the confidence intervals decreased. Increasing the number of refits incurs an inherent trade-off between interval width and coverage: The more refits are considered, the more accurate the learner-PFI and learner-PD become, and also the more certain the variance estimates become, scaling with $1/m$. However, there is a limit to the information in the data, such that additional refits falsely reduce the variance estimate and the confidence intervals become too narrow. To refit the model 10 - 20 times seemed to be an acceptable trade-off between coverage and interval width, as demonstrated in Figure 6. Below $\sim$ 10 refits, the confidence intervals were large and the mean PD/PFI estimates have a high variance. Above $\sim$ 20 refits, the widths no longer decreased substantially. The figures for the other scenarios can be found in Appendix H. With our simulation results, we could show that employing confidence intervals using the naive variance estimation (without correction) results in considerably too narrow intervals. While the simple correction term by Nadeau and Bengio [12] does not always provide the desired coverage probability, it is a vast improvement over the naive approach. We therefore recommend using the correction when computing confidence intervals for learner-PD and learner-PFI, as this is currently the best approach available. We also recommend refitting the model approximately 15 times. For more "cautious" confidence intervals,

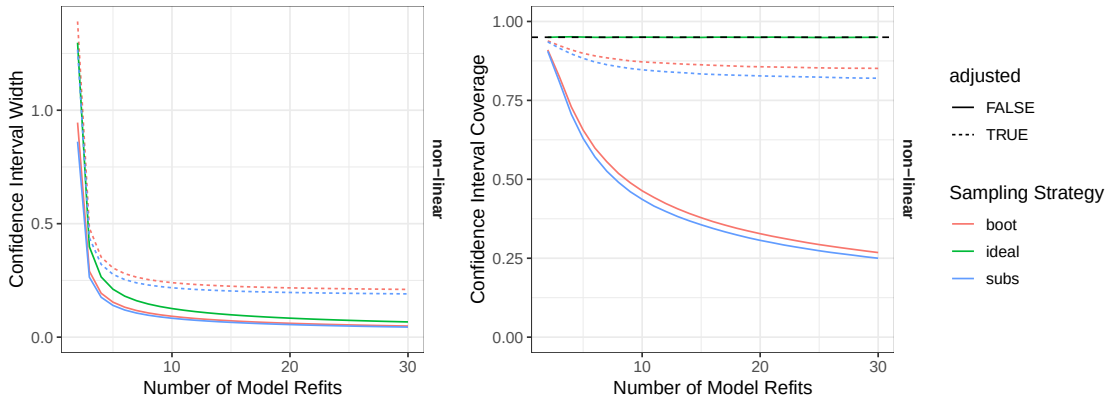


Figure 6 Average PD confidence band width (left) and coverage (right) as a function of the number of refitted models for the random forest on the *non-linear* DGP.

we recommend using confidence intervals based on resampling with replacement (bootstrap) over sampling without replacement (subsampling). However, besides wider confidence intervals, the bootstrap also requires additional attention when model-tuning with internal resampling is used; otherwise, data points may inadvertently be used in both training and validation datasets.

3.2 Comparison to Model-based Approaches

While our methods based on model-refits provide confidence intervals for PD and PFI in a model-agnostic manner, it is also possible to exploit (co)variance estimates of probabilistic models to construct confidence intervals. Here, we will, for the case of PD[7], compare our approach with the model-based approach of Moosbauer et al. [22] applied to a Gaussian Process (GP) and a linear model (LM). Below, we summarize the key theoretical concepts of the model-based approach and then investigate the differences to our approach based on two simulation settings. We find that our approach more reliably delivers better coverages that are closer to the $1-\alpha$ confidence level; this can be explained by the fact that the model-based approach ignores the variance in Monte Carlo integration.

Theoretical background

Moosbauer et al. [22] leverage the kernel of GPs to analytically calculate the model-based uncertainty contained in the PD function. Let $\hat{f}$ be a GP and $\hat{\boldsymbol{m}}(x) = \left(\hat{m}(x, x_C^{(i)})\right)_{i=1,\ldots,n_2}$ its estimated posterior mean and $\hat{\boldsymbol{K}}(x) = \left(\hat{k}\left((x, x_C^{(i)}), (x, x_C^{(j)})\right)\right)_{i,j=1,\ldots,n_2}$ its estimated posterior covariance on the test set D_{n_2} for fixed feature values $x \in X_S$. The PD estimate $\widehat{PD}$ of $\hat{f}$ can be seen as a random variable. Thus, the PD for the posterior mean function is given

[7] We do not know of any application of Moosbauer et al.'s [22] approach to PFI of probabilistic models.

by the expected value of $\widehat{PD}$:

$$\mathbb{E}_{\hat{f}}\left[\widehat{PD}(x)\right] = \mathbb{E}_{\hat{f}}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\hat{f}(x, x_C^{(i)})\right] = \frac{1}{n_2}\sum_{i=1}^{n_2}\hat{m}(x, x_C^{(i)}). \tag{10}$$

The variance of the PD is estimated accordingly and can be calculated straightforwardly by leveraging the posterior covariance of the GP:

$$\mathbb{V}_{\hat{f}}\left[\widehat{PD}(x)\right] = \mathbb{V}_{\hat{f}}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}\hat{f}(x, x_C^{(i)})\right] = \frac{1}{{n_2}^2}\mathbf{1}^\top \hat{\boldsymbol{K}}(x)\mathbf{1}. \tag{11}$$

Since the n_2 predictors $\hat{f}(x, x_C^{(i)})$ of the GP follow a Gaussian distribution, their sum is also normally distributed. Hence, we can construct confidence bands for the mean estimate in Eq. (10) by using the variance estimate in Eq. (11) together with the respective $1 - \alpha/2$ quantiles of the Gaussian distribution. This approach is applicable to any models (including non-GPs) that provide a fully specified covariance matrix between the predictions.

As Eq. (11) solely quantifies the variance w.r.t. the model given the observed data, the resulting confidence bands only capture model variance but not the variance induced by MC integration.

Simulation

We compare our approach for variance estimation to the model-based approach on the following two settings:

$$\text{DGP 1:} \quad Y = 4X_1 - 2X_2 + 2X_3 - X_4 + X_5 + \epsilon$$

$$\text{DGP 2:} \quad Y = 2sin(2\pi X_1) + cos(2\pi X_2) + exp(0.5X_3) - 2X_4^2 + \sqrt{X_5} + \epsilon$$

with $X_j \overset{i.i.d.}{\sim} U(0,1)$ for all $j \in \{1, ..., 5\}$. Given a DGP of the form $y = f(x) + \epsilon$ the distribution of ϵ is set to $\epsilon \sim N(0, (0.2\,\sigma(f(x)))^2)$.

While we calculate the DGP-PD for DGP 1 analytically, we approximate it for DGP 2 by averaging over PDs generated by 10,000 independent draws based on a linear model with correctly specified components.[8] The experiments are performed 1000 times for $n = 200$ and $n = 1000$, where a random sample of $n_1 = 0.632 \cdot n$ is used to fit the models and the remaining $n_2 = 0.368 \cdot n$ observations are used to calculate the PD. Since both DGPs (if features in DGP 2 are transformed accordingly) can be expressed by a linear model and since model-based variance estimates for linear models can be derived analytically based on the variance of their coefficients, we additionally compare these estimates to our resampling-based approach (i.e. the learner-PD) for a correctly specified linear model. While the model-based variance estimates can be calculated by one model fit per repetition, we draw 15 subsampled

[8] "Correctly specified components" means here that the features X_1 to X_5 are transformed according to the associated non-linear functions from DGP 2 before the LM is fitted.

data sets per repetition to compute the variance estimate for the resampling-based approach based on a marginal sampler (since we assume uncorrelated features in all scenarios).[9] We choose the grid points $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ and a confidence level of 0.95 to evaluate the mean and variance estimates of the PDs.

Table 3 shows the results for both the model-based (mod) and the adjusted subsampling-based (subs) approach. The results are averaged over all features and grid points. The coverage values for the model-based approach are lower than the confidence level of 0.95 as well as lower than the coverage values of the subsampling-based approach for all scenarios. The difference between the two approaches is particularly large for the correctly specified linear model. While the subsampling-based approach shows almost perfect coverages for the different settings, the model-based approach is far off the nominal level with values around 0.35. This gap can be explained by the MC integration variance which is not incorporated in the model-based approaches. Hence, if the MC error is relatively high compared to the model variance, coverages are bad. To illustrate this relationship, we calculated the average standard deviation of the MC integration variance estimator (see Eq. (5)) over all repetitions for the model-based approaches which are provided in the last column of Table 3. Since the confidence bands of these approaches only cover the model variance, the confidence width is directly proportional to the model variance. If we now compare the "MC se" column with the average widths of the model-based approach, it is observable that coverages are rather low (e.g., 0.34 for DGP 1 and LM with $n = 200$) in the case where "MC se" divided by width is rather high (e.g., $0.15/0.15 = 1$). On the other hand coverages are high (e.g., 0.87 for DGP 2 and GP with $n = 200$) if "MC se" divided by width is lower (e.g., $0.16/0.63 \approx 0.25$) and hence the model variance is rather high compared to MC error.

Thus, if the main goal is to quantify both uncertainty sources inherent in the PD estimation and thus to receive reasonable coverages, the model-based approach cannot be recommended since only one of two sources of variability are covered by the estimates. Even for the linear model, which is commonly used for inferential purposes, the confidence bands for the PD estimates might be far too conservative as shown in Table 3. The subsampling-based variance estimates we proposed in this work however cover both the learner variance and the MC error and provide satisfying coverage values.

4 Application

We apply our proposed estimators to the motivational example from Section 1.1. We supposed that a researcher predicted chronic heart disease [13] (Cleveland data, $n = 296$) from sociological and medical indicators such as age, blood

[9] We did not consider the bootstrapping approach in our experiments as we encountered numerical issues in the invertability of the covariance matrix (due to duplicated values introduced by bootstrap) [54].

Table 3 Coverage probabilities for 95% confidence bands of PD estimates for model-based (mod) and subsampling-based (subs) approaches. Results are averaged over all features and grid points on DGPs 1 and 2 for the GP and LM. The experiments were conducted on two different sample sizes n. Furthermore, mean (standard deviation) of confidence width are reported for both approaches. The last column contains the standard deviation of the MC error for the model-based approach.

dgp	model	n	coverage		width (sd)		
			mod	subs	mod	subs	mod
1	gp	200	0.66	0.95	0.36 (0.19)	0.48 (0.11)	0.15
1	gp	1000	0.71	0.97	0.28 (0.31)	0.24 (0.07)	0.07
1	lm	200	0.34	0.95	0.15 (0.03)	0.41 (0.10)	0.15
1	lm	1000	0.35	0.95	0.06 (0.01)	0.19 (0.05)	0.07
2	gp	200	0.87	0.88	0.63 (0.09)	0.48 (0.11)	0.16
2	gp	1000	0.85	0.92	0.31 (0.05)	0.26 (0.05)	0.08
2	lm	200	0.37	0.94	0.17 (0.04)	0.49 (0.13)	0.18
2	lm	1000	0.35	0.94	0.07 (0.02)	0.22 (0.06)	0.08

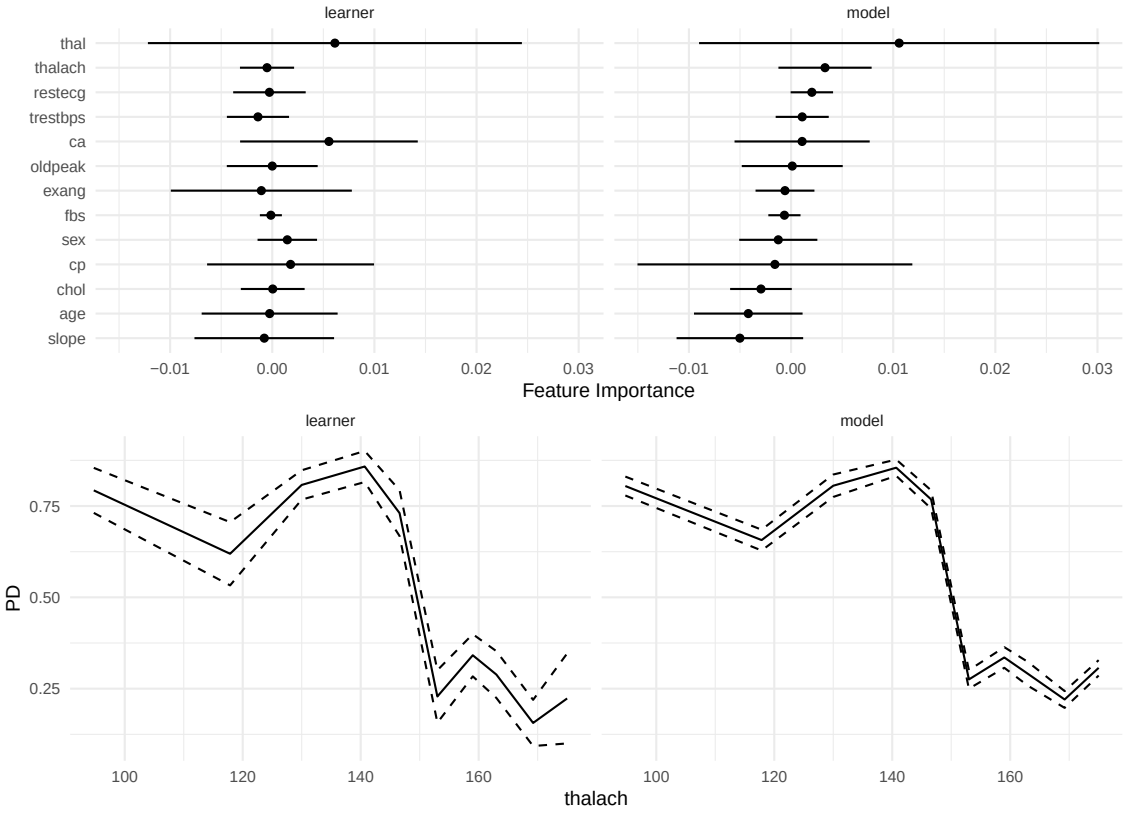


Figure 7 Top: Conditional Learner-PFI and model-PFI with point-wise 95%-confidence intervals for the random forest. Bottom: Conditional Learner-PDP and model-PDP with point-wise 95%-confidence bands for the random forest and feature `thalach`

pressure and maximum heart rate. She fitted one random forest and estimated conditional PFI and conditional PDPs to interpret the result.

Instead of only computing the conditional PFI and conditional PDP for one model, we estimate the proposed conditional model-PFI and conditional learner-PFI along with the proposed confidence intervals. For the learner-based insights, we therefore refitted the model 15 times on resampled training sets.

Figure 7 shows model and learner based conditional PFI and conditional PDP with the corresponding confidence intervals ($\alpha = 0.05$).

Conditional learner-PFI and model-PFI disagree on the ordering of the features: they agree that thalassemia (`thal`) is the most important feature; but conditional learner-PFI ranks number of major vessels (`ca`) and chest pain (`cp`) next, while conditional model-PFI ranks maximum heart rate (`thalach`) and resting state ECG (`restecg`) second and third. The confidence intervals for conditional model-PFI and learner-PFI indicate that both learner variance and the uncertainty stemming from the Monte Carlo integration are relatively high. All conditional model-PFI confidence intervals include zero, indicating that the high Monte Carlo uncertainty does not allow conclusions about the ground-truth conditional model-PFI. The conditional model-PFI cannot tell us to what extent the estimate varies due to learner variance; only the learner-PFI can quantify the learner variance. All conditional learner-PFI confidences intervals include zero as well (despite using the in comparison to bootstrapping less conservative resampling based confidence intervals), indicating that the high-ranked features may only appear relevant due to high model uncertainty.

Figure 7, bottom row, shows both the conditional model-PDP and the conditional learner-PDP for the maximum heart rate feature (`thalach`). We observe that individuals with high maximum heart rates are less likely to have chronic heart disease. Notably, the confidence bands of the conditional learner-PDP are wider than of the conditional model-PDP, and the conditional learner PDP confidences are especially wide for extreme values with little data support. Neglecting the learner variance would mean being overconfident about the partial dependence curve. In particular, the Monte Carlo approximation error decreases with $1/n$ as the sample size n for conditional PD and PFI estimation increases. Wrongly interpreted, this can lead to a false sense of confidence in the estimated effects and importance, since only one model is considered and learner variance is ignored.

5 Discussion

We related the PD and the PFI to the DGP, proposed variance and confidence intervals, and discussed conditions for inference. Our derivations were motivated by taking an external view of the statistical inference process and postulating that there is a ground truth counterpart to PD/PFI in the DGP. To the best of our knowledge, statistical inference via model-agnostic interpretable machine learning is already used in practice, but under-explored in theory.

A critical assumption for inference of effects and importance using interpretable machine learning is the unbiasedness of the learner. The learner bias is difficult to test, and can be introduced by e.g. choice of model class, regularization, and feature selection. For example, regularization techniques such as LASSO introduce a small bias *on purpose* [55] to decrease learner variance

and improve predictive performance. We must better understand how specific biases affect the prediction function and consequently PD and PFI estimates.

Another crucial limitation for inference of PD and PFI is the underestimation of variance due to data sharing between model refits. While we could show that a simple correction of the variance [12] vastly improves the coverage, a proper estimation of the variance remains an open issue. A promising approach relying on repeated nested cross validation to correctly estimate the variance was recently proposed by Bates et al. [56]. However, this approach is more computationally intensive by up to a factor of 1,000.

Furthermore, samplers are not readily available. Especially conditional sampling is a complex problem, and samplers must be trained using data. Training samplers even introduces another source of uncertainty to our estimates that we neglected in our work. It is difficult to separate this source of uncertainty from the uncertainty of the model learner, since trained samplers are correlated not only with each other, but possibly also with the trained models. We see integrating sampler uncertainty as an important step in providing reliable uncertainty estimates in practice, but we leave this to future work.

6 Statements and Declarations

Funding. This project is supported by the Bavarian State Ministry of Science and the Arts, the Bavarian Research Institute for Digital Transformation (bidt), the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A, by the German Research Foundation (DFG) – Emmy Noether Grant 437611051 to MNW, the Carl Zeiss Foundation (project on "Certification and Foundations of Safe Machine Learning Systems in Healthcare"), and by the Graduate School of Systemic Neurosciences (GSN) Munich. The authors of this work take full responsibilities for its content.

Conflicts of interest/Competing interests. No conflicts of interest.

Ethics approval. Not applicable.

Consent to participate. Not applicable

Consent for publication. Not applicable

Availability of data and material. The data used in the application is openly available and referenced in this paper.

Code availability. The code for visualizations, simulations and the application is written in the R programming language [57] and can be found in this repository: https://github.com/compstat-lmu/code_relating_pdp_pfi_to_dgp.

Authors' contributions. Contributions are reported using the CRediTtaxonomy. The authors are enumerated as follows: Christoph Molnar [1], Timo Freiesleben [2], Gunnar König [3], Julia Herbinger [4], Tim Reisinger [5], Giuseppe Casalicchio [6], Marvin N. Wright [7], Bernd Bischl [8].

Conceptualization: 1, 2, 3, 6, 8; *Methodology*: 1, 2, 3, 4, 5, 6; *Formal analysis and investigation*: 1, 2, 3, 5; *Writing - original draft preparation*: 1, 2; *Writing - review and editing*: 1, 2, 3, 4, 5, 6, 7, 8; *Visualization*: 1, 3; *Validation*: 1, 2, 3; *Software*: 1, 3; *Funding acquisition*: 1, 6, 7, 8; *Supervision*: 6, 7, 8.

Supplementary Material

Appendix A Bias and Variance of PD

The expected squared difference between model-PD and DGP-PD can be decomposed into bias and variance.

Proof

$$\begin{aligned}
\mathbb{E}_F[(PD_{\hat{f}}(x) - PD_f(x))^2] &= \mathbb{E}_F[PD_{\hat{f}}(x)^2] + \mathbb{E}_F[PD_f(x)^2] \\
&\quad - 2\mathbb{E}_F[PD_{\hat{f}}(x) PD_f(x)] \\
&= \mathbb{V}_F[PD_{\hat{f}}(x)] + \mathbb{E}_F[PD_{\hat{f}}(x)]^2 \\
&\quad + PD_f(x)^2 - 2\mathbb{E}_F[PD_{\hat{f}}(x) PD_f(x)] \\
&= \underbrace{(PD_f(x) - \mathbb{E}_F[PD_{\hat{f}}(x)])^2}_{\text{Bias}} + \underbrace{\mathbb{V}_F[PD_{\hat{f}}(x)]}_{\text{Variance}}
\end{aligned}$$

□

Appendix B Bias and Variance of PFI

The expected squared difference between model-PFI and DGP-PFI can be decomposed into bias and variance.

Proof

$$\begin{aligned}
\mathbb{E}_F[(PFI_{\hat{f}} - PFI_f)^2] &= \mathbb{E}_F[PFI_{\hat{f}}^2] + \mathbb{E}_F[PFI_f^2] \\
&\quad - 2\mathbb{E}_F[PFI_{\hat{f}} PFI_f] \\
&= \mathbb{V}_F[PFI_{\hat{f}}] + \mathbb{E}_F[PFI_{\hat{f}}]^2 \\
&\quad + PFI_f^2 - 2\mathbb{E}_F[PFI_{\hat{f}} PFI_f] \\
&= (PFI_f - \mathbb{E}_F[PFI_{\hat{f}}])^2 + \mathbb{V}_F[PFI_{\hat{f}}] \\
&= Bias_F^2[PFI_{\hat{f}}] + \mathbb{V}_F[PFI_{\hat{f}}]
\end{aligned}$$

□

Appendix C Model-PD Unbiasedness Regarding Theoretical PD

Proof By the law of large numbers, the Monte Carlo integration converges with $r \to \infty$ to the true integral. Assuming we have a fixed x, r identically distributed random draws $\tilde{X}_C^{(1,x)}, \ldots, \tilde{X}_C^{(r,x)} \sim \phi(x)$ and a model $\hat{f}$, the estimate is:

$$\begin{aligned}
\mathbb{E}_{\tilde{X}_C}[\widehat{PD}_{\hat{f}}(x)] &= \mathbb{E}_{\tilde{X}_C^{(1,x)}, \ldots, \tilde{X}_C^{(r,x)}} \left[\frac{1}{r} \sum_{i=1}^{r} \hat{f}(x, \tilde{X}_C^{(i,x)}) \right] \\
&= \frac{1}{r} r \mathbb{E}_{\tilde{X}_C}[\hat{f}(x, \tilde{X}_C)]
\end{aligned}$$

$$= PD_{\hat{f}}(x)$$

and therefore unbiased for the interval, i.e. the theoretical PD of the model. □

Appendix D Model-PD Unbiasedness Regarding DGP-PD

Proof Unbiasedness of the model $\hat{f}$ implies unbiasedness of the model-PD.

$$\begin{aligned}
\mathbb{E}_F[PD_{\hat{f}}(x)] &\overset{Def}{=} \int_F \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \phi(x)(\tilde{x}_c)\hat{f}(x, \tilde{x}_c)\, d\tilde{x}_c\, dP(F) \\
&\overset{Fub}{=} \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \int_F \phi(x)(\tilde{x}_c)\hat{f}(x, \tilde{x}_c)\, dP(F)\, d\tilde{x}_c \\
&\overset{const.}{=} \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \phi(x)(\tilde{x}_c) \int_F \hat{f}(x, \tilde{x}_c)\, dP(F)\, d\tilde{x}_c \\
&\overset{unbiased}{=} \int_{\tilde{x}_c \in \tilde{\mathcal{X}}_C} \phi(x)(\tilde{x}_c) f(x, \tilde{x}_c)\, d\tilde{x}_c \\
&\overset{def}{=} PD_f(x)
\end{aligned}$$

Fubini's theorem requires that $\int_{F,\tilde{X}_C} \mid \phi(x)(\tilde{X}_c)\hat{f}(\tilde{X}_c) \mid d\mathbb{P}_{F,X_C} < \infty$. One sufficient condition for this is when the model predictions have an upper bound $c :\mid \hat{f}(x) \mid < c < \infty$. □

Appendix E Model-PFI Regarding theoretical PFI

Proof As a function of random variables, the loss L itself is a random variable. We assume that the loss $L^{(i)}$ of observation i is a sample from the distribution of losses: $L^{(i)} \sim L$ and, similarly for the loss: $\tilde{L}^{(k,i)} \sim \tilde{L}$, where $L^{(i)} = L(y^{(i)}, \hat{f}(x^{(i)}))$ and $\tilde{L}^{(k,i)} = L(y^{(i)}, \hat{f}(\tilde{x}_S^{(k,i)}, x_C^{(i)}))$.

The expectation of our estimator is:

$$\begin{aligned}
\mathbb{E}_{\tilde{X}_S X_S X_C Y}[\widehat{PFI}_{\hat{f}}] &= \mathbb{E}_{\tilde{X}_S X_S X_C Y}\left[\frac{1}{n_2}\sum_{i=1}^{n_2}(\frac{1}{r}\sum_{k=1}^{r}(\tilde{L}^{(k,i)} - L^{(i)}))\right] \\
&= \frac{1}{n_2} n_2 \mathbb{E}_{\tilde{X}_S X_S X_C Y}[((\frac{1}{r} r\tilde{L}) - L)] \\
&= \mathbb{E}_{\tilde{X}_S X_C Y}[\tilde{L}] - \mathbb{E}_{X_S X_C Y}[L] \\
&= PFI_{\hat{f}}
\end{aligned}$$

In expectation, we retrieve the theoretical PFI of the model.

□

Appendix F PFI Biases for L2

In this proof, we use the conditional sampler ϕ_{cond} for both, the DGP-PFI and the model-PFI. Moreover, we assume that L is the squared loss $L(y, \hat{f}) =$

$(y - \hat{f}(x))^2$ and that $\mathbb{E}[Y \mid X]$ can be described by f with some additive, irreducible, error ϵ with $\mathbb{E}(\epsilon) = 0$ and $\mathbb{V}(\epsilon) = \sigma^2$. To further examine the bias for the PFI, we apply the Bias-Variance Decomposition additionally on the loss itself: In addition, we use that $\mathbb{E}_{XY}[Y] = \mathbb{E}_X[f(X)]$, $\mathbb{V}_Y[Y] = \sigma^2$ and $\mathbb{E}[A^2] = \mathbb{V}[A] + \mathbb{E}[A]^2$. We first derive the bias-variance decomposition of (i) permuted loss and (ii) original loss and therefrom derive the expected PFI.

For the permuted loss (i):

$$\begin{aligned}
\mathbb{E}_{F\tilde{X}_S XY}[\tilde{L}] &= \mathbb{E}_{F\tilde{X}_S XY}[(Y - \tilde{\hat{f}})^2] \\
&= \mathbb{E}_{\tilde{X}_S XY}[Y^2 - 2Y\mathbb{E}_F[\tilde{\hat{f}}] + \mathbb{E}_F[\tilde{\hat{f}}^2]] \\
&= \mathbb{E}_{\tilde{X}_S XY}[Y^2 - 2Y\mathbb{E}_F[\tilde{\hat{f}}] + \mathbb{E}_F[\tilde{\hat{f}}]^2 + \mathbb{V}_F[\tilde{\hat{f}}]] \\
&= \mathbb{V}_Y[Y] + \mathbb{E}_{\tilde{X}_S X}[f^2 - 2f\mathbb{E}_F[\tilde{\hat{f}}] + \mathbb{E}_F[\tilde{\hat{f}}]^2 + \mathbb{V}_F[\tilde{\hat{f}}]] \\
&= \underbrace{\sigma^2}_{\text{Data Var}} + \mathbb{E}_{\tilde{X}_S X} \underbrace{\left[(f - \mathbb{E}_F[\tilde{\hat{f}}])^2\right]}_{\text{Bias}^2} + \mathbb{E}_{\tilde{X}_S X} \underbrace{[\mathbb{V}_F[\tilde{\hat{f}}]]}_{\text{Variance}}
\end{aligned}$$

For the original loss (ii):

$$\begin{aligned}
\mathbb{E}_{FXY}[L] &= \mathbb{E}_{FXY}[(Y - \hat{f})^2] \\
&= \mathbb{E}_{XY}[Y^2 - 2Y\mathbb{E}_F[\hat{f}] + \mathbb{E}_F[\hat{f}^2]] \\
&= \mathbb{E}_{XY}[Y^2 - 2Y\mathbb{E}_F[\hat{f}] + \mathbb{E}_F[\hat{f}]^2 + \mathbb{V}_F[\hat{f}]] \\
&= \mathbb{V}_Y[Y] + \mathbb{E}_X[f^2 - 2f\mathbb{E}_F[\hat{f}] + \mathbb{E}_F[\hat{f}]^2 + \mathbb{V}_F[\hat{f}]] \\
&= \underbrace{\sigma^2}_{\text{Data Var}} + \mathbb{E}_X \underbrace{\left[(f - \mathbb{E}_F[\hat{f}])^2\right]}_{\text{Bias}^2} + \mathbb{E}_X \underbrace{[\mathbb{V}_F(\hat{f})]}_{\text{Variance}}
\end{aligned}$$

The expected PFI for feature X_S then is:

$$\begin{aligned}
\mathbb{E}_F[PFI_{\hat{f}}] &= \mathbb{E}_{F\tilde{X}_S XY}[\tilde{L}] - \mathbb{E}_{FXY}[L] \\
&\overset{(i)+(ii)}{=} \sigma^2 + \mathbb{E}_{\tilde{X}_S X}\left[(f - \mathbb{E}_F[\tilde{\hat{f}}])^2\right] + \mathbb{E}_{\tilde{X}_S X}[\mathbb{V}_F(\tilde{\hat{f}})] \\
&\quad - (\sigma^2 + \mathbb{E}_X\left[(f - \mathbb{E}_F[\hat{f}])^2\right] + \mathbb{E}_X[\mathbb{V}_F(\hat{f})]) \\
&= \mathbb{E}_{\tilde{X}_S X}\left[(f - \mathbb{E}_F[\tilde{\hat{f}}])^2\right] - \mathbb{E}_X\left[(f - \mathbb{E}_F[\hat{f}])^2\right] \\
&\quad + \mathbb{E}_{\tilde{X}_S X}[\mathbb{V}_F[\tilde{\hat{f}}]] - \mathbb{E}_X[\mathbb{V}_F[\hat{f}]]
\end{aligned}$$

We can derive the same L2 decomposition for the DGP-PFI by replacing $\hat{f}$ with f in the equation above. This yields $PFI_f = \mathbb{E}_{\tilde{X}_S X}[(f(X) - f(\tilde{X}_S, X_C))^2]$, since $\mathbb{V}_F[f] = \mathbb{V}_F[\tilde{f}] = 0$ and $\mathbb{E}_F[f] = f$ and $\mathbb{E}_F[\tilde{f}] = \tilde{f}$.

The bias of the model-PFI compared to the DGP-PFI is:

$$\mathbb{E}_F[PFI_{\hat{f}}] - PFI_f = \underbrace{\mathbb{E}_{\tilde{X}_S X}[(f - \mathbb{E}_F[\tilde{\hat{f}}])^2 - (f - \tilde{f})^2]}_{\text{Permutation Loss Bias}} \tag{F1}$$

$$- \underbrace{\mathbb{E}_X\left[(f - \mathbb{E}_F[\hat{f}])^2\right]}_{(\text{Learner Bias})^2} + \underbrace{\mathbb{E}_{\tilde{X}_S X}[\mathbb{V}_F[\hat{f}]] - \mathbb{E}_X[\mathbb{V}_F[\hat{f}]]}_{\text{Variance Inflation}} \tag{F2}$$

$$\overset{unbiased}{=} \underbrace{\mathbb{E}_{\tilde{X}_S X}[\mathbb{V}_F[\hat{f}]] - \mathbb{E}_X[\mathbb{V}_F[\hat{f}]]}_{\text{Variance Inflation}} \tag{F3}$$

$$\overset{\tilde{X}_S \sim X_S | X_C}{=} 0 \tag{F4}$$

The permutation loss bias and the squared learner bias are zero due to the unbiasedness assumption, i.e. $\mathbb{E}_F[\hat{f}] = f$. The variance inflation term is zero if $\tilde{X}_S \sim X_S \mid X_C$, which is here the case due to conditional sampling.

Appendix G conditional DGP-PFI minus model-PFI for L2

In this proof, we use the conditional sampler ϕ_{cond} for both, the DGP-PFI and the model-PFI.

$$\begin{aligned}
PFI_f - PFI_{\hat{f}} &= \mathbb{E}_{\tilde{X_S} X_C Y}[(Y - f)^2] - \mathbb{E}_{X_S X_C Y}[(Y - f)^2] \\
&\quad - \left(\mathbb{E}_{\tilde{X_S} X_C Y}[(Y - \hat{f})^2] - \mathbb{E}_{X_S X_C Y}[(Y - \hat{f})^2]\right) \\
&= \underbrace{\left(\mathbb{E}_{X_S X_C Y}[(Y - \hat{f})^2] - \mathbb{E}_{X_S X_C Y}[(Y - f)^2]\right)}_{\text{T1:=}} \\
&\quad + \underbrace{\left(\mathbb{E}_{\tilde{X_S} X_C Y}[(Y - f))^2] - \mathbb{E}_{\tilde{X_S} X_C Y}[(Y - \hat{f})^2]\right)}_{\text{T2:=}}
\end{aligned}$$

We know that for any $g : X \to Y$ holds:

$$\mathbb{E}_{X,Y}[(Y - g)^2] = \mathbb{E}_X[\mathbb{V}_{Y|X}[Y]] + \mathbb{E}_X[(\mathbb{E}_{Y|X}[Y] - g)^2]$$

Since $f = \mathbb{E}_{Y|X_S,X_C}[Y]$ we can conclude for our first term T1 that:

$$\begin{aligned}
\text{T1} &= \mathbb{E}_{X_S X_C}[\mathbb{V}_{Y|X_S,X_C}[Y]] + \mathbb{E}_{X_S X_C}[(f - \hat{f})^2] \\
&\quad - \left(\mathbb{E}_{X_S X_C}[\mathbb{V}_{Y|X_S,X_C}[Y]] + \underbrace{\mathbb{E}_{X_S X_C}[(f - f)^2]}_{=0}\right) \\
&= \mathbb{E}_{X_S X_C}[(f - \hat{f})^2]
\end{aligned}$$

We apply the same strategy to T2. Moreover, $Y \perp\!\!\!\perp \tilde{X}_S \mid X_C$.

$$\begin{aligned}
\text{T2} &= \mathbb{E}_{\tilde{X_S}X_C}[\mathbb{V}_{Y|\tilde{X}_S,X_C}[Y]] + \mathbb{E}_{\tilde{X_S}X_C}[(\mathbb{E}_{Y|\tilde{X}_S,X_C}[Y] - f)^2] \\
&- \left(\mathbb{E}_{\tilde{X_S}X_C}[\mathbb{V}_{Y|\tilde{X}_S,X_C}[Y]] + \mathbb{E}_{\tilde{X_S}X_C}[(\mathbb{E}_{Y|\tilde{X}_S,X_C}[Y] - \hat{f})^2]\right) \\
&= \mathbb{E}_{\tilde{X_S}X_C}[(\mathbb{E}_{Y|X_C}[Y] - f)^2] - \mathbb{E}_{\tilde{X_S}X_C}[(\mathbb{E}_{Y|X_C}[Y] - \hat{f})^2]
\end{aligned}$$

If we now set together the two terms again and use in the first step that $P(X_S, X_C) = P(\tilde{X}_S, X_C)$, we obtain:

$$\begin{aligned}
\text{T1+T2} &= \mathbb{E}_{X_SX_C}[(f - \hat{f})^2] + \mathbb{E}_{X_SX_C}[(\mathbb{E}_{Y|X_C}[Y] - f)^2] \\
&- \mathbb{E}_{X_SX_C}[(\mathbb{E}_{Y|X_C}[Y] - \hat{f})^2] \\
&= \mathbb{E}_{X_SX_C}\Big[f^2 - 2f\hat{f} + \hat{f}^2 + \mathbb{E}_{Y|X_C}[Y]^2 - 2\mathbb{E}_{Y|X_C}[Y]f + f^2 \\
&- \mathbb{E}_{Y|X_C}[Y]^2 + 2\mathbb{E}_{Y|X_C}[Y]\hat{f} - \hat{f}^2\Big] \\
&= 2\mathbb{E}_{X_SX_C}\Big[(f^2 - \mathbb{E}_{Y|X_C}[Y]f) - (f\hat{f} - \mathbb{E}_{Y|X_C}[Y]\hat{f})\Big] \\
&= 2\mathbb{E}_{X_C}\Big[\mathbb{E}_{X_S|X_C}\big[(f^2 - \mathbb{E}_{Y|X_C}[Y]f) - (f\hat{f} - \mathbb{E}_{Y|X_C}[Y]\hat{f})\big]\Big] \\
&\overset{*}{=} 2\mathbb{E}_{X_C}\Big[(\mathbb{E}_{X_S|X_C}[f^2] - \mathbb{E}_{Y|X_C}[Y]\mathbb{E}_{X_S|X_C}[f]) \\
&- (\mathbb{E}_{X_S|X_C}[f\hat{f}] - \mathbb{E}_{Y|X_C}[Y]\mathbb{E}_{X_S|X_C}[\hat{f}])\Big] \\
&\overset{**}{=} 2\mathbb{E}_{X_C}\Big[(\mathbb{E}_{X_S|X_C}[f^2] - \mathbb{E}_{X_S|X_C}[f]^2) \\
&- (\mathbb{E}_{X_S|X_C}[f\hat{f}] - \mathbb{E}_{X_S|X_C}[\hat{f}]\mathbb{E}_{X_S|X_C}[f])\Big] \\
&= 2\mathbb{E}_{X_C}\left[\mathbb{V}_{X_S|X_C}[f] - Cov_{X_S|X_C}[f, \hat{f}]\right]
\end{aligned}$$

At *, we use the fact that the random variable $\mathbb{E}_{Y|X_C}[Y]$ is measurable by the σ-Algebra generated from X_C, and we are inclined to pull it out of the expectation. In **, we use that from $f = \mathbb{E}_{Y|X_S,X_C}[Y]$ follows $\mathbb{E}_{X_S|X_C}[f] = \mathbb{E}_{Y|X_C}[Y]$.

Appendix H CI simulation results

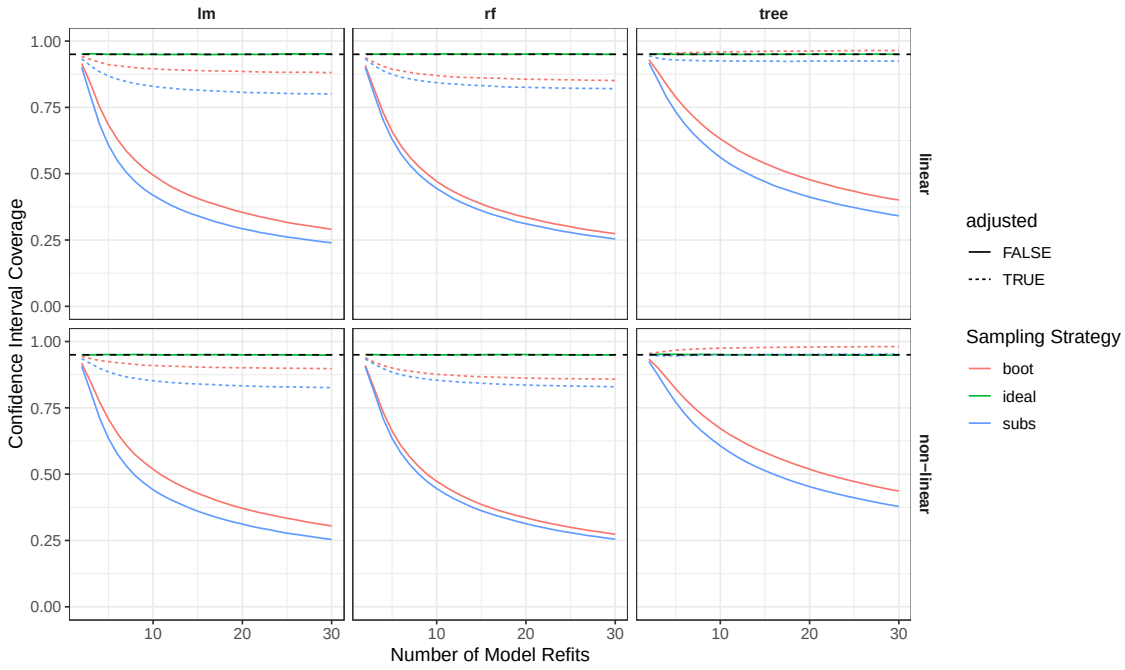


Figure H1 CI coverage for PD with n=100.

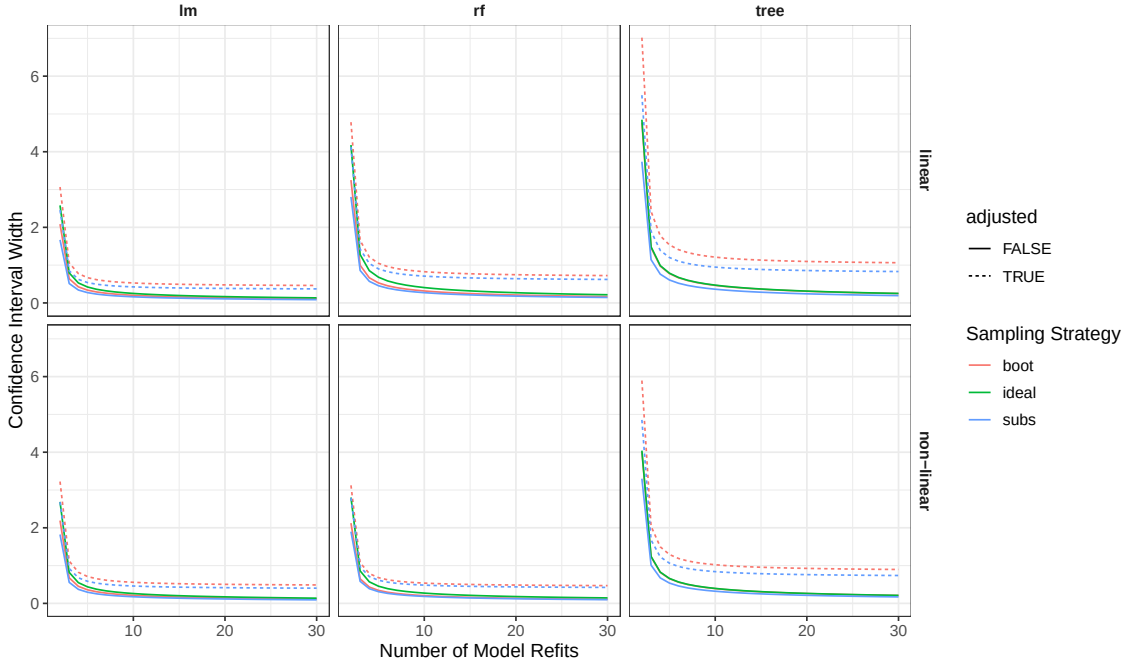


Figure H2 CI width for PD with n=100.

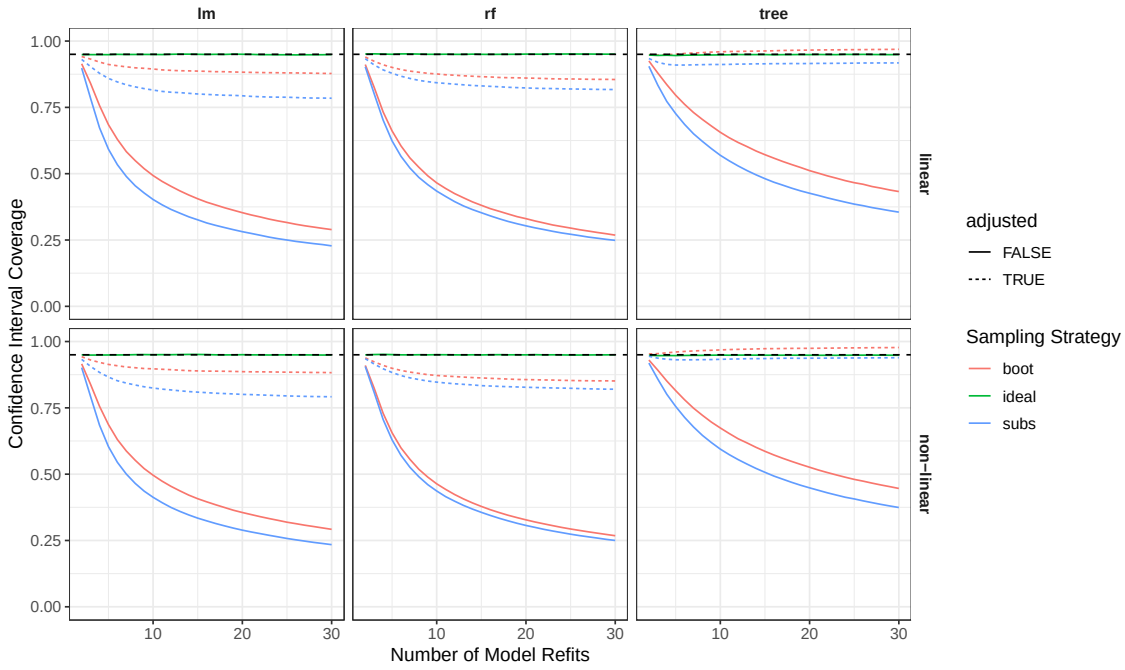


Figure H3 CI coverage for PD with n=1,000.

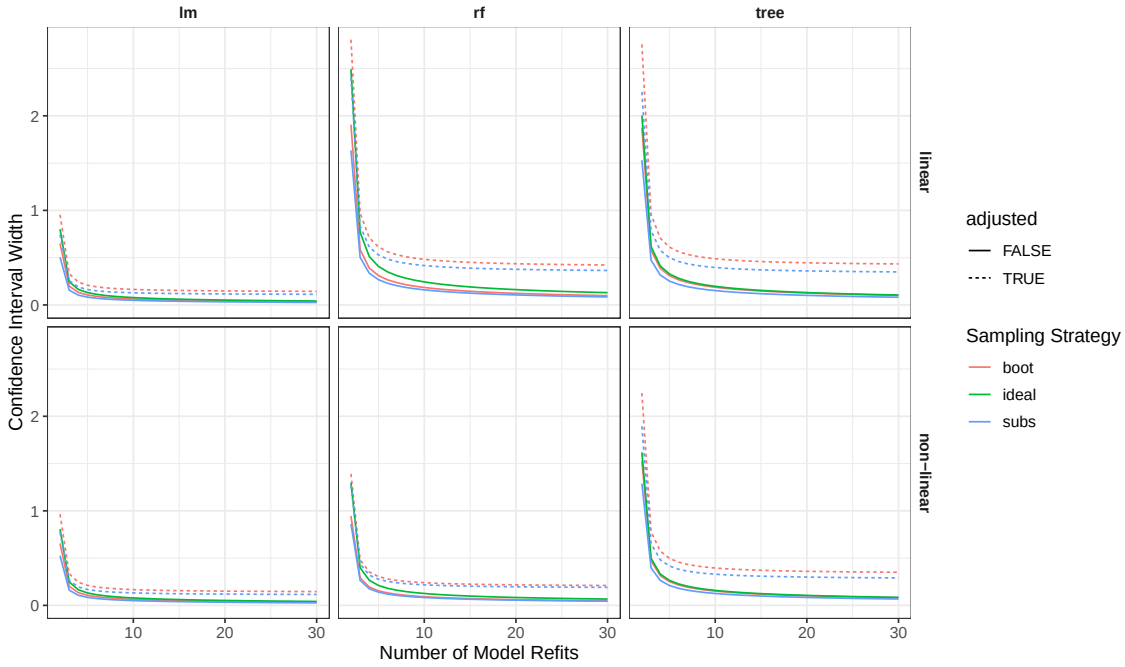


Figure H4 CI width for PD with n=1,000.

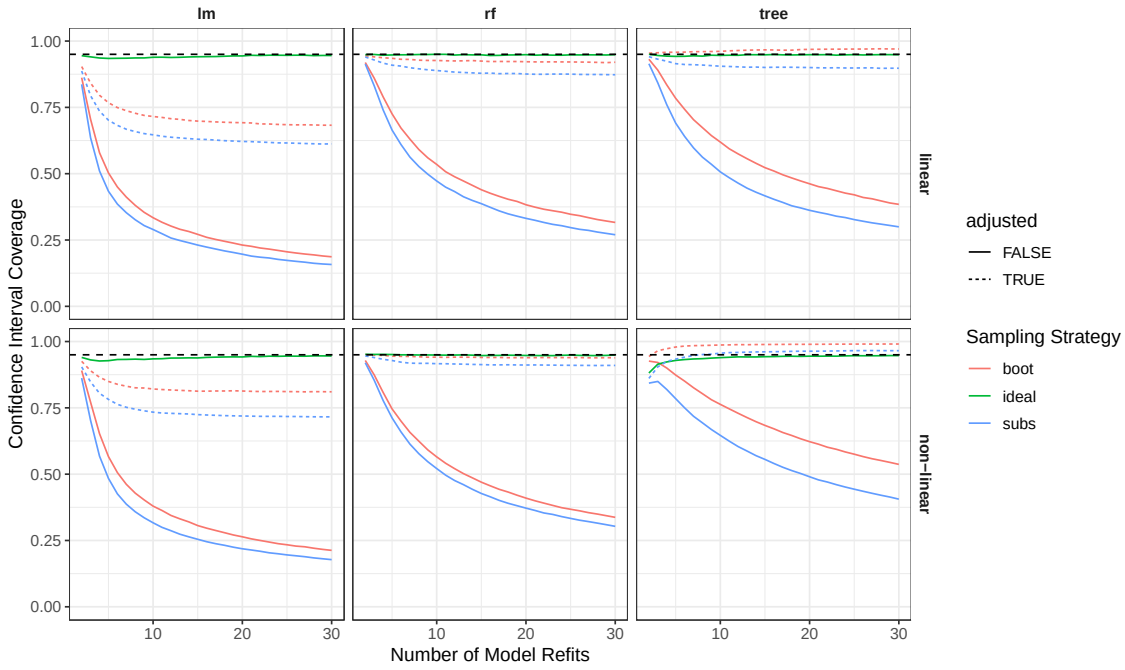


Figure H5 CI coverage for PFI with n=100.

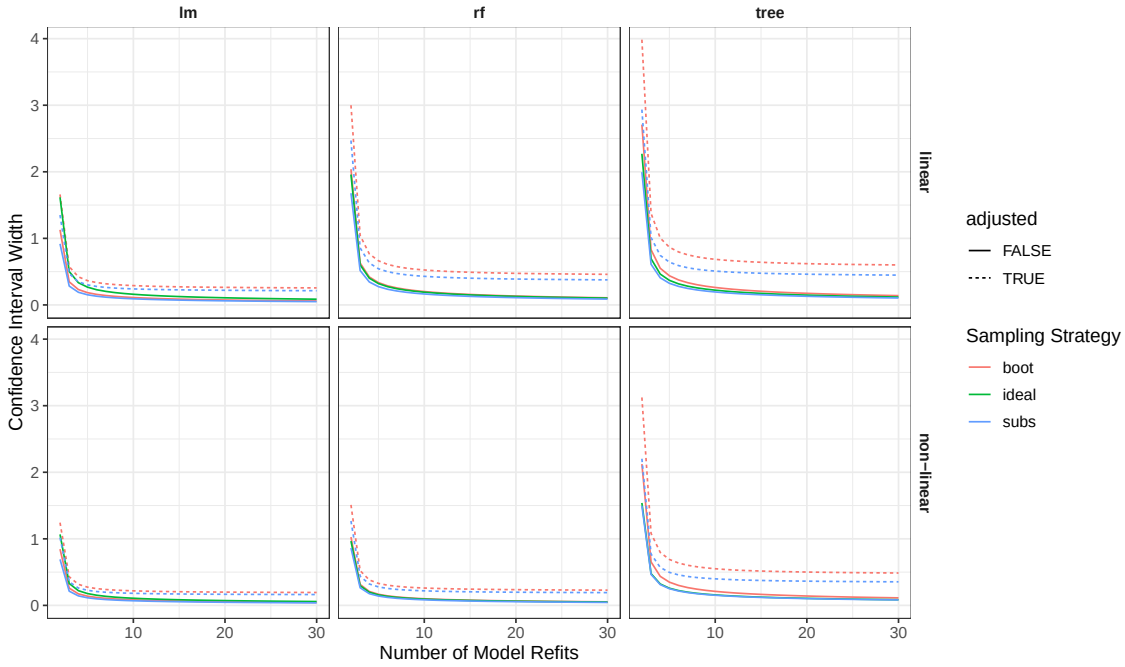


Figure H6 CI width for PFI with n=100.

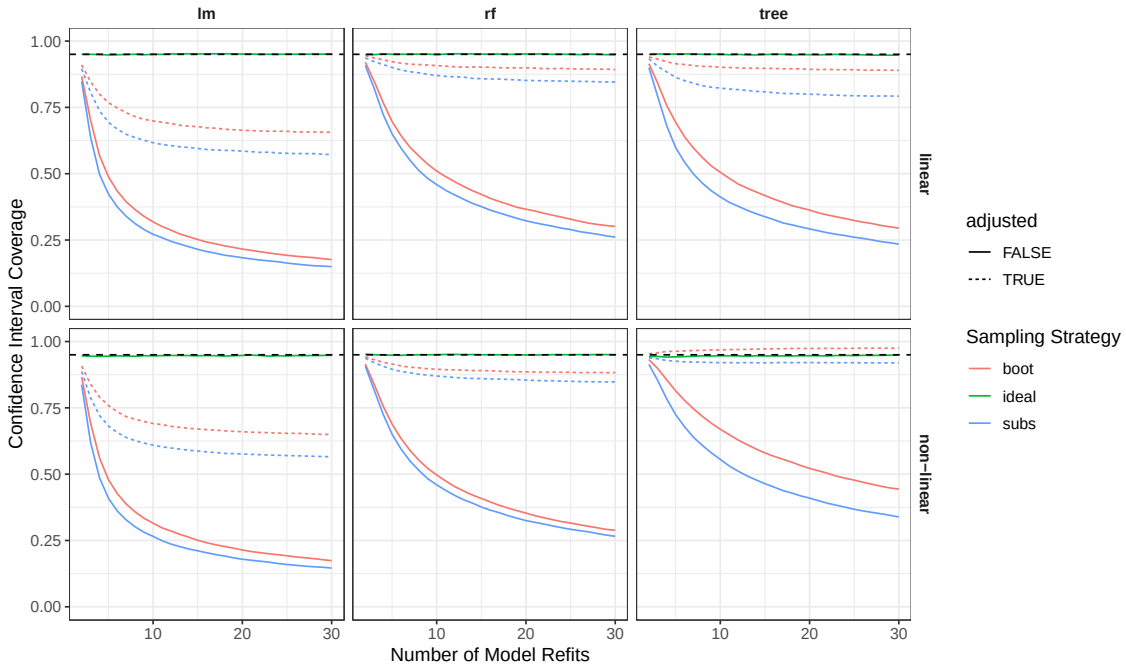


Figure H7 CI coverage for PFI with n=1,000.

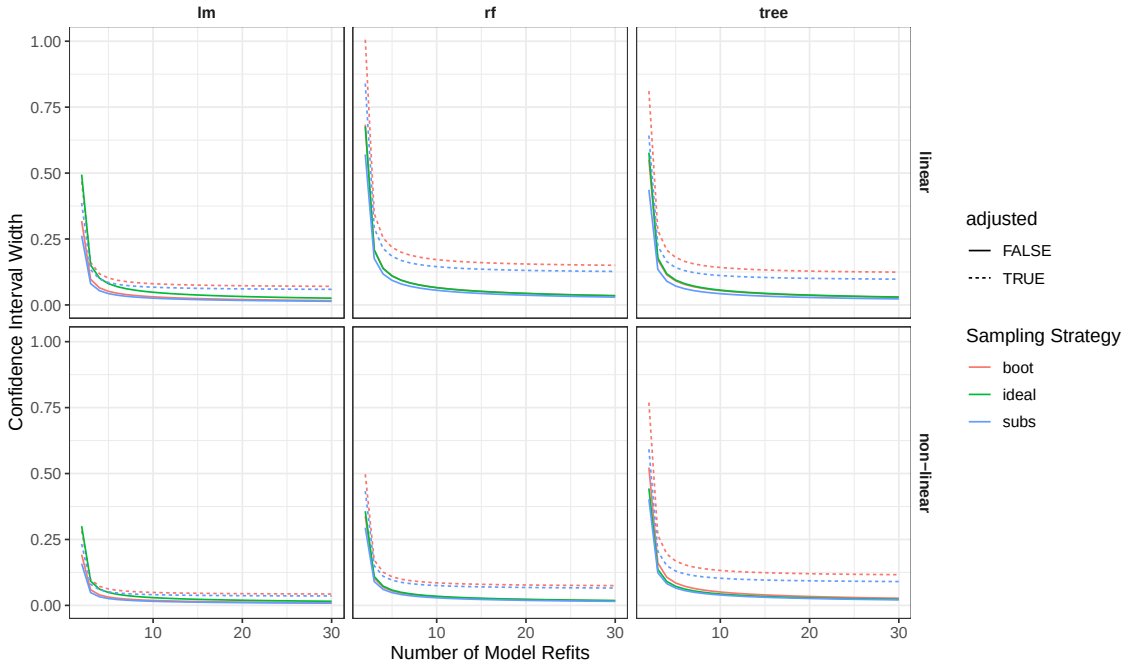


Figure H8 CI width for PFI with n=1,000.

References

[1] Fahrmeir, L., Kneib, T., Lang, S., Marx, B.: Regression, (2007). https://doi.org/10.1007/978-3-642-34333-9. Springer

[2] Bair, E., Ohrbach, R., Fillingim, R.B., Greenspan, J.D., Dubner, R., Diatchenko, L., Helgeson, E., Knott, C., Maixner, W., Slade, G.D.: Multivariable modeling of phenotypic risk factors for first-onset tmd: the oppera prospective cohort study. The Journal of Pain **14**(12), 102–115 (2013). https://doi.org/10.1016/j.jpain.2013.09.003

[3] Esselman, P.C., Stevenson, R.J., Lupi, F., Riseng, C.M., Wiley, M.J.: Landscape prediction and mapping of game fish biomass, an ecosystem service of michigan rivers. North American Journal of Fisheries Management **35**(2), 302–320 (2015). https://doi.org/10.1080/02755947.2014.987887

[4] Obringer, R., Nateghi, R.: Predicting urban reservoir levels using statistical learning techniques. Scientific Reports **8**(1), 1–9 (2018). https://doi.org/10.1038/s41598-018-23509-w

[5] Boulesteix, A.-L., Wright, M.N., Hoffmann, S., König, I.R.: Statistical learning approaches in the genetic epidemiology of complex diseases. Human Genetics **139**(1), 73–84 (2020). https://doi.org/10.1007/s00439-019-01996-9

[6] Stachl, C., Au, Q., Schoedel, R., Gosling, S.D., Harari, G.M., Buschek, D., Völkel, S.T., Schuwerk, T., Oldemeier, M., Ullmann, T., Hussmann, H., Bischl, B., Bühner, M.: Predicting personality from patterns of behavior collected with smartphones. Proceedings of the National Academy of Sciences **117**(30), 17680–17687 (2020). https://doi.org/10.1073/pnas.1920484117

[7] Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. ICML WHI '16 (2016) https://arxiv.org/abs/1606.05386

[8] Scholbeck, C.A., Molnar, C., Heumann, C., Bischl, B., Casalicchio, G.: Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. Communications in Computer and Information Science, 205–216 (2020). https://doi.org/10.1007/978-3-030-43823-4_18

[9] Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189–1232 (2001)

[10] Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001). https:

//doi.org/10.1023/A:1010933404324

[11] Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. Journal of Machine Learning Research **20**(177), 1–81 (2019)

[12] Nadeau, C., Bengio, Y.: Inference for the generalization error. Machine Learning **52**(3), 239–281 (2003). https://doi.org/10.1023/A:1024068626366

[13] Dua, D., Graff, C.: UCI Machine Learning Repository (2017)

[14] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees, (1984). https://doi.org/10.1201/9781315139470. CRC Press, Boca Raton

[15] Chen, H., Janizek, J.D., Lundberg, S., Lee, S.-I.: True to the model or true to the data? arXiv preprint arXiv:2006.16234 (2020)

[16] Freiesleben, T., König, G., Molnar, C., Tejero-Cantero, A.: Scientific inference with interpretable machine learning: Analyzing models to learn about real-world phenomena. arXiv preprint arXiv:2206.05487 (2022)

[17] König, G., Molnar, C., Bischl, B., Grosse-Wentrup, M.: Relative feature importance. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9318–9325 (2021). IEEE

[18] Watson, D.S., Wright, M.N.: Testing conditional independence in supervised learning algorithms. Machine Learning **110**, 2107–2129 (2021). https://doi.org/10.1007/s10994-021-06030-6

[19] Bates, S., Candès, E., Janson, L., Wang, W.: Metropolized knockoff sampling. Journal of the American Statistical Association **116**(535), 1413–1427 (2021)

[20] Blesch, K., Watson, D.S., Wright, M.N.: Conditional feature importance for mixed data. arXiv preprint arXiv:2210.03047 (2022)

[21] Cafri, G., Bailey, B.A.: Understanding variable effects from black box prediction: Quantifying effects in tree ensembles using partial dependence. Journal of Data Science **14**(1), 67–95 (2016)

[22] Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., Bischl, B.: Explaining hyperparameter optimization via partial dependence plots. Advances in Neural Information Processing Systems **34**, 2280–2291 (2021)

[23] Grange, S.K., Carslaw, D.C.: Using meteorological normalisation to detect interventions in air quality time series. Science of The Total Environment **653**, 578–588 (2019). https://doi.org/10.1016/j.scitotenv.2018.10.344

[24] Emrich, E., Pierdzioch, C.: Public goods, private consumption, and human capital: Using boosted regression trees to model volunteer labour supply. Review of Economics/Jahrbuch für Wirtschaftswissenschaften **67**(3) (2016). https://doi.org/10.1515/roe-2016-0004

[25] Page, W.G., Wagenbrenner, N.S., Butler, B.W., Forthofer, J.M., Gibson, C.: An evaluation of ndfd weather forecasts for wildland fire behavior prediction. Weather and Forecasting **33**(1), 301–315 (2018). https://doi.org/10.1175/WAF-D-17-0121.1

[26] Ishwaran, H., Lu, M.: Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival. Statistics in Medicine **38**(4), 558–582 (2019). https://doi.org/10.1002/sim.7803

[27] Archer, K.J., Kimes, R.V.: Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis **52**(4), 2249–2260 (2008). https://doi.org/10.1016/j.csda.2007.08.015

[28] Janitza, S., Celik, E., Boulesteix, A.-L.: A computationally fast variable importance test for random forests for high-dimensional data. Advances in Data Analysis and Classification **12**(4), 885–915 (2018). https://doi.org/10.1007/s11634-016-0276-4

[29] Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010). https://doi.org/10.1093/bioinformatics/btq134

[30] Williamson, B.D., Gilbert, P.B., Carone, M., Simon, N.: Nonparametric variable importance assessment using machine learning techniques. Biometrics (2019). https://doi.org/10.1111/biom.13392

[31] Williamson, B.D., Gilbert, P.B., Simon, N.R., Carone, M.: A unified approach for inference on algorithm-agnostic variable importance. arXiv preprint arXiv:2004.03683 (2020)

[32] Parr, T., Wilson, J.D., Hamrick, J.: Nonparametric feature impact and importance. arXiv preprint arXiv:2006.04750 (2020)

[33] Parr, T., Wilson, J.D.: A stratification approach to partial dependence for codependent variables. arXiv preprint arXiv:1907.06698 (2019)

[34] Zhang, L., Janson, L.: Floodgate: inference for model-free variable

importance. arXiv preprint arXiv:2007.01283 (2020)

[35] Hastie, T., Tibshirani, R., Friedman, J.H., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction vol. 2. Springer, ??? (2009)

[36] Hothorn, T., Leisch, F., Zeileis, A., Hornik, K.: The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics **14**(3), 675–699 (2005). https://doi.org/10.1198/106186005X59630

[37] Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C.A., Casalicchio, G., Grosse-Wentrup, M., Bischl, B.: In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W. (eds.) General pitfalls of model-agnostic interpretation methods for machine learning models, pp. 39–68. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-04083-2_4. https://doi.org/10.1007/978-3-031-04083-2_4

[38] Molnar, C., König, G., Bischl, B., Casalicchio, G.: Model-agnostic feature importance and effects with dependent features–a conditional subgroup approach. arXiv preprint arXiv:2006.04628 (2020)

[39] Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**(4), 1059–1086 (2020). https://doi.org/10.1111/rssb.12377

[40] Zhao, Q., Hastie, T.: Causal interpretations of black-box models. Journal of Business & Economic Statistics **39**(1), 272–281 (2021). https://doi.org/10.1080/07350015.2019.1624293

[41] Bishop, C.M.: Mixture density networks. Technical report, Aston University (1994)

[42] Bashtannyk, D.M., Hyndman, R.J.: Bandwidth selection for kernel conditional density estimation. Computational Statistics & Data Analysis **36**(3), 279–298 (2001)

[43] Sohn, K., Lee, H., Yan, X.: Learning structured output representation using deep conditional generative models. Advances in neural information processing systems **28** (2015)

[44] Trippe, B.L., Turner, R.E.: Conditional density estimation with bayesian normalising flows. arXiv preprint arXiv:1802.04908 (2018)

[45] Winkler, C., Worrall, D., Hoogeboom, E., Welling, M.: Learning likelihoods with conditional normalizing flows. arXiv preprint arXiv:1912.00042 (2019)

[46] Hothorn, T., Zeileis, A.: Predictive distribution modeling using transformation forests. Journal of Computational and Graphical Statistics **30**(4), 1181–1196 (2021)

[47] Zheng, W., van der Laan, M.J.: Cross-validated targeted minimum-loss-based estimation. In: Targeted Learning, pp. 459–474 (2011). https://doi.org/10.1007/978-1-4419-9782-1_27. Springer

[48] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J.: Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal **21**(1), 1–68 (2018). https://doi.org/10.1111/ectj.12097

[49] Hooker, G., Mentch, L.: Please stop permuting features: An explanation and alternatives. arXiv preprint arXiv:1905.03151 (2019)

[50] Candes, E., Fan, Y., Janson, L., Lv, J.: Panning for gold: 'model-x'knockoffs for high dimensional controlled variable selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **80**(3), 551–577 (2018). https://doi.org/10.1111/rssb.12265

[51] Groemping, U.: Model-agnostic effects plots for interpreting machine learning models. Reports in Mathematics, Physics and Chemistry, Department II, Beuth University of Applied Sciences Berlin. **Report 1/2020** (2020)

[52] Geman, S., Bienenstock, E., Doursat, R.: Neural networks and the bias/-variance dilemma. Neural Computation **4**(1), 1–58 (1992). https://doi.org/10.1162/neco.1992.4.1.1

[53] Mitchell, T.M.: The Need for Biases in Learning Generalizations, (1980). Department of Computer Science, Laboratory for Computer Science Research

[54] Roustant, O., Ginsbourger, D., Deville, Y.: Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. Journal of Statistical Software **51**(1), 1–55 (2012). https://doi.org/10.18637/jss.v051.i01

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

[56] Bates, S., Hastie, T., Tibshirani, R.: Cross-validation: what does it estimate and how well does it do it? arXiv preprint arXiv:2104.00673 (2021)

[57] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). R Foundation for Statistical Computing. https://www.R-project.org/

Chapter 4

Paper III: The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples

Freiesleben, T. (2022). The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds & Machines*, 32, 77-109, doi: https://doi.org/10.1007/s11023-021-09580-9

Author contributions:

T.F. is the only author.

Minds and Machines (2022) 32:77–109
https://doi.org/10.1007/s11023-021-09580-9

ORIGINAL RESEARCH

The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples

Timo Freiesleben[1]

Received: 30 April 2021 / Accepted: 22 October 2021 / Published online: 30 October 2021

Abstract

The same method that creates adversarial examples (AEs) to fool image-classifiers can be used to generate counterfactual explanations (CEs) that explain algorithmic decisions. This observation has led researchers to consider CEs as AEs by another name. We argue that the relationship to the true label and the tolerance with respect to proximity are two properties that formally distinguish CEs and AEs. Based on these arguments, we introduce CEs, AEs, and related concepts mathematically in a common framework. Furthermore, we show connections between current methods for generating CEs and AEs, and estimate that the fields will merge more and more as the number of common use-cases grows.

Keywords Counterfactual explanation · Adversarial example · XAI · AI-safety

1 Introduction

Machine Learning (ML) is transforming industry, science, and our society. Today, ML algorithms can fix a date at the hairdresser (Leviathan and Matias 2018), determine a protein's 3D shape from its amino-acid sequence (Senior et al. 2020), and even write news articles (Brown et al. 2020). Taking a sharp look at these developments, we observe a tendency towards more and more complex models. Different ML models are stacked together heuristically, with limited theoretical backing (Hutson 2018). In some applications, complexity may not be an issue as long as the algorithm performs well most of the time. However, in socially, epistemically, or safety-critical domains, complexity can rule out ML solutions—think of e.g. autonomous driving, scientific discovery, or criminal justice. Two of the major drawbacks of highly complex algorithms are the *opaqueness problem* (Lipton 2018) and *adversarial attacks* (Szegedy et al. 2014).

✉ Timo Freiesleben
timo.freiesleben@campus.lmu.de

[1] Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität, Ludwigstrasse 31, Munich, Germany

The opaqueness problem describes the limited epistemic access humans have to the inner workings of ML algorithms, especially concerning the semantic interpretation of parameters, the learning process, and the human-predictability of ML decisions (Burrell 2016). This lack of interpretability has gained a lot of attention recently, which gave rise to the field eXplainable Artificial Intelligence (XAI; Doshi-Velez and Kim 2017; Rudin 2019). Many techniques have been proposed to gain insights into ML systems (Adadi and Berrada 2018; Došilović et al. 2018; Das and Rad 2020). Especially model-agnostic methods have gained attraction since, unlike model-specific methods, their application is not restricted to a specific model type (Molnar 2019). Global model-agnostic interpretation techniques like Permutation Feature Importance (Fisher et al. 2019) or Partial Dependence Plots (Friedman et al. 1991) aim at understanding the general properties of ML algorithms. On the other side, local model-agnostic interpretation methods like LIME (Ribeiro et al. 2016) or Shapley Values (Štrumbelj and Kononenko 2014) aim at understanding the behavior of algorithms for particular regions. One way to explain a specific model-prediction is a Counterfactual Explanation (CE; Wachter et al. 2017). A CE explains a prediction by presenting a maximally close alternative input that would have resulted in a different (usually desired) prediction. CEs are the first class of objects we study in this paper.

The problem of adversarial attacks describes the fact that complex ML algorithms are vulnerable to deceptions (Papernot et al. 2016a; Goodfellow et al. 2015; Szegedy et al. 2014). Such malfunctions can be exploited by attackers to e.g. harm model-employers or endanger end-users (Song et al. 2018). The field that investigates adversarial attacks is called adversarial ML (Joseph et al. 2018). If the attack happens during the training process by inserting mislabeled training data, the attack is called poisoning. If an attack happens after the training process, it is commonly called an adversarial example (AE; Serban et al. 2020). AEs are inputs that resemble real data but are misclassified by a trained ML model, e.g., the image of a turtle is classified as a riffle (Athalye et al. 2018). Hence, misclassified means here that the algorithm assigns the wrong class/value compared to some (usually human-given) ground-truth (Elsayed et al. 2018). AEs are the second class of objects relevant to our study.

Even though the opaqueness problem and the problem of adversarial attacks seem unrelated at first sight, there are good reasons to study them jointly. AEs show where an ML model fails, and examining these failures deepens our understanding of the model (Tomsett et al. 2018; Dong et al. 2017). Explanations on the other hand can shed light on how ML algorithms can be improved to make them more robust against AEs (Molnar 2019). As a downside, explanations may enclose too much information about the model, thereby allowing AEs to be constructed and the model attacked (Ignatiev et al. 2019; Sokol and Flach 2019). CEs are even stronger connected to AEs than other explanations. CEs and AEs can be obtained by solving the same optimization problem[1] (Wachter et al. 2017; Szegedy et al. 2014):

[1] x describes the original input, x' the counterfactual/adversarial vector, f the ML model, y_{des} the desired classification, $d(\cdot,\cdot)$ and $d'(\cdot,\cdot)$ distances, and λ a trade-off scalar. For details, see Sect. 4.3.

$$\underset{x' \in X}{\operatorname{argmin}}\ d(x, x') + \lambda\, d'(f(x'), y_{des}). \tag{1}$$

Term 1 has led to various confusions concerning the relationship between CEs and AEs in the research community.[2] We aim to resolve them and give a detailed analysis of the relationship between the two fields.

The aim of the present paper is twofold. Our first goal is the *clarification of concepts*. Commonly used concepts such as CE/AE, flipping/misclassifying, process/model-level, and closeness/distance are often misunderstood or not clearly defined. We define these terms properly in one mathematical framework, aiming for more clarity and unification. The second goal is to *familiarize researchers* of each of the respective fields *with its neighboring area*. Even in one of the fields, it is hard to keep track of developments and new ideas, in both it is worse. Since there are many ways in which each of the fields can profit from the other, both methodologically and conceptually, we aim to provide a guide connecting the two literatures.

We will start by providing an intuition to the reader with two standard use cases of CEs/AEs and give an overview of relevant other applications in Sect. 2. In Sect. 3, we present the (historical) background of CEs and AEs, including the current debate around their relationship. Next, we present arguments in what sense the current understanding of the relation between CEs and AEs is flawed in Sect. 4.1. In Sect. 4.2, we will argue that the notions of misclassification and maximal proximity are the central properties that distinguish CEs from AEs. Based on that, we introduce in Sect. 4.3 our more fine-grained formal definitions of CEs, AEs, and related concepts. In Sect. 5, we discuss connections between the solution approaches for finding CEs/AEs in the literature. We conclude in Sect. 6 by discussing the relevance and limitations of our work.

2 Examples and Use Cases

Before we get into the technical and conceptual details, let us look at two use cases where both CEs and AEs have been successfully deployed. This provides an intuition to the reader and will moreover serve explanatory purposes in the later sections. The first example is among the most prominent use-cases of CEs, automated lending. The second example shows one prominent use-case of AEs, image-classification of hand-written digits.

Loan Application imagine a scenario where person P wants to obtain a loan and applies for it through a bank's online portal. She has to enter several of her properties into the user-interface e.g. her age, salary, capital, number of open loans, and number of pets. The portal uses an automated, algorithmic decision system, which decides that P will not receive the loan. However, she would have liked to obtain it and therefore demands an explanation. An example of a potential CE would be:

[2] We discuss these confusions in more detail in Sect. 4.1.

> If P had a 5, 000 € p.a. higher salary and an outstanding loan less, her loan application would have been accepted.

She can use this information to guide her future actions or potentially to contest the algorithmic decision. Clearly, CEs are not restricted to that setting. If P were the model engineer instead of the customer, she could also use the explanation to raise her understanding of the model or to debug it.

Now, suppose that P wants to trick the system to get the credit. Assume the decision system was constructed from an ML algorithm, trained on historic data of the companies loan admission policy. From the data, the algorithm has learned that the number of pets is positively associated with repaying the credit and consequently the system uses the information in its decision making.[3] One potential way to trick the system with an AE in such a case could for example look as follows:

> P indicates two more pets on the application form than she actually has to obtain the loan.

P has changed a feature that the model deems causally relevant for creditworthiness but which is only spuriously correlated, thus, P has tricked the model. Moreover, she probably does not even have to prove the feature to the bank as there is often no official legal document for the ownership of e.g. fish or birds. This change allowed her to obtain the loan, even though none of her properties have changed.

Hand-Written Digits Recognition imagine a simplified scenario in which a postal service employs an image recognition algorithm. This algorithm takes as input gray-scale 28×28 pixel images and assigns them the number between 0 and 9 they depict. This procedure eases the work of the postal service a lot. Cases of errors are rare but costly, as the postal service must pay the sender 5€ if a letter or package is sent to the wrong address. Therefore, the postal service is interested in improving the algorithm.

One way of improving the system would be to generate CEs for specific instances, evaluate how useful they are, and adjust the algorithm. Such CEs can be found in the first two columns of Fig. 1. One can see e.g. that the images in the first row show that the algorithm assigns major importance to the lower-left line to distinguish between a six and a five. The postal service might derive that the algorithm already has a robust understanding of digits.

Now, assume we take the perspective of an attacker who is interested in exploiting the 5€ per error system. Such an attacker will be interested in generating AEs, put them on letters/parcels and gain money. Examples of such AEs are presented in the last column of Fig. 1. One can see e.g. that the system has problems when random dots appear around a 0 and misclassifies the input as the number 5. While the attacker will aim to accomplish many successful attacks, the postal service will

[3] Reasons for such an association in the data might be that pets are expensive and hence associated with capital/salary or that people with more pets have also kids and are therefore more reliable. The example is inspired by Ballet et al. (2019).

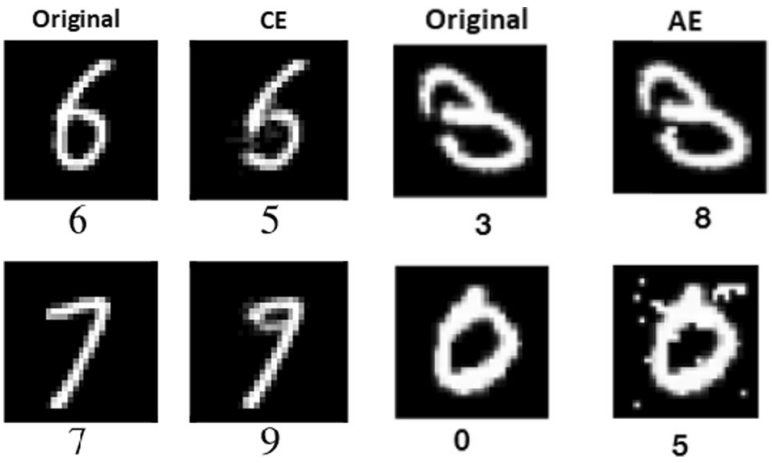


Fig. 1 The images are taken from Van Looveren and Klaise (2019) and Papernot et al. (2017). They are generated from CNNs trained on the MNIST dataset. The first an the third column depict original images from the MNIST dataset. Column two depicts the corresponding CEs and column four shows the corresponding AEs

try to limit the deceivability of its algorithm by making it more robust or excluding unrealistic outliers in classification.

The Relevance of These Use-Cases loan applications are among the most popular example use-cases in the CE literature (Wachter et al. 2017; Dandl et al. 2020; Grath et al. 2018). The example is particularly valuable as it describes a technically and ethically complex decision situation, in which explanations are a requirement. Interestingly, the lending use-case gains more and more interest also in the AE literature since it depicts the safety troubles of ML systems. Ballet et al. (2019) introduce a new notion of imperceptibility for these scenarios which got quickly picked up by others (Cartella et al. 2021; Hashemi and Fathi 2020).

Hand-Written-Digits classification is the classical use-case among all image-classification tasks. Many methods to generate AEs discuss it at least as a test case (Wang et al. 2019; Szegedy et al. 2014; Papernot et al. 2017). The feature space is comparatively small and the problem itself well studied, therefore, generating AEs is computationally cheap and conceptually informative. However, security threats cannot be as easily depicted from this use case (that is why we created the fictional scenario from above). Because of its simplicity, it has also been used as a starting point in the CE literature. The difficulty lies here in finding semantically meaningful notions of similarity for images. Three papers proposed approaches to that problem, Van Looveren and Klaise (2019) use prototypes to generate realistic CEs, Poyiadzi et al. (2020) use allowed paths, and Goyal et al. (2019) use differently classified images to identify regions that shift the classification.

Other Use Cases there are common use-cases for CEs other than loan approval, such as university applications, diabetes diagnosis (Wachter et al. 2017), adult-income prediction (Mothilal et al. 2020), or predicting student performances in law-school (Russell 2019). Most of the common use-cases focus on tabular data settings, as it is easier to make sense of CEs in these scenarios (Verma et al. 2020). Changes in semantically meaningful variables are easy to convey. Moreover, the scenarios considered often describe high-stakes decisions with an ethical dimension. There

are few non-classification, non-tabular settings in which CEs have been applied, such as image recognition (Goyal et al. 2019; Van Looveren and Klaise 2019), NLP-tasks (Akula et al. 2019), regression problems (Anjomshoae et al. 2019) and non-supervised learning settings (Olson et al. 2021).

The AE community on the other hand has largely focused on image classification tasks (Serban et al. 2020). Many AEs focus particularly on the state-of-the-art image classifiers from Google, Amazon, or Facebook (Serban et al. 2020). Well-known examples include AEs on road signs (Eykholt et al. 2018), the 3-D print of a turtle classified as a rifle (Athalye et al. 2018), and the adversarial patch, a sticker that fools image recognition software into classifying it as a toaster (Brown et al. 2017). One reason why image classifiers lie at the center of the study of AEs is that the imperceptibility of changes and the true class label are easy to define (Ballet et al. 2019). Moreover, since image recognition models focus on models like CNNs, AEs help to assess the limitations of opaque deep learning algorithms. However, there is also work on AEs in other task environments e.g. audio/video-classification (Carlini et al. 2016; Carlini and Wagner 2018; Wei et al. 2018), regression problems (Balda et al. 2019), and non-supervised learning settings (Behzadan and Munir 2017; Huang et al. 2017).

3 Background on CEs and AEs

This section provides a background on where CEs and AEs have historically come from, discusses their roles in ML, and presents the discussions about the relationship between the two. The historic background and roles of CEs/AEs provide the basis for understanding the discussions around the relationship between the two fields, which motivate our proposal.

3.1 Historic Background

History of CEs CEs have their roots in Philosophy as so-called *subjunctive counterfactual conditionals*. They describe conditionals of the form

$$\text{If } S \text{ was the case } Q \text{ would have been the case,} \tag{2}$$

where S and Q are events. Importantly, event S did not in fact occur. The truth-condition for conditional 2 is hotly debated in philosophy until today (Starr 2019). The approach that was taken up by the XAI community (Wachter et al. 2017) builds on the work of Lewis (1973) and Stalnaker (1968). In their framework, conditional 2 holds if and only if the closest possible world[4] $\omega' \in \Omega$ to the actual world $\omega \in \Omega$ in which S is the case[5] also Q is the case. The notion of similarity between possible worlds is critical in assessing a counterfactual conditional and Lewis discusses

[4] Ω denotes the set of possible worlds.

[5] S is false in ω.

similarity in more detail in Lewis (1979). He argues that between close worlds laws of nature must be preserved, widespread, diverse violations should be avoided, and facts stay congruent for maximal time. Particular facts on the other side can be changed without significantly increasing dissimilarity. Despite these specifications, Lewis himself admits that the under-specified notion of similarity between possible worlds remains the crucial weak-spot of his framework (Lewis 1983).

It is very important to keep in mind that Lewis aimed to describe causal dependence via counterfactual conditionals (Menzies and Beebee 2019). The idea is that Q' causally depends on S' if and only if, if S were not the case Q would not have been the case.[6] Even though CEs are not necessarily causal (Reutlinger 2018), the connection to causality is the main factor that underlies the explanatory force of CEs in XAI. We can see a textual CE in XAI as a true counterfactual conditional in which the antecedent describes a change in input features and the consequent a corresponding change in the classification.

Research on CEs in Psychology concerning human-to-human interaction is another root and inspiration of the discussion in XAI (Byrne 2016; Miller 2019). Humans use CEs in their daily life when they explain behavior or phenomena to each other, often in the form of a contrastive explanation highlighting the differences to the real scenario. Byrne (2019) summarized the central findings on CEs in Psychology and evaluates their relevance to XAI. She points out that people tend to create CEs that: add information rather than delete, show better rather than worse outcomes, identify relevant cause–effect relationships, and change antecedents that are exceptional, controllable, action-based, recent, and not highly improbable.

Using Lewis's account of counterfactuals for generating explanations for the decisions of ML algorithms was first proposed by Wachter et al. (2017) who also drew the connection to the philosophical/psychological tradition of CEs. They argue that CEs have three intuitive functions: *raise understanding*, *give guidance for future actions*, and *allow to contest decisions*.[7] Also, they highlighted the legal relevance of CEs and argued that they satisfy the requirements proposed in the so-called 'right to explanation' as it is defined in Recital 71 of the European General Data Protection Regulation (GDPR). This law guarantees European citizens the right to obtain an explanation in cases they are subject to the fully automated decision-making of an algorithm (Voigt and Von dem Bussche 2017).

History of AEs AEs have a less rich philosophical tradition, but instead a strong history in the robustness and reliability literature in computer science (Joseph et al. 2018). Fernandez et al. (2005) describes robustness as "the ability of a software to keep an 'acceptable' behavior [...] in spite of exceptional or unforeseen execution conditions." The reliability and robustness of computer systems have always been

[6] Interestingly, Pearl (2009) turns this story around and defines counterfactuals via causal graphs. Instead of comparing similar worlds, he directly focuses on the underlying mechanisms defined by a structural equation. However, as Woodward (2002) and Hitchcock (2001) pointed out that is a matter of interpretation as we can instead also understand Pearl's structural equations as sets of primitive counterfactuals. Also, Pearl's notion has found its way into the XAI literature in the form of algorithmic recourse (Karimi et al. 2020c, b).

[7] It is not necessarily the case that all of these functions are or can be satisfied by one CE (Russell 2019).

major concerns, especially in safety-critical applications such as health or the military sector. Critical elements can be the human interactors, hardware (e.g. sensors, hard drives, or processors), and the software. All kind of software is prone to erroneous behavior (Kizza et al. 2013), however, adversarial ML focuses particularly on the robustness of ML software.

For classical 'rule-based' software, the robustness can often be tested by formal verification (D'silva et al. 2008). This becomes more difficult if systems interact dynamically with their environment or learn from data. Statistical Learning Theory tries to extend the idea of formal verification to statistical learning methods and gives theoretical guarantees for the performance of specific model-classes (Vapnik 2013). Unfortunately, good guarantees become unattainable for very broad and powerful model-classes such as for Deep Neural Networks and learning procedures like Stochastic Gradient Descent (Goodfellow et al. 2016). What is special about the robustness of complex ML algorithms compared to others is that they are vulnerable to attacks even if common errors in model-selection have been avoided (Bishop 2006; Claeskens et al. 2008; Good and Hardin 2012). Moreover, the kind of attacks they are vulnerable to is highly unexpected, which even has led to the question of whether they learn anything meaningful at all (Szegedy et al. 2014). The study of adversarial ML is not restricted to Deep Learning but also applies to classical ML models e.g. logistic regression (Dalvi et al. 2004).

The research in adversarial ML focuses on attacks on ML models by manipulated inputs and the defenses against such attacks. An AE describes an input to a model that is deliberately designed to effectively "fool" the model into misclassifying[8] it. AEs occur even for ML algorithms with strong performances in testing-conditions. Since the changes from the original to the adversarial input are mostly *imperceptible to humans*, AEs have been compared to optical illusions tailored to ML models (Elsayed et al. 2018).

Szegedy et al. (2014) and Goodfellow et al. (2015) contributed milestones in the literature on AEs by not only providing ways to generate AEs but also attempting to explain their existence. Szegedy et al. (2014) argued that AEs live mainly in spaces of low probability in the data-manifold. Therefore, they do not appear in either the training or the test dataset. Hence, artificial neural networks (ANNs) can have a low generalization error despite the existence of AEs. Goodfellow et al. (2015) refuse this thesis and argue that AEs arise instead due to the linearity of many ML models including ANNs with semi-linear activations. Tanay and Griffin (2016) disagree and show that linearity is neither sufficient nor necessary to explain AEs. Instead, they claim that AEs lie slightly outside the real-data distribution close to tilted decision boundaries. They argue that the decision boundary is continuous outside the data-manifold and can therefore easily be crossed by AEs. A radically different view is proposed by Ilyas et al. (2019) who show that AEs arise from highly predictive but non-robust features present in the training data. Hence, AEs are a human-centered

[8] From now on, we will mainly talk about misclassification and classifying. However, this is only to simplify our language usage. AEs are not restricted to classification tasks but also work on regression problems.

phenomenon, the ML models, however, just rely on useful information in the data humans do not use.[9]

3.2 Role in ML

Due to the theoretical foundation, practical applicability, and legal significance, the CE approach was quickly adopted by the XAI community as one method to explain individual predictions of ML models to end-users (Verma et al. 2020). Nevertheless, the method remains controversial and has often been accused of giving misleading explanations (Laugel et al. 2019a; Barocas et al. 2020; Páez 2019).

The trust we have in AI systems is and will be closely linked to the extent to which adversarial attacks are possible (Toreini et al. 2020). On the negative side, AEs can cause severe damage and security threats (Eykholt et al. 2018). On the positive side, AEs can help us understand how the algorithm works (Ignatiev et al. 2019; Tomsett et al. 2018) and therefore to understand what it has actually learned (Lu et al. 2017a). AEs can even concretely improve models (Bekoulis et al. 2018; Stutz et al. 2019).

Both CEs and AEs play a great role in the ML landscape, namely for the trust people have in ML (Shin 2021; Toreini et al. 2020). CEs and AEs contribute to improving model understanding, identifying biases, and even offer methods to eliminate these biases through adversarial/counterfactual-training (Bekoulis et al. 2018; Sharma et al. 2020). However, while improving understanding and highlighting algorithmic problems is usually only a byproduct of AEs, it is the focus of CEs. The deception of a system, on the other hand, is essential for AEs, but a potential byproduct of CEs in cases where they disclose too much information about the algorithm (Sokol and Flach 2019).

3.3 The Relation Between CEs and AEs

As mentioned in Sect. 1, CEs and AEs derive from solutions to the same optimization problem 1. While the close mathematical relationship between CEs and AEs has been frequently pointed out, their exact relationship remains controversial and there are a variety of opinions on the matter we present here in more detail.

In one of the early papers on CEs, Wachter et al. (2017) note that an AE can be described as "a counterfactual by a different name" (Wachter et al. 2017, p. 852). They see one difference between counterfactuals and adversarials in the applied notion of distance arising from the misaligned aims, e.g. sparsity vs. imperceptibility. The other difference they argue for is that while counterfactuals ought to

[9] Since it is extremely controversial why AEs exist, it is also hard to defend a system against them. It is even difficult to formulate the desired property an ML model should have concerning AEs (Bastani et al. 2016; Biggio and Roli 2018). Classical verification methods have to be modified because they explode computationally in the high-dimensional input spaces we are dealing with in ML. Since defense techniques are not relevant for CEs, we will not discuss them in the present paper. We advise the interested reader to Serban et al. (2020).

describe closest possible worlds, AEs often result from 'impossible worlds' in the Lewisian sense i.e. unrealistic data-points. Additionally, they hint at methodological synergies between the two approaches, especially with respect to optimization techniques.

Browne and Swift (2020) reject the two difference makers between CEs and AEs highlighted by Wachter et al. (2017) (distance metrics, possibility of worlds) as not definitional. They argue that using the "wrong" notion of distance may favor less relevant counterfactuals, but these are still ultimately potential explanations. Moreover, they reject the claim that adversarials must describe impossible worlds by pointing out that adversarial attacks can be carried out in real-world settings. Instead, they view counterfactuals and adversarials as formally equivalent. They argue that the key difference between CEs and AEs is not mathematical, but relies on the semantic properties of the input space. They point out that: "Mathematically speaking, there is no difference between a vector of pixel values and a vector of semantically rich features" (Browne and Swift 2020, p. 6). They highlight the role of semantics in human-to-human explanation and claim that this difference makes CEs for image-data adversarials as AEs cannot be conveyed to an explainee in human-understandable terms.

Verma et al. (2020) see the terms CE and AE as non-interchangeable due to the different desiderata they must account for. They highlight tensions between the adversarial desideratum of imperceptibility and counterfactual desiderata like sparsity, closeness to the data-manifold, and actionability. According to Grath et al. (2018) CEs and AEs are similar as both are example-based approaches. They describe the distinction between CEs and AEs as the difference between flipping and explaining decisions. They remark that CEs inform about the changes, while AEs aim at hiding those. Laugel et al. (2019b) agree that the two concepts show strong mathematical similarities. However, they also point to the difference in purpose and application. They note that CEs are mainly considered in the context of low-dimensional tabular data scenarios, whereas AEs are considered in less-structured domains like image/audio data. Dandl et al. (2020) and Molnar (2019) describe AEs as special CEs with the aim of deception. Sokol and Flach (2019) discuss CEs in the context of AI safety. They make the case that CEs can disclose too much information about the model and thereby lead to AEs.

4 Defining Concepts

This section consists of three parts: (1) a critical assessment of the accounts from Sect. 3.3; (2) our conceptual proposal; (3) our formal proposal. In the first part, we will argue why none of the afore-mentioned accounts can properly explain the difference between CEs and AEs. As we will point out, one problem is that they focus on the optimization problem 1 as the defining mathematical term for CEs/AEs. Instead, we will explain why solving Eq. 1 leads to counterfactuals in tabular settings and adversarials in the image-domain. Moreover, we propose that the relation of the counterfactual/adversarial to the true label and the proximity to the original data-point present the definitional distinction between CEs and AEs. Since these

two distinguishing properties are not captured by Eq. 1 we will consequently present novel mathematical definitions of CEs/AEs in part three.

In our arguments, we assume that the reader is familiar with the ideas behind *decision boundaries, data manifolds, meaningless/unrealistic/unseen inputs, and distance metrics*. For readers who are not familiar with these concepts, we have provided a short glossary in Appendix A where we explain these concepts with an illustrative example.

4.1 Conceptual Discussion of Other Accounts

Two Names for the Same Objects Taking the optimization problem from Eq. 1 as definitional, Wachter et al. (2017) and Browne and Swift (2020) conclude that they are the same mathematical objects. To evaluate this claim, imagine a model, e.g. an image classifier that, for all inputs for which a ground truth exists, assigns exactly this ground truth. Now, consider a particular prediction of this perfect algorithm. Via solving the optimization problem in Eq. 1 we can generate counterfactuals. The CEs would be pointing to another input that receives a different assignment e.g. instead of the original image of a 3, it shows a 9 looking similar to that 3. However, the system cannot be fooled by a modified image because it is always correct. Therefore, no AEs exist in that case and none of the generated counterfactuals is an AE. The case of a perfect algorithm shows that there are models for which we can reasonably generate CEs but no AEs. Consequently, they cannot generally be the same objects with different names. This shows that while there may be some cases where a vector can be called both counterfactual and adversarial, there must be a definitional difference between the two concepts.

*The Two Differ in Aims*Verma et al. (2020) point out that the terms are not interchangeable because "while the optimization problem is similar to the one posed in counterfactual-generation, the desiderata are different" (Verma et al. 2020, p.4). By desiderata they mean additional requirements that are enforced on adversarials (like imperceptibility) or counterfactuals (e.g. sparsity, closeness to the data-manifold and feasibility. See also Sect. 5). These different desiderata are realized in the different distance metrics applied. This difference in aims corresponds to what Wachter et al. (2017) mean by claiming that AEs are not making use of appropriate distance metrics. So even though counterfactuals and adversarials share the same formal definition, they can be distinguished by their notion of distance i.e. the applied metric.

We agree with Browne and Swift (2020) that the applied distances do not indicate a definitional difference between CEs/AEs. We contend that whether the desiderata overlap or not, depends on the respective aims the user has with a CE/AE. Agents might also be interested in generating CEs to get guidance on how to deceive the system (Sokol and Flach 2019). In such cases, imperceptibility will indeed be relevant, while sparsity or closeness to the data-manifold will be less relevant. Moreover, attackers could be interested in creating realistic AEs because they are harder to detect. In such scenarios, closeness to the data-manifold or feasibility constraints are desirable properties of AEs. Also, both CEs and AEs can be relevant to better understand the model at hand and to improve it.

If the desiderata are similar, so is the mathematical approach. In such scenarios, good counterfactuals and adversarials may actually align and describe the same objects. However, a proper definitional distinction between concepts should be universal, objective, and independent of the agent's intentions. It requires necessary (and sufficient) criteria that make an object an instantiation of one object-class rather than another. The various desiderata are insufficient to account for differences between counterfactuals and adversarials in this strong sense.

*Flipping and Explaining*Grath et al. (2018) draws the distinction between CEs and AEs as the difference between explaining and flipping a decision. While CEs point to changes in a meaningful way, AEs try to hide those. We think that this is a solid observation, however, it shows a difference in presentation and not in definition. If the presentation style would be the whole difference, we would agree that CEs and AEs could mathematically be described as the same objects by a different name.

*Low vs. High-Dimensional Use Cases*Laugel et al. (2019b), Wachter et al. (2017), and Browne and Swift (2020) highlight the difference in use-cases. They argue, that while for CEs mainly low-dimensional and semantically meaningful features are used, AEs are mostly considered for high-dimensional image data with little semantic meaning of individual features. Therefore, the difference is not a difference of mathematical objects but rather a difference of semantic structure of the input space provided to generate an explanation/attack. In that sense, an AE is a CE that points to semantically non-interpretable factors.

However, as discussed in Sect. 2 the use-cases are increasingly overlapping. So, if Browne and Swift (2020) would be right that the provided semantics in the input spaces is the crucial difference, authors studying AEs in low-dimensional setups would just directly use the approaches from the CE literature instead of developing new methods. According to their argumentation, the two approaches should be equivalent for low-dimensional setups. But, what we can notice is that e.g. Ballet et al. (2019) uses expert knowledge to generate imperceptible AEs for structured data by asking for features they find irrelevant for the decision at hand. Moreover, Goyal et al. (2019) and Poyiadzi et al. (2020) manage to give, as it seems, meaningful CEs also for high-dimensional input spaces without making use of higher-level semantic concepts the model creates while Browne and Swift (2020) thought this is inevitable. These examples show that the semantic structure of the input space cannot account for a definitional distinction. Nevertheless, we agree that the difference between CEs and AEs is semantic in nature.

4.2 Our Proposal

After our critical assessment, we found that all approaches so far have failed to show definitional differences between counterfactuals and adversarials. This is not surprising bearing in mind that all of them take Eq. 1 as definitional for CEs and AEs. If one starts with the same definition for both approaches, one can either claim that counterfactuals and adversarials are identical or point to the elements within the optimization problem that differ such as the applied distances (i.e. the aims) or

the structure of the input space. However, just because two object classes contain solutions to the same optimization problem, does not mean that they are identical.[10] We propose two definitional differences between CEs and AEs that have so far been overseen. Moreover, we argue why nevertheless Eq. 1 can generate both CEs and AEs in different contexts.

Misclassification one obvious distinction that has largely been overseen by researchers is that adversarials must be necessarily misclassified while counterfactuals are agnostic in that respect. A correctly classified counterfactual is acceptable and often even desirable. On the other hand, if an adversarial were correctly classified, no one would call it an adversarial as it would provide no means to attack a target system. Consequently, misclassification is a necessary condition that any object called an adversarial must meet. This is different from the desiderata discussed above, which depend only on the goals of the agent with a CE or AE. Misclassification as a definitional distinction has been overseen since CEs and AEs can be generated by solving the same optimization problem 1. How can it be that the same optimization problem is used to generate CEs for tabular-data models and AEs for image-data models? This is the crucial question that has to be assessed. It is strongly connected to the riddle the existence of AEs poses as discussed in Sect. 3.1, therefore, our analysis bases on the ideas of Szegedy et al. (2014) and Tanay and Griffin (2016).

We must look at image-classification models to answer why solutions to Eq. 1 are mostly misclassified in that scenario. Complex image classifiers perform reasonably well on training data and highly similar inputs. In "unseen regions", on the other hand, they have to extrapolate and therefore perform worse. Since the input space is incredibly high-dimensional, the training data and therefore the data-manifold the algorithm approximates is comparably tiny. That means, there are many more meaningless, unrealistic, and unseen inputs than there are points in the training-data. The assignment of these inputs is not trustworthy and does not necessarily match the assignment of other nearby inputs. At the same time, there is usually a strongly limited number of classes that inputs are assigned to. Moreover, the training-data assigned to different classes have great distances. Hence, if we search for an input from another class but close to a given input, the probability is high that it is an input the algorithm has not seen, is unrealistic, or is meaningless and therefore where the algorithm is not reliable. Thus, the model will with high probability misclassify this input. Often these close inputs are neither unrealistic nor meaningless as thought by Wachter et al. (2017), but realistic. Completely unrealistic or meaningless inputs are at greater distance from the original input. Realistic but unseen data-points make up the dangerous AEs.

This explains why misclassified adversarials are generated in input spaces with high-dimensionality and little structure. The effect is even stronger if distances are applied that do not reflect what humans consider to be close inputs in the high-dimensional case. Minimal changes according to conceptually less-justified

[10] For example, both local maxima and minima minimize the absolute derivative of a differentiable function. Nevertheless, the two object classes can be formally distinguished.

distances break the dependencies between variables present in the real world and therefore search for inputs in regions with less training-data support. This line of thought might suggest that the main reason why mostly adversarials are obtained by Eq. 1 for image-classification is the use of distance metrics with little conceptual justification. Whether the right distance metric would yield fewer adversarials is, in our opinion, an empirical question that we cannot settle here. However, we will present our thoughts on this in Sect. 6.2.

There are several reasons why counterfactuals generated in structured, low-dimensional input spaces are not generally adversarials. First, the models are often more robust and extrapolate better in unseen regions, also because background knowledge can more easily enter the model. Second, the real-world variables have a much simpler dependence structure compared to the high-dimensional image-data case. Additionally, as distances are chosen that favor sparse rather than distributed changes, these dependencies are often preserved by the manipulations to the input vectors. Third, often additional constraints are added that make sure that the generated input stays close/within the data-manifold i.e. in regions where the model performs well (further discussions of these constraints can be found in Sect. 5.2).

Summed up, both counterfactuals and adversarials can be generated using the same method. However, that does not entail that they describe the same object class. Counterfactuals are agnostic with respect to the true label, whereas adversarials must be misclassified. From this perspective, counterfactuals could be considered the more general object-class. However, this conclusion would be drawn too early, since there is a second definitional difference.

Proximity to the Original Input additionally to misclassification, we want to highlight a second, minor distinction between counterfactuals and adversarials, which is their tolerance with respect to proximity to the original input.

Closeness to the original input is usually a benefit for adversarials to make them less perceptible. However, an adversarial can still be used to attack a system if it is a little bit more distal to x than another adversarial (Goodfellow et al. 2015). Dependent on the aim of the attacker, this might even be desirable. Adversarials with greater distance to the decision boundary transfer better between different models, are often more effective, or more meaningful (Zhang et al. 2019; Elsayed et al. 2018).

For counterfactuals on the other side, closeness to the original input plays a significant role in the causal interpretation as discussed in Sect. 3.1. Without maximal closeness, a counterfactual shows only a sufficient scenario for a different classification but not a necessary one. For example, assume we are in the loan-application setting from Sect. 2, where one point describes a maximally close counterfactual and the other a relatively close alternative input to x, both assigned to the same class. Assume moreover that the only difference between them is a change in gender from female to male. Then, even though such a change in gender would not impact the model-prediction, it would appear as a cause for the explainee receiving the alternative input. Such alternative inputs are less valuable than actual counterfactuals not only to data-subjects but also for model-developers examining the model. Thus, accepting 'close enough' but not maximally close inputs with a different classification as counterfactuals means either ignoring better CEs or admitting that the used distance is not perfectly adjusted for relevance in the given context.

Despite that difference in their tolerance with regards to proximity, we do not see this difference as equally essential as misclassification. If closeness is handled more loosely to generate "CEs", we might not gain real CEs, but still possibly relevant explanations. Thus, we do not entirely leave the category of objects. If on the other side we generate correctly classified inputs, we left the realm of attacks.

4.3 Our Definitions

As we argued, we must not take Eq. 1 as definitional for CEs/AEs. To account for the definitional differences we proposed in Sect. 4.2, we require novel definitions that include misclassification and the tolerance with respect to proximity to the original input. The definitions we will offer satisfy these requirements, offer useful conceptual extensions, and are grounded in the recent literature (e.g. Verma et al. 2020; Stepin et al. 2021 for CEs and Szegedy et al. 2014; Serban et al. 2020 for AEs). We try to be maximally inclusive to the usage of the terms in the general literature, however, due to the great number of papers on both fields (Yuan et al. 2019; Verma et al. 2020; Serban et al. 2020; Stepin et al. 2021) our framework will probably not be able to cover all usages.

Before we can define CEs and AEs, we need to know what we aim to explain or attack, namely ML models or the processes in which they are employed. We will restrict ourselves here to the highly common supervised learning setup. Moreover, we will focus on classification tasks. These restrictions have mainly the purpose to keep the analysis accessible. Many notions can be easily extended to other learning-paradigms.

Machine Learning Algorithms and Models assume we consider the relation of variables $\mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ and a (often one-dimensional) variable $\mathcal{Y}$. We can see these variables as random variables standing in a causal relation to each other. Let X and Y denote the co-domain of $\mathcal{X}$ respectively $\mathcal{Y}$. A (supervised) *ML algorithm* $\boldsymbol{\Phi}$ is a procedure that based on a set of models $\mathcal{M}$, a labeled training dataset $\mathcal{D}_{Tr} := \{(x^1, y^1), \ldots, (x^n, y^n)\}$ with $n \in \mathbb{N}$, some hyperparameters $\mathcal{H}$, an optimization method $\mathcal{O}$, and a loss function $\mathcal{L}$ outputs a model $f \in \mathcal{M}$. This procedure $\boldsymbol{\Phi}$ intuitively speaking searches for a model f in the set $\mathcal{M}$, using method $\mathcal{O}$ and hyperparameters $\mathcal{H}$, that has a low prediction loss $\mathcal{L}$ on the training dataset $\mathcal{D}_{Tr}$.

The model $f \in \mathcal{M}$ that is obtained by running the procedure $\boldsymbol{\Phi}$ on a given input is called the *machine learning model*. It can be described as a function $f : X \to Y$. This model ideally has a low bias measured by the loss function on the training dataset $\mathcal{D}_{Tr}$ and, moreover, a low generalization error on an unseen test dataset $\mathcal{D}_{Te} := \{(x^{n+1}, y^{n+1}), \ldots, (x^l, y^l)\}$ with $l > n$. That means that f does predict values of $\mathcal{Y}$ from $\mathcal{X}$ in cases it has seen the correct assignment, but also for cases that have not been part of the training dataset $\mathcal{D}_{Tr}$.

Counterfactuals and Adversarials unlike other authors, we distinguish between the mathematical objects that induce a CE/AE and the explanations/examples themselves. First, we will define the mathematical objects. For all the following

definitions, assume we consider a fixed ML model f, a particular vector[11] $x \in X$ that is mapped by f to a value $f(x) \in Y$, and a semi-metric[12] $d(\cdot, \cdot)$ on space X.

Definition We call $x' \in X$ an *alternative* to x if $f(x') \neq f(x)$.

In simple terms, x' is an alternative to x if it gets a different assignment by f.

Definition Let $\epsilon > 0$. We call x'_ϵ an *ϵ-alternative* to x if

$$d(x'_\epsilon, x) < \epsilon \text{ and } x'_\epsilon \text{ is an alternative to } x.$$

We can think of x'_ϵ as a step away from x for which we cross a decision boundary of the model but stay within a local ϵ-environment around x.

Definition We call $c_x \in X$ a *counterfactual* to x if

$$d(c_x, x) \text{ is minimal subject to } f(c_x) \neq f(x).$$

Staying in the narrative, a counterfactual describes the shortest[13] step that crosses a decision boundary. Notice that this closest vector does not have to be unique, there might exist a variety of vectors in equal distance.

A *true label* $y_{x',true} \in Y$ for a vector $x' \in X$ describes the objectively correct label that the input-vector x' should be assigned to. This ground-truth is often given by expert human evaluation. Not for all inputs there exists such a true label. The reason might be that the correct assignment is controversial even among expert evaluators or the considered input is unrealistic. Why are such unrealistic inputs relevant? As introduced above, image$(\mathcal{X}) \subseteq X$. That means that in cases where the subset-relation is strict, our model f is defined on data-points that do not realistically occur in the real world.

Definition We call a vector $x' \in X$ *misclassified* if $f(x') \neq y_{x',true}$.

A misclassification describes a mistake made by the algorithm relative to an expert-human assignment.

Definition Let $\epsilon > 0$. We call $a_{x,\epsilon} \in X$ an *adversarial* to x if

$$a_{x,\epsilon} \text{is an } \epsilon\text{-alternative and misclassified.}$$

In the literature, no clear definitional distinction is drawn between counterfactuals and adversarials. However, as we have argued in Sect. 4.2, we believe that the distinctions we have introduced are conceptually necessary. The definitions

[11] This vector x describes mostly a real-data instance.

[12] A semi-metric on a space X is a function $d : X \times X \to \mathbb{R}$ such that for all $x, x' \in X$ $d(x, x') \geq 0$, $d(x, x') = 0 \Leftrightarrow x = x'$, and $d(x, x') = d(x', x)$.

[13] With respect to $d(\cdot, \cdot)$.

of counterfactuals and adversarials differ in two aspects: the relation to the true instance label and the constraint of how close the respective data-point must be. The misclassification of adversarials enters the definition by enforcing it as an additional necessary condition. Note that this entails that only inputs for which a ground-truth exists can in our definition be called adversarials.

The second definitional difference we introduce is that counterfactuals must be maximally close data-points, while adversarials need only be within an ϵ-environment around the original input x. This relaxed condition on adversarials is introduced via defining them as ϵ-alternatives. This means, whether an input is called an adversarial or not, depends on how close the attacker requires the input to be. If the constraint is put too strong i.e. if ϵ is too small, there might not exist any adversarials within that environment. If, on the other side, the constraint-parameter is set very high, even inputs rather dissimilar to the original input can count as proper adversarials. Unlike adversarials, counterfactuals always exist as long as there exists an alternative to x. Moreover, only maximally close alternatives count as proper counterfactuals.

Especially counterfactuals, but also adversarials are often *targeted* i.e. the generated vector should not only be assigned to a different class than the original vector but to a specific *desired class*. The desired class imposes an additional relevance constraint. For counterfactuals, this may be from the perspective of the end-user who wants to get her loan application accepted rather than rejected or the model-engineer who wants to check whether the model can distinguish an input from other inputs of a specific object-class. For adversarials, this may be the desired classification from the perspective of the attacker of the system (e.g. Whatever is next to this sticker is a toaster Brown et al. 2017). In cases where a desired class exists and is imposed, we talk about *targeted* (ϵ)*-counterfactuals/ adversarials*. More formally, let $y_{des} \in Y$ with $f(x) \neq y_{des}$ denote the desired outcome of a stakeholder given such a desired outcome exists.

Definition We call an alternative $x' \in X$ to x y_{des}*-targeted* if $f(x') = y_{des}$.

The notion of targeted vectors has relevance when it comes to generating counterfactuals/adversarials (see Sect. 5). Moreover, we can see the y_{des}-targeted property as a further specification of a counterfactual/adversarial that informs about the relevant class. In the case of counterfactuals, targetedness also has definitional relevance. Not every y_{des}-targeted counterfactual is also a "normal" counterfactual. There are cases where c_x is a vector with minimal distance to x that belongs to class y_{des}, however, there still exist inputs x' closer to x than c_x that change the classification to a different class $f(x) \neq f(x') \neq y_{des}$. Consider a loan application scenario in which a poor rejected applicant does not only want to get his loan accepted but be classified as a high-credibility premium client with better conditions. In such a case, the targeted counterfactual would not be among the more realistic "normal" counterfactuals. For adversarials on the other side, every targeted adversarial is also a "normal" adversarial given we consider the same ϵ environment.

CEs and AEs so far, we have only discussed vectors living in a space X. How do we get from these vectors to explanations or attacks?

Definition

- A *contrastive explanation (CON)* is a presentation of an alternative x' in contrast to x understandable to a human agent.
- A *counterfactual explanation (CE)* is a presentation of a counterfactual c_x understandable to a human agent.
- An *adversarial example (AE)* is the depiction of an adversarial a_x.

Notice that while every counterfactual and every adversarial describes an alternative, not every CE or AE is a CON. CONs must be presented as a contrast between x' and x. Possible presentation styles for CEs/AEs include the presentation in form of an (English-)conditional of type III for tabular data, an image for visual-data, or a sound for auditory-data. For tabular data, we use the property that the input features in such scenarios are interpretable. That means they have semantic meaning and can be expressed by human language concepts.

Assume we are in such a tabular-data scenario where $x = (x_1, \dots, x_n)$ describes the original vector and $c_x = (c_{x_1}, \dots, c_{x_n})$ one of its targeted-counterfactuals. Now, consider the vector $c_x - x$. $p \leq n$ of this vector's values will be non zero. Assume $k_1, \dots, k_p$ describe the names of these non-zero entries of the vector and $e_{k_1}, \dots, e_{k_p}$ their respective values. The (contrastive) CE in this scenario would be:

> If P had a $e_{k_1}, \dots, e_{k_p}$ higher/lower value in $k_1, \dots, k_p$, she would have reached her desired classification instead of $f(x)$.

For image-data, we can use the fact that vectors in such spaces can be visualized directly in their image representation. Examples have been shown both for CEs and AEs in Sect. 2. The same holds for auditory data-inputs which can be presented as a sound.

As mentioned above, often there is not one unique counterfactual to a given vector x. Therefore, there is not one unique correct CE. Worse, often different CEs are incompatible. The fact that there are several equally "good" explanations for the same prediction is called the *Rashomon effect* (Molnar 2019). Several ways to deal with this problem have been proposed. Mothilal et al. (2020), Moore et al. (2019), Wachter et al. (2017), and Dandl et al. (2020) propose to present various CEs dependent on the specific aim of a user. However, then the question arises, how many and which ones? Others propose to select a single CE according to relevance (Fernández-Loría et al. 2020) or a quality standard set by the user, such as complexity (Sokol and Flach 2019). We think the question the Rashomon effect poses is still open to debate. AEs are unique neither. However, as AEs must not cohere, nor be necessarily presented to humans, this plays no role.

Model-Level and Real-World one distinction that is often overlooked is the difference between an explanation/attack on the model-level and the real-world.

We need to be clear about whether we want to explain/attack the model or the modeled process. Generally, the former is much easier to accomplish than the latter. We can only move from a model explanation/attack to a process explanation/attack if the model itself, and also the translation of our inputs, preserve the essential structure of the process (Molnar et al. 2020). There are two scenarios for which the distinction between the two levels is relevant: it is relevant for CEs if a user is interested in recourse to attain a desired outcome (Karimi et al. 2020c). It is relevant for AEs if an attacker aims to deceive an ML system deployed in the physical world (Kurakin et al. 2016).

To give two examples that highlight the difference between model-level and real-world explanations/attacks, we reconsider the examples from Sect. 2. The presented CE in the loan application setting was: "If P had a 5, 000 € p.a. higher salary and an outstanding loan less, her loan application would have been accepted." This explanation clearly tells us something about the employed model, namely about the assignment for a particular alternative. However, P could take this as an action recommendation in the sense that if she raises her salary and paid her outstanding loan, she will receive the loan she applied for. Unfortunately, things are not that simple in the real world. P has to work hard to raise her salary and pay her open loan, this does not happen in zero time. By the time she reaches the required threshold, she may be five years older and her loan application will be rejected again, this time due to her advanced age or because a different algorithm is now used (Venkatasubramanian and Alfano 2020). So the transfer from the model explanation to an action recommendation for recourse is not as easy.

A similar example can be shown for AEs. Consider the Hand-Written Digits Recognition Scenario from Sect. 2 where an attacker aims to money-pump the postal service. The AEs presented are clearly inputs that trick the model. However, if she now aims to make the step to a real-world fraud, she has to print them out. A bad printing, different colors, alternative background, changed angles, or the camera employed by the postal service will impact which input the model receives. Thus, the AE might not work in the postal-service hand-written digits recognition service but only in the artificial setting where we can directly manipulate the input the model receives.

For both CEs and AEs, we need to know the employment context and the required functionality in order to be clear about what level we are dealing with. The work of Karimi et al. (2020c) and Mahajan et al. (2019) on algorithmic recourse and the work of Kurakin et al. (2016), Lu et al. (2017b), and Athalye et al. (2018) AEs in the physical world have alerted the CE and AE communities to the importance of the two different levels. The two levels collapse only for artificial settings in which the model perfectly matches the process (Karimi et al. 2020c) and the interventions truly lead to improvements in the target (König et al. 2021).

Definition We say a CE/AE operates at the *real-world level* if it describes changes in $\mathcal{X}$ that result in changes in $\mathcal{Y}$. We say that a CE/AE operates at the *model-level* if it describes changes in X that result in changes in Y.

5 Generation of CEs/AEs

So far we have motivated and discussed the formal definitions of CEs/AEs. Now, we move from the definition to their generation. Again, we will focus on the connections between the two fields. Before we start, it is important to note that the generation methods for AEs do generally not guarantee success i.e. it is unclear whether the generated input vector is misclassified. Instead, misclassification is particularly in image-classification still often reached accidentally as discussed in Sect. 4.2.

5.1 General Approaches

Optimization Problem the most common approach to find CEs/AEs is to formulate and solve an optimization problem. Such a problem formulation is already present in the definition of CEs/AEs, however, this is an optimization under side conditions and therefore not easy to solve. Instead, the standard formulation as a single objective optimization problem is Eq. 1 that led to the confusions discussed in Sect. 4.1.

For both (targeted/untargeted) CEs and AEs there exist many other formulations as an optimization problem (Serban et al. 2020; Verma et al. 2020). For example, for CEs Poyiadzi et al. (2020), Kanamori et al. (2020), and Van Looveren and Klaise (2019) add additional terms to Eq. 1 encoding further desiderata (see aims and distances below), Dandl et al. (2020) instead add these desiderata by formulating a multi-objective optimizations problem, and Karimi et al. (2020a) formulate a search for the smallest intervention on the variables needed to attain a change in classification. Similar to the former formulations for CEs, there exist approaches to AEs like (Carlini and Wagner 2017; Moosavi-Dezfooli et al. 2017) which modify the objective from Eq. 1 to obtain desired properties like computational efficiency or universality of an AE. Other optimization problems also take into account transformations of background or objects and generate AEs whose classification is invariant under such transformations (Eykholt et al. 2018; Brown et al. 2017; Athalye et al. 2018).

Generative Networks a second way to generate CEs/AEs that has been fruitfully applied is the use of generative networks that generate CEs/AEs for a given input. This technique is widespread for both AEs (Goodfellow et al. 2014; Zhao et al. 2017; Yuan et al. 2019) and CEs (Mahajan et al. 2019; Van Looveren and Klaise 2019; Pawelczyk et al. 2020).

Sensitivity Analysis a third approach that is almost exclusively used by the AEs community is sensitivity analysis. Information about the gradient (Goodfellow et al. 2015; Lyu et al. 2015) or Jacobian (Papernot et al. 2016b) of the function in the specific input is used to make a step in the direction of the decision boundary to a different class. Moore et al. (2019) is the only example we are aware of who use this approach to generate CEs. One reason why such approaches have probably not been picked up in the CE-literature is that it has limited conceptual justification, e.g. with respect to minimal distance, as we discuss in Sect. 6.

5.2 Distances

All approaches to generate AEs necessitate an underlying notion of distance, mainly for the inputs space but often also for the output space. Researchers worked with a high variety of distances. Often the distances encode specific desiderata researchers want CEs/AEs to satisfy. For both fields, the question for the right distance for a given use-case is considered an open problem (Serban et al. 2020; Verma et al. 2020). Since every norm induces a metric, we will use the names of the norms and generally talk about distances.

Sparsity and Imperceptibility since explanations often need to be understandable to people with limited time and cognitive resources, it is desirable for CEs to point out only few relevant features. Therefore, distances are preferred that take into account sparsity. For adversarials on the other side, a common aim is imperceptibility. Changes from the original input to the modified input should be hard to grasp for human observers. While these desiderata often lead to conflicting notions of distance, they also can coincide. For example, the L_0 and L_1 norm have both been fruitfully been applied to generate sparse counterfactuals (Dandl et al. 2020; Wachter et al. 2017) and imperceptible AEs (Su et al. 2019; Tramer and Boneh 2019; Pawelczyk et al. 2020).

However, some distances to attain sparsity of counterfactuals have not been used to reach imperceptibility of AEs. One way by which sparsity can be guaranteed is to explicitly put a constraint on the number of features allowed to change (Kanamori et al. 2020; Ustun et al. 2019; Sokol and Flach 2019). Another is to constrain the number of actions that can be taken, but not the number of the corresponding feature changes (Karimi et al. 2020c). To attain imperceptibility of AEs, the distances are more diverse. Common examples are the L_2 (Moosavi-Dezfooli et al. 2016) and L_∞ (Goodfellow et al. 2015; Elsayed et al. 2018) norm for distributed changes which often makes AEs look identical to the input they origin from. Other norms, more inspired by human perception are the Wasserstein-distance (Wong et al. 2019), using physical parameters underlying the image formation process (Liu et al. 2018), or the Perceptual Adversarial Similarity Score (Rozsa et al. 2016).

Plausibility and Misclassification in many contexts, end-users want to use explanations for guiding their future actions. In such scenarios, CEs should not present an entirely unrealistic alternative scenario to the explainee. Instead, the recommended alternative should be within reach and if possible it should be feasible for agents to perform actions based on these alternatives. This often means that the counterfactual lies in the natural data-distribution. AEs must by definition be misclassified, which as discussed in Sect. 4.2, is often easier to reach on the edges or slightly outside the natural data-manifold. We see an antagonism between the goal of realism of CEs and the misclassification of AEs. Thus, progress in one of them (especially concerning the applied distances) can easily inform progress in the other, only with reversed sign in the optimization.

One common way to attain plausibility is to take into account the distance of the CE to the closest training-datapoint (Kanamori et al. 2020; Dandl et al. 2020; Sharma et al. 2020) or the allowed path to the counterfactual (Poyiadzi et al. 2020). Often, additional constraints are posed such that only actionable features should be

changed to avoid non-helpful recommendations (Ustun et al. 2019). Another way is to take into account the causal structure of the real-world features. If a counterfactual arises realistically from an intervention on some of these features, the corresponding CE is plausible (Mahajan et al. 2019; Karimi et al. 2020c).

To attain misclassified inputs, it is generally reasonable to search in low-probability areas of the data-manifold (Szegedy et al. 2014) or even outside of it (Tanay and Griffin 2016). Therefore, most distances for AEs do not respect the causal structure between the corresponding real-world variables. Some even act directly against the causal structure and modify only irrelevant features (Ballet et al. 2019) or, just as for CEs, put constraints on the potential changes (Cartella et al. 2021). Particularly noteworthy is the distance of Moosavi-Dezfooli et al. (2017) who encode the robustness of the flip in classification and also the work of Carlini and Wagner (2017) who compare the misclassification between different applied distances. Interestingly, it has been found that a greater distance to the given decision boundary guarantees more robustness of misclassification, hence, many do not search for minimal but only close adversarials (Zhang et al. 2019).

Contestability and Misclassification CEs should allow explainees to detect adverse or wrong decisions. If the explainee is an end-user, this could be the case if she feels judged unfairly (Kusner et al. 2017; Asher et al. 2020). On the other side, if the explainee is the model-engineer, this could mean CEs reveal bugs. Again, AEs must be misclassified. Decision-making mistakes are the common denominator of the contestability reached by CEs and misclassification provided by AEs. Various ways have been proposed to encode these aims.

Russell (2019) provide contestability by presenting a range of diverse CEs in which different features were modified. This increases the chance that some CEs are presented that provide grounds to contest the decision. Sharma et al. (2020) define protected properties like ethnicity and focus on changes in these features in their distance. Laugel et al. (2019b) discuss how standard norms like L_1 can lead to unjustified CEs since they arise from inputs outside the training-data. Hashemi and Fathi (2020) combines CEs and AEs to evaluate the weaknesses of a given model. They use both, the L_0 and L_2 norm plus focus on protected features in search for realistic but misclassified counterfactuals. In a similar vein, Ballet et al. (2019) assign importance weights to features and through these weights they define weighted L_p norm where changes in more important features have a lower weight and are therefore more likely to change in the optimization process. Cartella et al. (2021) extend their work and put additional constraints to keep the adversarials realistic but still fraudulent. Especially the last three examples show the great overlap between the goals of contestability and misclassification.

5.3 Model-Access

As we have discussed above, we do not need to define an optimization problem to generate counterfactuals or adversarials. However, different solution methods differ in the degree of model-access they need. We distinguish between black-box and white-box scenarios. In a black-box scenario, explainers/attackers can only query

the model for some inputs they provide and receive the corresponding output. In a white-box scenario, the explainer/attacker has full model access. We can further distinguish between methods that only work for a particular model-class and methods that are model-agnostic. All black-box solvers work for any model. For white-box solvers, some only need access to gradients and therefore require a differentiable model and those that are specific to a particular model-class e.g. linear models but can therefore often handle mixed-data. Interestingly, even though white-box scenarios are more realistic for explainers and black-box scenarios more commonly occur for attackers, the literature shows tendencies in the opposite directions.[14]

Many solvers rely on access to the models gradients e.g. for CEs (Wachter et al. 2017; Mothilal et al. 2020; Pawelczyk et al. 2020; Mahajan et al. 2019) or for AEs (Szegedy et al. 2014; Athalye et al. 2018; Brown et al. 2017; Ballet et al. 2019). Other solvers for CEs are model-specific and require full model-access such as mixed-integer linear program solvers (Ustun et al. 2019; Russell 2019; Kanamori et al. 2020) or solvers tailored for decision trees (Tolomei et al. 2017). For AEs some solvers require neural network feature representations (Sabour et al. 2016). However, several solvers can deal with a black-box setup. Common in both literatures are evolutionary algorithms e.g. for CEs (Sharma et al. 2020; Dandl et al. 2020) and for AEs (Guo et al. 2019; Alzantot et al. 2019; Su et al. 2019). Very prominent for AEs are also the approximation of gradients by symmetric differences (Chen et al. 2017) and the usage of surrogate models (Papernot et al. 2017). Especially the latter approach is interesting as it is based on the transferability of AEs between different models optimized for the same task.

We see that many solvers are fruitfully used in both domains. It will be seen whether surrogate model-based approaches also find their way into the CE literature. We find the use of them for CEs conceptually controversial as the faithfulness to the model is more critical for an explanation than for an attack (also see Sect. 6 for a short discussion of this point)).[15]

6 Discussion

In this paper, we discussed the relationship between CEs and AEs. We argued that the definitional difference between the two object classes consists in their relation to the true data labels (i.e. adversarials must necessarily be misclassified) and their proximity to the original data-point (i.e. counterfactuals must be maximally close to the original input). Based on this, we introduced formal definitions for the key concepts of the fields. In addition, we have highlighted similarities and differences between the two fields in terms of use cases, solution methods, and distance metrics.

[14] See Serban et al. (2020) and Verma et al. (2020) who notice the respective tendencies in their surveys. They explain this by the chances to explore more in white box settings and the computational problems of black-box attacks in high-dimensional use cases (see Sect. 2).

[15] A first approach to use a surrogate model to generate similar explanations to CEs was proposed by Guidotti et al. (2018).

6.1 Relevance

Our work adds a new viewpoint to the discussion of the relationship between CEs and AEs. Eventually, we hope that our work can form the basis for merging the two fields. Based on our arguments and the formal definitions inspired by them, adversarials can be seen as special cases of (more distal) misclassified counterfactuals. Especially when it comes to CEs for which misclassification is a desirable property, such as CEs for contesting adverse decisions, detecting bugs, or improving model-robustness, we see potential synergies. We believe that a solid conceptual discussion becomes more important as these functions of CEs are emphasized and as application domains overlap (e.g., AEs in lending, CEs for image classification).

Our work also has a clear practical relevance. The conceptual arguments for the maximal proximity of counterfactuals make clear that generating counterfactuals via sensitivity analysis, as proposed by Moore et al. (2019), or using surrogate model approaches could be problematic. In the case of sensitivity analysis, maximal proximity to the original input is not guaranteed and hence the corresponding CEs have less explanatory power. Surrogate models might not be sufficiently faithful to the original model and therefore lead to bad/misleading explanations. As we discussed, solution methods to find CEs can also generate AEs, but the reverse can be problematic.

What we have shown in terms of the current literature is that there is a large amount of overlap. We have also suggested which parts are good candidates for transfers. However, as we have made clear, such transfers of mathematical frameworks or approaches require conceptual justification. While transferring gradient-based solution techniques from the AE literature to generate counterfactuals, as proposed by Wachter et al. (2017), is conceptually unproblematic, using counterfactuals to measure the robustness of a model, as suggested by Sharma et al. (2020), will not work for tabular data scenarios.

6.2 Limitations and Open Problems

Misclassification Formalized our work points to an important weak spot of the current AE literature: misclassification is achieved more or less by accident in the image domain, but is not clearly formalized. Such a formalization of misclassification would greatly advance the merging process between CEs and AEs. It may be considered a limitation of our work that we have not provided this formalization but instead referred to the true data-labels, which are either expensive to obtain or simply unknown. Nevertheless, we want to provide a roadmap of what such a formalization might look like.

We believe that Ballet et al. (2019) made the first solid contributions to a formal representation without requiring the ground-truth data labels. In our opinion, a good candidate framework for generalizing their approach is causal modeling (Pearl 2009). If we have a true causal model, misclassification is obtained by modifying a correctly classified input sufficiently to change the classification, but in a way that

violates the causal structure. We suggest that adversarials can be viewed as small modifications in causally irrelevant features that unjustly influence the prediction.

Unfortunately, approaching the problem of misclassification from a causal modeling perspective also comes with strong requirements: we need a structural causal model or at least a causal graph. Obtaining such models is extremely difficult (Pearl 2009; Schölkopf 2019), and when dealing with conceptually lower-order features such as pixels or sounds, causal models might even be the wrong descriptive language. Still, we think that even limited causal knowledge about, e.g. parts of the causal graph or some of the structural equation, might suffice in many contexts to prove that a change in classification is unjustified. Moreover, for conceptually less-structured feature spaces, higher-order causal models (Beckers and Halpern 2019) where features such as objects are supervened by lower-order features such as pixels may provide the right level of description to define misclassification.

Distances on Unstructured Spaces in our discussion in Sect. 4.2 on misclassification, we gave reasons why most inputs that solve Eq. 1 are misclassified. We argued that theoretically poorly justified distance metrics are one of the reasons for this phenomenon. However, we did not address whether this might be the only reason for this behavior and whether this would still be the case if we had conceptually well-justified distances on high dimensional spaces with little semantics such as pixel spaces.

We believe that this is an empirical question we could not settle in this paper. The standard way for approaching it would be to move the distances from raw features such as pixels to higher-order features such as object properties. It has often been pointed out that deep-learning algorithms based on convolutional neural networks (CNNs; Goodfellow et al. 2016) automatically find semantically meaningful features in layers close to the output space (Zhang and Zhu 2018; Bau et al. 2017). For example, one could define a distance function on the feature space just before the so-called dense layer in CNNs, which is responsible for classification.

While we consider this a promising direction for future research, there are good reasons to remain skeptical. First, unfortunately, it is not so easy to assign specific semantic meaning to these high-level features, since some of them are poly-semantic and are triggered by quite different inputs (Olah et al. 2020). Distance measures on such features may therefore also be conceptually unjustified and the problem remains. Second, examples of AEs, such as those given by Szegedy et al. (2014) or Goodfellow et al. (2015), seem to show images that are almost identical to the original image. Hence, conceptually well-justified distance functions should also assign a low distance to these images, and consequently they will still be generated by solving Eq. 1. Following (Ilyas et al. 2019), we think that AEs are generated by Eq. 1 not only because we apply the wrong distance function, but also because the ML model has not really learned the robust concepts that humans use to distinguish objects.

Explanations and Deceptions we have not discussed the conceptual relationship between illusions and explanations more generally (e.g. the relation between everyday life explanations and cognitive biases or optical illusions), but have focused only on CEs/AEs in ML. In what sense can an illusion explain a phenomenon? How can an explanation lead to a deception? Is there an underlying conceptual or even cognitive connection between explaining and deceiving? We do find these questions, and

the possible embedding of our CE/AE discussion within them, intriguing. For now, however, we leave these deep and difficult philosophical/psychological questions to other researchers.

Glossary

We will shortly explain the following terms with the help of the example depicted in Fig. 2. As in Sect. 4.3 we call $f : X \to Y$ the classifier, X the input space, and Y the output space.

Decision boundary in the example, the decision boundary is described by the blue line. All inputs above the decision boundary are labeled "approved", all inputs below the blue line are labeled "rejected". Crossing the decision boundary means that a point is moved from one side of the decision boundary to the other. For example, the individual represented by the black "x" at position (1, 28) might cross the decision boundary by moving his salary up 1000€ or by buying two more pets.

More generally, we can describe a decision boundary as a hypersurface in space Y that separates one class from another. These hypersurfaces are induced by the classification model $f : X \to Y$.

Data-Manifold in our example, the green and red points lie within the data-manifold of realistic data samples. However, there is no point number or negative number of pets, so such instances would lie outside the data-manifold.

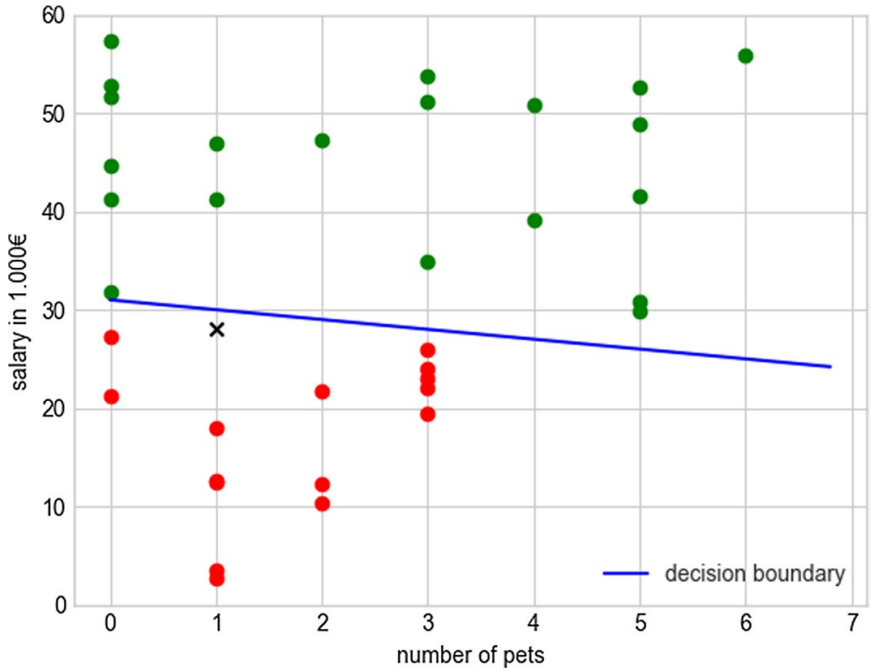


Fig. 2 This figure depicts the decision behavior of a simple classifier. It describes the scenario from Sect. 2, which is inspired by Ballet et al. (2019). The classifier uses two features, salary and number of pets, to decide whether to approve or reject a loan application. The green dots are the training data labeled as approved, the red dots are the training data labeled as rejected. The blue line describes the decision boundary of the classifier

More generally, a data-manifold describes a subset (often a hypersurface) of the spaces $X \times Y$ that arises naturally from a data-generating mechanism. A data-manifold encompasses the statistical population. The training and test data are usually a sample from this population.

Meaningless, unrealistic, or unseen inputs:

- *Meaningless* an example of a meaningless input in our scenario would be a person with a negative number of pets. It describes an input that makes no sense to us, but is contained in the space X.
- *Unrealistic* an example of an unrealistic input in our scenario would be a person with five million pets. It describes an input we can understand, but that most likely does not occur in the real world.
- *Unseen but realistic* an example of an unseen input in our scenario would be a person who earns 29,000€ and has four pets. It describes an input that may realistically occur in the real world, but was not part of the training data.

Conceptually (un-)justified distance metrics conceptually unjustified distance metrics assign small distances to inputs that are not similar from a conceptual standpoint. In our example, a distance function might assign a small distance to the points $x_1 = (0, 10)$ and $x_2 = (22, 10)$. This would make x_2, which lies far outside the data manifold and is assigned to the "approved" class by the model, a potential counterfactual for x_1. However, x_2 is highly unrealistic as 20 pets are a lot and it breaks the dependence that 20 pets are probably too expensive for an income of 10,000€ per year. This dependency problem is more severe for pixel spaces, since pixels have strong dependencies in the real world with their neighboring pixels. Moreover, an image in the form of a set of pixels represents an image of objects to humans, a fact that is difficult to account for with a metric.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by the Graduate School of Systemic Neuroscience (GSN) of the LMU Munich.

Data Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of interest The author has no conflict of interest.

References

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

Akula, A. R., Todorovic, S., Chai, J. Y., & Zhu, S. C. (2019). Natural language interaction with explainable AI models. In *CVPR workshops* (pp. 87–90).

Alzantot, M., Sharma, Y., Chakraborty, S., Zhang, H., Hsieh, C. J., & Srivastava, M. B. (2019). Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the genetic and evolutionary computation conference* (pp. 1111–1119).

Anjomshoae, S., Främling, K., & Najjar, A. (2019). Explanations of black-box model predictions by contextual importance and utility. In D. Calvaresi, A. Najjar, M. Schumacher & K. Främling (Eds.), *Explainable, transparent autonomous agents and multi-agent systems* (pp. 95–109). Springer .

Asher, N., Paul, S., & Russell, C. (2020). Adequate and fair explanations. arXiv preprint arXiv:200107578.

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In *International conference on machine learning, PMLR* (pp. 284–293).

Balda, E. R., Behboodi, A., & Mathar, R. (2019). Perturbation analysis of learning algorithms: Generation of adversarial examples from classification to regression. *IEEE Transactions on Signal Processing, 67*(23), 6078–6091.

Ballet, V., Renard, X., Aigrain, J., Laugel, T., Frossard, P., & Detyniecki, M. (2019). Imperceptible adversarial attacks on tabular data. arXiv preprint arXiv:191103274.

Barocas, S., Selbst, A.D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAT* '20, p 80–89, https://doi.org/10.1145/3351095.3372830

Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A. V., & Criminisi, A. (2016). Measuring neural net robustness with constraints. In *Proceedings of the 30th international conference on neural information processing systems* (pp. 2621–2629).

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6541–6549).

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, pp. 2678–2685).

Behzadan, V., & Munir, A. (2017). Vulnerability of deep reinforcement learning to policy induction attacks. In *International conference on machine learning and data mining in pattern recognition* (pp. 262–275). Springer.

Bekoulis, G., Deleu, J., Demeester, T., & Develder, C. (2018). Adversarial training for multi-context joint entity and relation extraction. arXiv preprint arXiv:180806876.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84,* 317–331.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Brown, T. B., Mané, D., Roy, A., Abadi, M., & Gilmer, J. (2017). Adversarial patch. arXiv preprint arXiv:171209665.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:200514165.

Browne, K., & Swift, B. (2020). Semantics and explanation: Why counterfactual explanations produce adversarial examples in deep neural networks. arXiv:2012.10076.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data and Society, 3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512.

Byrne, R. M. (2016). Counterfactual thought. *Annual Review of Psychology, 67,* 135–157.

Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning. In *IJCAI* (pp. 6276–6282).

Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy* (pp. 39–57). IEEE.

Carlini, N., & Wagner, D. (2018). Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)* (pp. 1–7). IEEE.

Carlini, N., Mishra, P., Vaidya, T., Zhang, Y., Sherr, M., Shields, C., Wagner, D., & Zhou, W. (2016). Hidden voice commands. In *25th USENIX security symposium (USENIX security 16)* (pp. 513–530).

Cartella, F., Anunciacao, O., Funabiki, Y., Yamaguchi, D., Akishita, T., & Elshocht, O. (2021). Adversarial attacks for tabular data: Application to fraud detection and imbalanced data. arXiv preprint arXiv:210108030.

Chen, P. Y., Zhang, H., Sharma, Y., Yi, J., & Hsieh, C. J. (2017). ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security* (pp. 15–26).

Claeskens, G., Hjort, N. L., et al. (2008). *Model selection and model averaging*. Cambridge Books. https://doi.org/10.1017/CBO9780511790485.

Dalvi, N., Domingos, P., Sanghai, S., & Verma, D. (2004). Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 99–108).

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In T. Bäck, M. Preuss, A. Deutz, H. Wang, C. Doerr, M. Emmerich & H. Trautmann (Eds.), *Parallel problem solving from nature—PPSN XVI* (pp. 448–469). Springer.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. arXiv preprint arXiv:200611371.

Dong, Y., Su, H., Zhu, J., & Bao, F. (2017). Towards interpretable deep neural networks by leveraging adversarial examples. arXiv preprint arXiv:170805493.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:170208608.

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE.

D'Silva, V., Kroening, D., & Weissenbacher, G. (2008). A survey of automated techniques for formal software verification. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 27*(7), 1165–1178.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. In *Advances in neural information processing systems* (pp. 3910–3920).

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1625–1634).

Fernandez, J. C., Mounier, L., & Pachon, C. (2005). A model-based approach for robustness testing. In *IFIP international conference on testing of communicating systems* (pp. 333–348). Springer.

Fernández-Loría, C., Provost, F., & Han, X. (2020). Explaining data-driven decisions made by AI systems: The counterfactual approach. arXiv:2001.07417.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81. http://jmlr.org/papers/v20/18-760.html.

Friedman, J. H., et al. (1991). Multivariate adaptive regression splines. *The Annals of Statistics, 19*(1), 1–67. https://doi.org/10.1214/aos/1176347963.

Good, P. I., & Hardin, J. W. (2012). *Common errors in statistics (and how to avoid them)*. Wiley. https://doi.org/10.1002/9781118360125.

Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *International conference on learning representations*. arXiv:1412.6572.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. arXiv preprint arXiv:14062661.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning, PMLR, proceedings of machine learning research* (Vol. 97, pp. 2376–2384). http://proceedings.mlr.press/v97/goyal19a.html.

Grath, R. M., Costabello, L., Van, C. L., Sweeney, P., Kamiab, F., Shen, Z., & Lecue, F. (2018). Interpretable credit application predictions with counterfactual explanations. arXiv preprint arXiv:181105245.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. arXiv preprint arXiv:180510820.

Guo, C., Gardner, J. R., You, Y., Wilson, A. G., & Weinberger, K. Q. (2019). Simple black-box adversarial attacks. arXiv preprint arXiv:190507121.

Hashemi, M., & Fathi, A. (2020). Permuteattack: Counterfactual explanation of machine learning credit scorecards. arXiv:2008.10138.

Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy, 98*(6), 273–299.

Huang, S., Papernot, N., Goodfellow, I., Duan, Y., & Abbeel, P. (2017). Adversarial attacks on neural network policies. arXiv preprint arXiv:170202284.

Hutson, M. (2018). Ai researchers allege that machine learning is alchemy. *Science, 360*(6388), 861.

Ignatiev, A., Narodytska, N., & Marques-Silva, J. (2019). On relating explanations and adversarial examples. In *Advances in neural information processing systems* (pp. 15883–15893).

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Madry, A. (2019). Adversarial examples are not bugs, they are features. In *Advances in neural information processing systems* (pp. 125–136).

Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. (2018). *Adversarial machine learning*. Cambridge University Press.

Kanamori, K., Takagi, T., Kobayashi, K., & Arimura, H. (2020). DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20. International joint conferences on Artificial Intelligence Organization* (pp. 2855–2862).

Karimi, A. H., Barthe, G., Balle, B., & Valera, I. (2020a). Model-agnostic counterfactual explanations for consequential decisions. In *International conference on artificial intelligence and statistics, PMLR* (pp. 895–905).

Karimi, A. H., Barthe, G., Schölkopf, B., & Valera, I. (2020b). A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. arXiv preprint arXiv:201004050.

Karimi, A. H., Schölkopf, B., & Valera, I. (2020c). Algorithmic recourse: From counterfactual explanations to interventions. In *37th International conference on machine learning (ICML)*.

Kizza, J. M., & Kizza, W. (2013). *Guide to computer network security*. Springer.

König, G., Freiesleben, T., & Grosse-Wentrup, M. (2021). A causal perspective on meaningful and robust algorithmic recourse. arXiv preprint arXiv:210707853.

Kurakin, A., Goodfellow, I., & Bengio, S., et al. (2016). Adversarial examples in the physical world

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Advances in neural information processing systems* (pp. 4066–4076).

Laugel, T., Lesot, M. J., Marsala, C., Renard, X., & Detyniecki, M. (2019a). The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence, IJCAI-19, international joint conferences on Artificial Intelligence Organization* (pp. 2801–2807). https://doi.org/10.24963/ijcai.2019/388.

Laugel, T., Lesot, M. J., Marsala, C., Renard, X., & Detyniecki, M. (2019b). Unjustified classification regions and counterfactual explanations in machine learning. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 37–54). Springer.

Leviathan, Y., & Matias, Y. (2018). Google duplex: An AI system for accomplishing real-world tasks over the phone. https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs, 13*(4), 455–476.

Lewis, D. (1983). *Philosophical papers* (Vol. I). Oxford University Press.

Lewis, D. K. (1973). *Counterfactuals*. Blackwell.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57.

Liu, H. T. D., Tao, M., Li, C. L., Nowrouzezahrai, D., & Jacobson, A. (2018). Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer. arXiv preprint arXiv:180802651.

Lu, J., Issaranon, T., & Forsyth, D. (2017a). SafetyNet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE international conference on computer vision* (pp. 446–454).

Lu, J., Sibai, H., Fabry, E., & Forsyth, D. (2017b). No need to worry about adversarial examples in object detection in autonomous vehicles. arXiv preprint arXiv:170703501.

Lyu, C., Huang, K., & Liang, H. N. (2015). A unified gradient regularization family for adversarial examples. In *2015 IEEE international conference on data mining* (pp. 301–309). IEEE.

Mahajan, D., Tan, C., & Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv preprint arXiv:191203277.

Menzies, P., & Beebee, H. (2019). Menzies, P., & Beebee, H. (2019). Counterfactual theories of causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* winter 2019 edition. Metaphysics Research Lab, Stanford University.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267,* 1–38.

Molnar, C. (2019). 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. (2020). Pitfalls to avoid when interpreting machine learning models. arXiv:2007.04131.

Moore, J., Hammerla, N., & Watkins, C. (2019). Explaining deep learning models with constrained adversarial examples. In *Pacific Rim international conference on artificial intelligence* (pp. 43–56). Springer.

Moosavi-Dezfooli, S. M., Fawzi, A., & Frossard, P. (2016). DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2574–2582).

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765–1773).

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the ACM conference on fairness, accountability, and transparency*.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. Distill 5(3):e00024–001

Olson, M. L., Khanna, R., Neal, L., Li, F., & Wong, W. K. (2021). Counterfactual state explanations for reinforcement learning agents via generative deep learning. *Artificial Intelligence, 295,* 103455.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*(3), 441–459.

Papernot, N., McDaniel, P., & Goodfellow, I. (2016a). Transferability in machine learning: From phenomena to black-box attacks using adversarial samples. arXiv preprint arXiv:160507277.

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016b). The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)* (pp. 372–387). IEEE.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security* (pp. 506–519).

Pawelczyk, M., Broelemann, K., & Kasneci, G. (2020). Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of the web conference*, 2020 (pp. 3126–3132).

Pearl, J. (2009). *Causality*. Cambridge University Press.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 344–350).

Reutlinger, A. (2018). Extending the counterfactual theory of explanation. In *Explanation beyond causation: Philosophical perspectives on non-causal explanations* (pp. 74–95). Oxford University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp .1135–1144). ACM. https://doi.org/10.1145/2939672.2939778.

Rozsa, A., Rudd, E. M., & Boult, T. E. (2016). Adversarial diversity and hard positive generation. In *Proceedings of the IEEE conference on computer vision and recognition workshops* (pp. 25–32).

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Russell, C. (2019). Efficient search for diverse coherent explanations. In *Proceedings of the conference on fairness, accountability, and transparency, FAT* '19*, New York, NY, USA (pp. 20–28). Association for Computing Machinery. https://doi.org/10.1145/3287560.3287569.

Sabour, S., Cao, Y., Faghri, F., & Fleet, D. J. (2016). Adversarial manipulation of deep representations. In Y. Bengio & Y. LeCun (Eds.), *4th International conference on learning representations, ICLR 2016*, San Juan, Puerto Rico, May 2–4, 2016, Conference Track Proceedings. arXiv:1511.05122.

Schölkopf, B. (2019). Causality for machine learning. arXiv preprint arXiv:191110500.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature, 577*(7792), 706–710.

Serban, A., Poll, E., & Visser, J. (2020). Adversarial examples on object recognition: A comprehensive survey. *ACM Computing Surveys (CSUR), 53*(3), 1–38.

Sharma, S., Henderson, J., & Ghosh, J. (2020). CERTIFAI: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society*. https://doi.org/10.1145/3375627.3375812.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human–Computer Studies, 146,* 102551.

Sokol, K., & Flach, P. A. (2019). Counterfactual explanations of machine learning predictions: Opportunities and challenges for AI safety. In *Proceedings of the AAAI workshop on artificial intelligence safety*.

Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., & Kohno, T. (2018). Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*.

Stalnaker, R. C. (1968). A theory of conditionals. In *IFS* (pp. 41–55). Springer.

Starr, W. (2019). Counterfactuals. In: E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.

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.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems, 41*(3), 647–665. https://doi.org/10.1007/s10115-013-0679-x.

Stutz, D., Hein, M., & Schiele, B. (2019) Confidence-calibrated adversarial training: Generalizing to unseen attacks. arXiv preprint arXiv:191006259.

Su, J., Vargas, D. V., & Sakurai, K. (2019). One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation, 23*(5), 828–841.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In *International conference on learning representations*. arXiv:1312.6199.

Tanay, T., & Griffin, L. (2016). A boundary tilting perspective on the phenomenon of adversarial examples. arXiv preprint arXiv:160807690.

Tolomei, G., Silvestri, F., Haines, A., & Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 465–474).

Tomsett, R., Widdicombe, A., Xing, T., Chakraborty, S., Julier, S., Gurram, P., Rao, R., & Srivastava, M. (2018). Why the failure? How adversarial examples can provide insights for interpretable machine learning. In *21st International conference on information fusion (FUSION)* (pp. 838–845). IEEE.

Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C. G., & van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 272–283).

Tramer, F., & Boneh, D. (2019). Adversarial training and robustness for multiple perturbations. arXiv preprint arXiv:190413000.

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 10–19).

Van Looveren, A., & Klaise, J. (2019). Interpretable counterfactual explanations guided by prototypes. arXiv preprint arXiv:190702584.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer.

Venkatasubramanian, S., & Alfano, M. (2020). The philosophical basis of algorithmic recourse. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 284–293).
Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual explanations for machine learning: A review. arXiv preprint arXiv:201010596.
Voigt, P., & Von dem Bussche, A. (2017). *The EU general data protection regulation (GDPR). A practical guide* (1st ed.). Springer 10:3152676.
Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv JL & Tech, 31,* 841.
Wang, X., He, K., & Hopcroft, J.E. (2019). AT-GAN: A generative attack model for adversarial transferring on generative adversarial nets. *CoRR*. arXiv:abs/190407793.
Wei, X., Liang, S., Chen, N., & Cao, X. (2018). Transferable adversarial attacks for image and video object detection. arXiv preprint arXiv:181112641.
Wong, E., Schmidt, F., & Kolter, Z. (2019). Wasserstein adversarial examples via projected Sinkhorn iterations. In *International conference on machine learning, PMLR* (pp. 6808–6817).
Woodward, J. (2002). What is a mechanism? A counterfactual account. *Philosophy of Science, 69*(S3), S366–S377.
Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems, 30*(9), 2805–2824.
Zhang, H., Chen, H., Song, Z., Boning, D., Dhillon, I.S., & Hsieh, C.J. (2019). The limitations of adversarial training and the blind-spot attack. arXiv preprint arXiv:190104684
Zhang, Q., & Zhu, S. C. (2018). Visual interpretability for deep learning: A survey. arXiv preprint arXiv:180200614.
Zhao, Z., Dua, D., & Singh, S. (2017). Generating natural adversarial examples. arXiv preprint arXiv:171011342.

Chapter 5

Paper IV: Improvement-focused Causal Recourse (ICR)

König, G., **Freiesleben, T.** and Grosse-Wendrup, M. (forthcoming 2023). Improvement-focused Causal Recourse (ICR). *Proceedings of the AAAI Conference on Artificial Intelligence 2023*.

Author contributions:

G.K. had the initial idea, G.K. and **T.F.** developed the story and the philosophical foundation together. G.K. wrote large parts of the paper, developed the proofs and wrote the code. **T.F.** came up with the running example, lead writing Section 4, checked the proofs and contributed to Sections 1, 2, 9 and 10. **All authors** helped to revise and proofread the paper.

Improvement-Focused Causal Recourse (ICR)

Gunnar König, [1,2] **Timo Freiesleben,** [4,5,6] **Moritz Grosse-Wentrup** [2,3]
[1] Munich Center for Machine Learning (MCML), LMU Munich
[2] Research Group Neuroinformatics, University of Vienna
[3] Data Science @ Uni Vienna, Vienna CogSciHub
[4] Munich Center for Mathematical Philosophy (MCMP), LMU Munich
[5] Cluster of Excellence Machine Learning, University of Tübingen
[6] Graduate School of Systemic Neurosciences, LMU Munich
g.koenig.edu@pm.me

Abstract

Algorithmic recourse recommendations, such as Karimi et al.'s (2021) causal recourse (CR), inform stakeholders of how to act to revert unfavorable decisions. However, there are actions that lead to acceptance (i.e., revert the model's decision) but do not lead to improvement (i.e., may not revert the underlying real-world state). To recommend such actions is to recommend fooling the predictor. We introduce a novel method, Improvement-Focused Causal Recourse (ICR), which involves a conceptual shift: Firstly, we require ICR recommendations to guide towards improvement. Secondly, we do not tailor the recommendations to be accepted by a specific predictor. Instead, we leverage causal knowledge to design decision systems that predict accurately pre- and post-recourse. As a result, improvement guarantees translate into acceptance guarantees. We demonstrate that given correct causal knowledge ICR, in contrast to existing approaches, guides towards both acceptance and improvement.

1 Introduction

Predictive systems are increasingly deployed for high-stakes decisions, for instance in hiring (Raghavan et al. 2020), judicial systems (Zeng, Ustun, and Rudin 2017), or when distributing medical resources (Obermeyer and Mullainathan 2019). A range of work (Wachter, Mittelstadt, and Russell 2017; Ustun, Spangher, and Liu 2019; Karimi, Schölkopf, and Valera 2021) develops tools that offer individuals possibilities for so-called algorithmic recourse (i.e. actions that revert unfavorable decisions). Joining previous work in the field, we distinguish between reverting the model's prediction $\hat{Y}$ (acceptance) and reverting the underlying real-world state Y (improvement) and argue that recourse should lead to acceptance *and improvement* (Ustun, Spangher, and Liu 2019; Barocas, Selbst, and Raghavan 2020). Existing methods, such as counterfactual explanations (CE; Wachter, Mittelstadt, and Russell (2017)) or causal recourse (CR; Karimi, Schölkopf, and Valera (2021)), ignore the underlying real-world state and only optimize for acceptance. Since ML models are not designed to predict accurately in interventional environments (i.e. environments where actions have changed the data distribution), acceptance does not necessarily imply improvement.

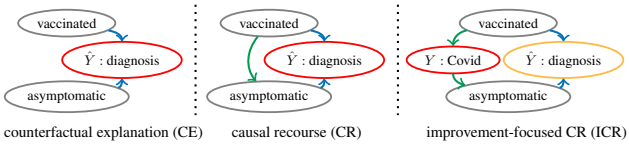


Figure 1: Directed Acyclic Graph (DAG) illustrating the perspectives of counterfactual explanations (CE, left) and causal recourse (CR, center) in contrast to improvement-focused recourse (ICR, right). Green edges represent real-world causal links, blue edges the prediction model. Gray nodes represent covariates, and the red (yellow) node the primary (secondary) recourse target. CR respects causal relationships, but solely between input features; only ICR takes the target Y into account. While CE and CR aim to revert the prediction $\hat{Y}$, ICR aims to revert the target Y.

Let us consider an example. We aim to predict whether hospital visitors without test certificate are infected with Covid to restrict access to tested and low-risk individuals. Here, the model's *prediction* $\hat{Y}$ represents whether someone is classified to be infected, whereas the *target* Y represents whether someone is actually infected. Target and prediction differ in how they are affected by actions: Intervening on the *symptoms* may change the model's diagnosis $\hat{Y}$, but will not affect whether someone is infected (Y).
Both counterfactual explanations (CE) and causal recourse (CR) only target $\hat{Y}$ (Figure 1). Therefore, CE and CR may suggest altering the *symptoms* (e.g., by taking cough drops) and thereby may recommend to *game* the predictor: Although the intervention leads to acceptance, the actual Covid risk Y is not improved.[1]
One may argue that this is an issue of the prediction model and may adapt the predictor to make gaming less lucrative than improvement (Miller, Milli, and Hardt 2020). However, such adaptions may come at the cost of predictive performance – even in light of causal knowledge. The reason is that gameable variables can be highly predictive (Shavit, Edelman, and Axelrod 2020); In our example, the model's reliance on the symptom state would need to be reduced. Thus, we tackle the problem by adjusting the explanation instead.

[1] In E.1, the case is formally demonstrated.

Contributions We present improvement-focused causal recourse (ICR), the first recourse method that targets improvement instead of acceptance. Since estimating the effects of actions is a causal problem, causal knowledge is required. More specifically, we show how to exploit either knowledge of the structural causal model (SCMs) or the causal graph to guide towards improvement (Section 5). On a conceptual level we argue that the individual's improvement options should not be limited by an acceptance constraint (Section 4). In order to nevertheless yield acceptance, we show how to exploit said causal knowledge to design post-recourse decision systems that in expectation recognize improvement (Section 6), such that improvement guarantees translate into acceptance guarantees (Section 7). On synthetic and semi-synthetic data, we demonstrate that ICR, in contrast to existing approaches, leads to improvement and acceptance (Section 8).

2 Related Work

Constrastive Explanations Contrastive explanations explain decisions by contrasting them with alternative decision scenarios (Karimi et al. 2020a; Stepin et al. 2021); a well known example are counterfactual explanations (CE) that highlight the minimal feature changes required to revert the decision of a predictor $\hat{f}(x)$ (Wachter, Mittelstadt, and Russell 2017; Dandl et al. 2020). However, CEs are ignorant of causal dependencies in the data and therefore in general fail to guide action (Karimi, Schölkopf, and Valera 2021). In contrast, the causal recourse (CR) framework by Karimi et al. (2022) takes the causal dependencies between covariates into account: More specifically, Karimi et al. (2022) use structural causal models or causal graphs to guide individuals towards acceptance.[2] The importance of improvement was discussed before (Ustun, Spangher, and Liu 2019; Barocas, Selbst, and Raghavan 2020), but as of now no improvement-focused recourse method was proposed.

Strategic Classification The related field of strategic modeling investigates how the prediction mechanism incentivizes rational agents (Hardt et al. 2016; Tsirtsis and Gomez Rodriguez 2020). A range of work (Bechavod et al. 2020; Chen, Wang, and Liu 2020; Miller, Milli, and Hardt 2020) thereby distinguishes models that incentivize *gaming* (i.e., interventions that affect the prediction $\hat{Y}$ but not the underlying target Y in the desired way) and *improvement* (i.e., actions that also yield the desired change in Y). Strategic modeling is concerned with adapting the model, where except for special cases the following three goals are in conflict: incentivizing improvement, predictive accuracy, and retrieving the true underlying mechanism (Shavit, Edelman, and Axelrod 2020).

3 Background and Notation

Prediction model We assume binary probabilistic predictors and cross-entropy loss, such that the optimal score function $h^*(x)$ models the conditional probability $P(Y = 1|X = x)$, which we abbreviate as $p(y|x)$. We denote the estimated score function as $\hat{h}(x)$, which can be transformed into the binary decision function $\hat{f}(x) := [\hat{h}(x) \geq t]$ via the decision threshold t.

Causal data model We model the data generating process using a structural causal model (SCM) $\mathcal{M} \in \Pi$ (Pearl 2009; Peters, Janzing, and Schölkopf 2017). The model $\mathcal{M} = \langle X, U, \mathbb{F} \rangle$ consists of the endogenous variables $X \in \mathcal{X}$, the mutually independent exogenous variables $U \in \mathcal{U}$, and structural equations $\mathbb{F} : \mathcal{U} \to \mathcal{X}$. Each structural equation f_j specifies how X_j is determined by its endogenous causes and the corresponding exogenous variable U_j. The SCM entails a directed graph $\mathcal{G}$, where variables are connected to their direct effects via a directed edge.

The index set of endogenous variables is denoted as D. The parent indexes of node j are referred to as $pa(j)$ and the children indexes as $ch(j)$. We refer to the respective variables as $X_{pa(j)}$. We write $X_{pa(j)}$ to denote all parents excluding Y and $(X, Y)_{pa(j)}$ to denote all parents including Y. All ascendant indexes of a set S are denoted as $asc(S)$, its complement as $nasc(S)$, all descendant indexes as $d(S)$, and its complement as $nd(S)$.

SCMs allow to answer causal questions. This means that they cannot only be used to describe (conditional) distributions (observation, rung 1 on Pearl's ladder of causation (Pearl 2009)), but can also be used to predict the (average) effect of actions $do(x)$ (intervention, rung 2) and imagine the results of alternative actions in light of factual observation $(x, y)^F$ (counterfactuals, rung 3).

As such, we model actions as structural interventions $a : \Pi \to \Pi$, which can be constructed as $do(a) = do(\{X_i := \theta_i\}_{i \in I})$, where I is the index set of features to be intervened upon. A model of the interventional distribution can be obtained by fixing the intervened upon values to θ_I (e.g. by replacing the structural equation $f_I := \theta_I$). Counterfactuals can be computed in three steps (Pearl 2009): First, the factual distribution of exogenous variables U given the factual observation of the endogenous variables x^F is inferred (*abduction*) (i.e., $P(U_j|X^F)$). Second, the structural interventions corresponding to $do(a)$ are performed (*action*). Finally, we can sample from the counterfactual distribution $P(X^{SCF}|X = x^F, do(a))$ using the abducted noise and the intervened-upon structural equations (*prediction*).

4 The Two Tales of Contrastive Explanations

In the introduction we have demonstrated that CE and CR may suggest to game the predictor (i.e. guide towards acceptance without improvement). To tackle the issue, we will introduce a new explanation technique called improvement-focused causal recourse (ICR) in Section 5.

In this section we lay the conceptual justification for our method. More specifically, we argue that for recourse the acceptance constraint of CR should be *replaced* by an improvement constraint. Therefore, we first recall that a multitude of goals may be pursued with contrastive explanations (Wachter, Mittelstadt, and Russell 2017) and separate two purposes of contrastive explanations: *contestability of algorithmic decisions* and *actionable recourse*. We then argue

[2] For the interested reader, we formally introduce CR in our notation in A.4.

that improvement is an essential requirement for recourse and that the individual's options for improvement should not be limited by acceptance constraints.

Contestability and recourse are distinct goals. *Contestability* is concerned with the question of whether the algorithmic decision is correct according to common sense, moral or legal standards. Explanations may help model authorities to detect violations of such standards or enable explainees to contest unfavorable decisions (Wachter, Mittelstadt, and Russell 2017; Freiesleben 2021). Explanations that aim to enable contestability must reflect the model's rationale for an algorithmic decision. *Recourse recommendations* on the other hand need to satisfy various constraints unrelated to the model, such as causal links between variables (Karimi, Schölkopf, and Valera 2021) or their actionability (Ustun, Spangher, and Liu 2019). Consequently, explanations geared to contest are more complete and true to the model while recourse recommendations are more selective and true to the underlying process.[3] We believe that the selectivity and reliance of recourse recommendations on factors besides the model itself is not a limitation but an indispensable condition for making explanations more relevant to the explainee.

In the context of recourse, improvement is desirable for model authority and explainee. We consider improvement to be an important normative requirement for recourse, both with respect to explainee and model authority. Valuable recourse recommendations enable explainees to plan and act; thus, such recommendations must either provide indefinite validity or a clear expiration date (Wachter, Mittelstadt, and Russell 2017; Barocas, Selbst, and Raghavan 2020; Venkatasubramanian and Alfano 2020). Problematically, when model authorities give guarantees for non-improving recourse, this constitutes a binding commitment to misclassification. However, if model authorities do not provide recourse guarantees over time, this diminishes the value of recourse recommendations to explainees. They might invest effort into non-improving actions that ultimately do not even lead to acceptance because the classifier changed.[4] In contrast, improvement-focused recourse is honored by any accurate classifier. We conclude that, given these advantages for both model authority and explainee, recourse recommendations should help to improve the underlying target Y.[5]

Improvement should come first, acceptance second. Taken that we constrain the optimization on improvement, how to guarantee acceptance remains an open question. One approach would be to constrain the optimization on both improvement and acceptance. However, a restriction on acceptance is either redundant or, from our moral standpoint, questionable: If improvement already implies acceptance, the constraint is redundant. In the remaining cases, we can predict improvement with the available causal knowledge but would withhold these (potentially less costly) improvement options because of the limitations of the observational predictor. To ensure that acceptance ensues improvement, we instead suggest to exploit the assumed causal knowledge for accurate post-recourse prediction (Section 6), such that acceptance guarantees can be made (Section 7).

[3]We do not claim that recourse and contestability always diverge, we only describe a difference in focus. If contesting is successful it may even provide an alternative route towards recourse.

[4]For instance, in the introductory example, an intervention on the symptom state would only be honored by a refit of the model on pre- and post-recourse data for the small percentage of individuals who were already vaccinated, as documented in more detail in E.1. Also, gaming actions may not be robust concerning model multiplicity, as seen in the experiments (Section 8).

[5]We do not claim that gaming is necessarily bad; it may be justified when predictors perform morally questionable tasks.

5 Improvement-Focused Causal Recourse (ICR)

We continue with the formal introduction of ICR, an explanation technique that targets improvement ($Y = 1$) instead of acceptance ($\hat{Y} = 1$). Therefore we first define the improvement confidence γ, which can be optimized to yield ICR. Like previous work in the field (Karimi et al. 2020b), we distinguish two settings: In the first setting, knowledge of the SCM can be assumed, such that we can leverage structural counterfactuals (rung 3 on Pearl's ladder of causation) to introduce the individualized improvement confidence γ^{ind}. In the second setting only the causal graph is known, which we exploit to propose the subpopulation-based improvement confidence γ^{sub} (rung 2).

Individualized improvement confidence For the individualized improvement confidence γ^{ind} we exploit knowledge of a SCM. SCMs can be used to answer counterfactual questions (rung 3). In contrast to rung-2-predictions, counterfactuals are tailored to the individual and their situation (Pearl 2009): They ask what would have been if one had acted differently and thereby exploit the individual's factual observation. Given unchanged circumstances, counterfactuals can be seen as individualized causal effect predictions.
In contrast to existing SCM-based recourse techniques (Karimi et al. 2022) we include both the prediction $\hat{Y}$ and the target variable Y as separate variables in the SCM. As a result, the SCM can be used not only to model the individualized probability of acceptance, but also the individualized probability of improvement.

Definition 1 (Individualized improvement confidence)**.** *For pre-recourse observation x^{pre} and action a we define the individualized improvement confidence as*

$$\gamma^{ind}(a) = \gamma(a, x^{pre}) := P(Y^{post} = 1 | do(a), x^{pre}).$$

Since the pre-recourse (factual) target Y cannot be observed, standard counterfactual prediction cannot be applied directly. However, we can regard the distribution as a mixture with two components, one for each possible state of Y. We can estimate the mixing weights using h^* and each component using standard counterfactual prediction. Details including pseudocode are provided in B.1.

Subpopulation-based improvement confidence For the estimation of the individualized improvement confidence γ^{ind} knowledge of the SCM is required. If the SCM is not specified, but the causal graph is known instead and there are no unobserved confounders (causal sufficiency), we can still estimate the effect of interventions (rung 2).
In contrast to counterfactual distributions (rung 3), interventional distributions describe the whole population and therefore provide limited insight into the effects of actions on specific individuals. Building on Karimi et al. (2020b), we thus narrow the population down to a subpopulation of similar individuals, for which we then estimate the subpopulation-based causal effect. More specifically, we consider individuals to belong to the same subgroup if the variables that are not affected by the intervention take the same values. For action a, we define the subgroup characteristics as $G_a := nd(I_a)$ (i.e., the non-descendants of the intervened-upon variables in the causal graph).[6] More formally, we define the subpopulation-based improvement confidence γ^{sub} as the probability of Y taking the favorable outcome in the subgroup of similar individuals (Definition 2).

Definition 2 (Subpopulation-based improvement confidence)**.** *Let a be an action that potentially affects Y, i.e. $I_a \cap asc(Y) \neq \emptyset$.[7] Then we define the subpopulation-based improvement confidence as*

$$\gamma^{sub}(a) = \gamma(a, x^{pre}_{G_a}) := P(Y^{post} = 1 | do(a), x^{pre}_{G_a}).$$

The set G_a is chosen for practical reasons. In order to make the estimation more accurate, we would like to condition on as many characteristics as possible. However, without access to the SCM, one can only identify interventional distributions for subgroups of the population by conditioning on their (unobserved) post-intervention characteristics (but not by conditioning on their pre-intervention characteristics) (Pearl 2009; Glymour, Pearl, and Jewell 2016). If we were to select a subgroup from a post-recourse distribution by conditioning on pre-recourse characteristics that are affected by a (e.g. strong pre-recourse symptoms), we yield a group that the individual may not be part of (e.g. people with strong post-recourse symptoms). In contrast, for X_{G_a} pre- and post-intervention values coincide, such that we can estimate γ^{sub}: Assuming causal sufficiency, the standard procedure to sample interventional distributions can be applied, only that additionally $X^{post}_{G_a} := x^{pre}_{G_a}$. Based on the sample γ^{sub} can be estimated (as detailed in B.3).
The estimation of γ^{sub} does not require knowledge of the SCM, but is less accurate than γ^{ind}. In the introductory example, for the action *get vaccinated* the set of subgroup-characteristics G_a is empty. As such, γ^{sub} is concerned with the effect of a vaccination over the whole population. If we were to observe *zip code*, a variable that is not affected by *vaccination*, γ^{sub} would indicate the effect of vaccination for subjects that share the explainee's *zip code*. In contrast, γ^{ind} also takes the explainee's *symptom state* into account.

Optimization problem To generate ICR recommendations, we can optimize Equation 1. We aim to find actions that meet a user-specified improvement target confidence $\overline{\gamma}$ with minimal cost for the recourse seeking individual. The cost function $\text{cost}(a, x^{pre})$ captures the effort the individual requires to perform action a (Karimi et al. 2020b).
As for CE or CR, the optimization problem for ICR is computationally challenging (B.4). It can be seen as a two-level problem, where on the first level the intervention targets I_a, and on the second level the corresponding intervention values θ_a are optimized (Karimi et al. 2020b). Since we target improvement, we can restrict I_a to causes of Y. Following Dandl et al. (2020), we use the genetic algorithm NSGA-II (Deb et al. 2002) for optimization.

$$\text{argmin}_{a=do(X_I=\theta)} \quad \text{cost}(a, x^{pre}) \quad \text{s.t.} \quad \gamma(a) \geq \overline{\gamma}. \quad (1)$$

6 Accurate Post-Recourse Prediction

Recourse recommendations should not only lead to improvement Y but also revert the decision $\hat{Y}$. Whether acceptance guarantees naturally ensue from γ depends on the ability of the predictor to recognize improvements. As follows, we demonstrate how the assumed causal knowledge can be exploited to design accurate post-recourse predictors. We find that an individualized post-recourse predictor is required to translate γ^{ind} into an individualized acceptance guarantee, but curiously that the observational predictor is sufficient in supopulation-based settings.

Individualized post-recourse prediction If we were to use the optimal pre-recourse observational predictor h^* for post-recourse prediction, there would be an imbalance in predictive capability between ML model and individualized ICR: ICR individualizes its predictions using x^{pre} and the SCM. This knowledge is not accessible by the predictor h^*, which only makes use of x^{post}. As such, improvement that was accurately predicted by ICR is not necessarily recognized by h^* and γ^{ind} cannot be directly translated into an acceptance bound. We demonstrate the issue at an Example in E.3.[8]
In order to settle the imbalance between ICR and the predictor, we suggest to leverage the SCM not only when generating individualized ICR recommendations but also when predicting post-recourse, such that the predictor is at least as accurate as γ^{ind}. More formally, we suggest to estimate the post-recourse distribution of Y conditional on x^{pre}, $do(a)$, and the post-recourse observation $x^{post,a}$ (Definition 3). This post-recourse prediction resembles the counterfactual distribution, except that we additionally take the factual post-recourse observation of the covariates into account.

[6] The estimand resembles the conditional treatment effect with G_a being effect modifiers (Hernán MA 2020).

[7] If a cannot affect Y, we can predict $P(Y|x^{pre}, do(a)) = P(Y|x^{pre})$ using the optimal observational predictor h^*.

[8] One may also argue that standard predictive models are not suitable since optimality of the predictor in the pre-recourse distribution does not necessarily imply optimality in interventional environments (as Example 1, E.1 demonstrates). We can refute this criticism using Proposition 3, where we learn that $\hat{h}^*$ is stable with respect to ICR actions.

Definition 3 (Individualized post-recourse predictor)**.** *We define the individualized post-recourse predictor as*

$$h^{*,ind}(x^{post}) = P(Y^{post} = 1|x^{post}, x^{pre}, do(a))$$

For SCMs with invertible equations, $h^{*,ind}$ can be estimated using a closed form solution. Otherwise we can sample from the counterfactual post-recourse distribution $p(y^{post}, x^{post}|x^{pre}, do(a))$ (as we did for the estimation of γ^{ind}), select the samples that conform with x^{post} and compute the proportion of favorable outcomes (details in B.2).
For the individualized post-recourse predictor, improvement probability and prediction are closely linked (Proposition 1). More specifically, the expected post-recourse prediction $h^{*,ind}$ is equal to the individualized improvement probability $\gamma(x^{pre}, a)$. We will exploit Proposition 1 in Section 7, where we derive acceptance guarantees for ICR.

Proposition 1. *The expected individualized post-recourse score is equal to the individualized improvement probability* $\gamma^{ind}(x^{pre}, a) := P(Y^{post} = 1|x^{pre}, do(a))$*, i.e.*

$$E[\hat{h}^{*,ind}(x^{post})|x^{pre}, do(a)] = \gamma^{ind}(a).$$

Subpopulation-based post-recourse prediction Curiously we find that for ICR actions a the optimal observational pre-recourse predictor h^* remains accurate: in the subpopulation of similar individuals the expected post-recourse prediction corresponds to the improvement probability $\gamma^{sub}(a)$ (Proposition 3). This allows us to derive acceptance guarantees for h^* in Section 7.
This result is in contrast to the negative results for CR, where actions may not affect prediction and the underlying target coherently, such that the predictive performance deteriorates (as demonstrated in the introduction, and more formally in E.1). The key difference to CR is that ICR actions exclusively intervene on causes of Y: Interventions on non-causal variables may lead to a shift in the conditional distribution $P(Y|X_S)$ (where $S \subseteq D$ is any set of variables that allows for optimal prediction). In contrast, given causal sufficiency, the conditional $P(Y|X_S)$ is stable to interventions on causes of Y.

Proposition 2. *Given nonzero cost for all interventions, ICR exclusively suggests actions on causes of* Y*. Assuming causal sufficiency, for optimal models the conditional distribution of* Y *given the variables* X_S *that the model uses (i.e.* $P(Y|X_S)$*) is stable w.r.t interventions on causes. Therefore, optimal predictors are intervention stable w.r.t. ICR actions.*

Proposition 3. *Given causal sufficiency and positivity*[9]*, for interventions on causes the expected subgroup-wide optimal score* h^* *is equal to the subgroup-wide improvement probability* $\gamma^{sub}(a) := P(Y^{post} = 1|do(a), x^{pre}_{G_a})$*, i.e.*

$$E[\hat{h}^*(x^{post})|x^{pre}_{G_a}, do(a)] = \gamma^{sub}(a).$$

[9]Positivity ensures that the post-recourse observation lies within the observational support (Neal 2020), where the model was trained (i.e., $p^{pre}(x^{post}) > 0$)).

Link between CR and ICR: Proposition 2 has further interesting consequences. For CR actions a that only intervene on causes of Y and that are guaranteed to yield a predicted score ζ in the subpopulation, we can infer that $\gamma^{sub}(a) \geq \zeta$. For instance, if acceptance with respect to a 0.5 decision threshold can be guaranteed, that implies improvement with at least 50% probability. As such, in subpopulation-based settings (1) improvement guarantees can be made for CR if only interventions on causes are lucrative, and (2) CR can be adapted to also guide towards improvement by a restricting actions to intervene on causes.

7 Acceptance Guarantees

For the presented accurate post-recourse predictors, improvement guarantees translate into acceptance guarantees (Proposition 4). The reason is that the post-recourse prediction is linked to γ (Propositions 1 and 3).

Proposition 4. *Let* g *be a predictor with* $E[g(x^{post})|x^{pre}_S, do(a)] = \gamma(x^{pre}_S, a)$*. Then for a decision threshold* t *the post-recourse acceptance probability* $\eta(t; x^{pre}_S, a) := P(g(x^{post}) > t|x^{pre}_S, do(a))$ *is lower bounded by the respective improvement probability:*

$$\eta(t; x^{pre}_S, a, g) \geq \frac{\gamma(x^{pre}_S, a) - t}{1 - t}.$$

Proof (sketch): We decompose the expected prediction (γ) into true positive rate (TPR), false negative rate (FNR) and acceptance rate. By bounding TPR and FNR we yield the presented acceptance bound. The proof is provided in D.4.

Using Proposition 4, we can tune confidence γ and the model's decision threshold to yield a desired acceptance rate. For instance, we can guarantee acceptance with (subgroup-wide) probability $\eta \geq 0.9$ given $\gamma = 0.95$ and a global decision threshold $t = 0.5$.
Furthermore we can leverage the sampling procedures that we use to compute γ to estimate the individualized or subpopulation-based acceptance rate $\eta(t; x^{pre}_S, a, g)$ (as detailed in B.1 and B.3). To guarantee acceptance with certainty, the decision threshold can be set to $t = 0$.
For the explainee, it is vital that the acceptance guarantee is presented in a human-intelligible fashion. In contrast to previous work in the field, we suggest to communicate the acceptance guarantee in terms of a probability.[10] Furthermore, for subpopulation-based recourse, the set of subgroup characteristics should be transparent. In the hospital admission example, the subpopulation-based acceptance guarantee could be communicated as follows: *Within a group of individuals that share your zip code, a vaccination leads to acceptance with at least probability* η*.*

8 Experiments

In the experiments we evaluate the following questions, assuming correct causal knowledge and accurate models of the conditional distributions in the data:

[10]For CR, the acceptance confidence is encoded in a hyperparameter, as explained in E.2.

Q1: Do CE, CR and ICR lead to improvement?
Q2: Do CE, CR and ICR lead to acceptance (by pre- and post- post-recourse predictor)?
Q3: Do CE, CR and ICR lead to acceptance by other predictors with comparable test error?[11]
Q4: How costly are CE, CR and ICR recommendations?

Setup We evaluate CE, individualized and subpopulation-based CR and ICR with various confidence levels, over multiple runs, and on multiple synthetic and semi-synthetic datasets with known ground-truth (listed below).[12] Random forests were used for prediction, except in the *3var* settings where logistic regression models were used. Following Dandl et al. (2020), we use NSGA-II (Deb et al. 2002) for optimization. For a full specification of the SCMs including the linear cost functions we refer to C.2. Details on the implementation and access to the code are provided in C.1.

3var-causal: A linear gaussian SCM with binary target Y, where all features are causes of Y.
3var-noncausal: The same setup as *3var-causal*, except that one of the features is an effect of Y.
5var-skill: A categorical semi-synthetic SCM where programming skill-level is predicted from causes (e.g. *university degree*) and non-causal indicators extracted from GitHub (e.g. *commit count*).
7var-covid: A semi-synthetic dataset inspired by a real-world covid screening model (Jehi et al. 2020; Wynants et al. 2020).[13] The model includes typical causes like *covid vaccination* or *population density* and symptoms like *fever* and *fatigue*. The variables are mixed categorical and continuous with various noise distributions. Their relationships include nonlinear structural equations.

Results The results are visualized in Figures 3-5 and provided in tabular form in C.3. For each setting CE, CR and ICR explanations were computed over 10 runs on 200 individuals each. For CR and ICR the confidences $0.75, 0.85, 0.9, 0.95$ were targeted (for CR: $\overline{\eta}$, for ICR: $\overline{\gamma}$). For CE no slack is allowed, such that the results correspond to a confidence level of 1.0. Values are plotted on quadratic scales.

Q1 (Figure 3): In scenarios where gaming is possible and lucrative (*3var-noncausal*, *5var-skill* and *7var-covid*) ICR reliably guides towards improvement, but CE and CR game the predictor and yield improvement rates close to zero. For instance, on *5var-skill* CE and CR exclusively suggest to tune the GitHub profile (e.g. by adding more commits). Since the employer offered recourse it should be honored although the applicants remain unqualified. In contrast, ICR

Figure 2: Left: Causal graphs. Right: Legend for color (SCM) and linestyle (recourse type) in Figures 3, 4 and 5.

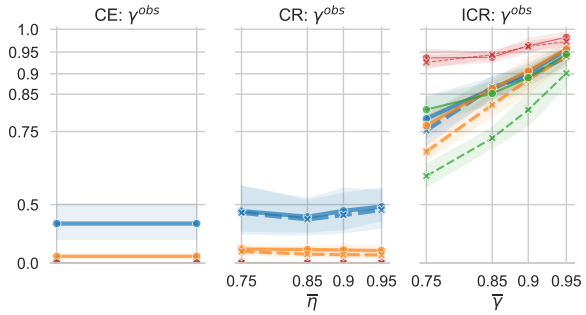


Figure 3: Observed improvement rates γ^{obs} (Q1).

suggests to get a degree or to gain experience, such that recourse implementing individuals are suited for the job.
On *3var-causal*, where gaming is not possible, CR also achieves improvement. However, since acceptance w.r.t to a decision treshold $t = 0.5$ is targeted, only improvement rates close to 50% are achieved (the expected predicted score translates into γ^{sub} (Proposition 3)).
For subp. ICR, γ^{obs} is below $\overline{\gamma}$, because the subpopulation may include individuals that were already accepted pre-recourse, such that γ^{sub} and γ^{obs} may not coincide.

Q2 (Figure 4): All methods yield the desired acceptance rates w.r.t. to the pre-recourse predictor.[14] For CE and CR η^{obs} is higher than for ICR, and for ind. recourse higher than for subp. recourse. Curiously, although no acceptance guarantees could be derived for the pre-recourse predictor and ind. ICR, we find that both pre- and ind. post-recourse predictor reliably lead to acceptance.[15]

Q3 (Figure 5): We observe that CE and CR actions are unlikely to be honored by other model fits with similar performance on the same data. This result is highly relevant to practitioners, since models deployed in real-world scenarios are regularly refitted. As such, individuals that implemented acceptance-focused recourse may not be accepted after all, since the decision model was refitted in the meantime. In contrast, ICR acceptance rates are nearly unaffected by refits. The result confirms our argument that improvement-focused recourse may be more desirable for explainees (Section 4).

[11]The problem that refits on the same data with similar performance have different mechanism is known as the Rashomon problem or model multiplicity (Breiman 2001; Pawelczyk, Broelemann, and Kasneci 2020; Marx, Calmon, and Ustun 2020).

[12]For ground-truth counterfactuals, simulations are necessary (Holland 1986).

[13]The real-world screening model is used to decide whether individuals need a test certificate to enter a hospital. It can be accessed via https://riskcalc.org/COVID19/.

[14]ICR holds the acceptance rates from Proposition 4, as analyzed in more detail in C.3.

[15]Given that the ind. post-recourse predictor is much more difficult to estimate, the pre-recourse predictor in combination with individualized acceptance guarantees (B.1) may cautiously be used as fallback.

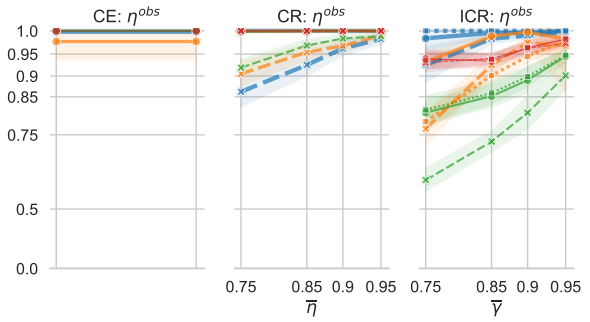


Figure 4: Observed acceptance rates η^{obs} w.r.t. h^*; for ind. ICR additionally w.r.t. $h^{*,ind}$ (Q2).

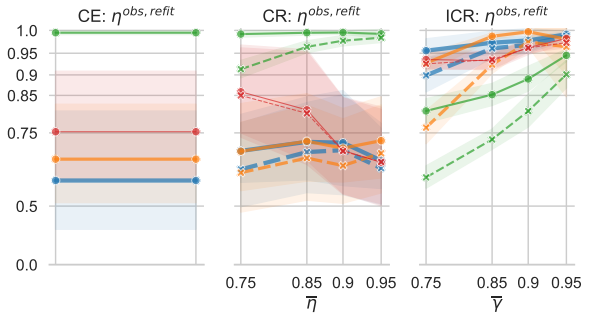


Figure 5: Observed acceptance rates for other fits with comparable test set performance $\eta^{obs,\text{refit}}$ (Q3).

Q4 (Table 1): CR actions are cheaper than ICR actions, since improvement may require more effort than gaming. As such, CR has benefits for the explainee: For instance, on *5var-skill*, CR suggests to tune the GitHub profile (e.g. by adding more commits), which requires less effort than earning a degree or gaining job experience. Detailed results on cost are reported in C.3.

In conclusion, ICR actions require more effort than CR, but lead to improvement and acceptance while being more robust to refits of the model.

9 Limitations and Discussion

Causal knowledge and assumptions Individualized ICR requires a fully specified SCM; Subpopulation-based ICR is less demanding but still requires the causal graph and causal sufficiency. SCMs and causal graphs are rarely readily available in practice (Peters, Janzing, and Schölkopf 2017) and causal sufficiency is difficult to test (Janzing et al. 2012). Research on causal inference gives reason for cautious optimism that the difficulties in constructing SCMs and causal graphs can eventually be overcome (Spirtes and Zhang 2016; Peters, Janzing, and Schölkopf 2017; Heinze-Deml, Maathuis, and Meinshausen 2018; Malinsky and Danks 2018; Glymour, Zhang, and Spirtes 2019).

There are further foundational problems linked to causality that affect our approach: causal cycles, an ontologically vague target Y (e.g. in hiring), disparities in our data, or causal model misspecification (Barocas and Selbst 2016; Barocas, Hardt, and Narayanan 2017; Bongers et al. 2021). All of these factors are considered difficult open problems and may have detrimental impact on our, as well as on any other, recourse framework.

Guiding action without causal knowledge is impossible; when causal knowledge is available, our work provides a normative framework for improvement-focused recourse recommendations. Thus, we join a range of work in explainability (Frye, Rowat, and Feige 2020; Heskes et al. 2020; Wang, Wiens, and Lundberg 2021; Zhao and Hastie 2021) and fairness (Kilbertus et al. 2017; Kusner et al. 2017; Zhang and Bareinboim 2018; Makhlouf, Zhioua, and Palamidessi 2020) that highlights the importance of causal knowledge.

Contestability Improvement-focused recourse guides individuals towards actions that help them to improve, e.g., it recommends a vaccination to lower the risk to get infected with Covid. If, however, a explainee is more interested in contesting the algorithmic decision, (improvement-focused) recourse recommendations are not sufficient. Think of an individual who is denied entrance to an event because of their high Covid risk prediction, which is based on a non-causal, spurious association with their country of origin[16]. In such situations, we suggest to additionally show explainees diverse explanations, which enable to contest the decision. For example, such an explanation could be: if your country of origin would be different, your predicted Covid risk would have been lower.

Table 1: Recourse cost (Q4).

CE	ind. CR	sub. CR	ind. ICR	sub. ICR
1.8 ± 1.1	1.3 ± 1.1	1.7 ± 1.0	4.3 ± 3.3	4.2 ± 3.3

10 Conclusion

In the present paper, we took a causal perspective and investigated the effect of recourse recommendations on the underlying target variable. We demonstrated that acceptance-focused recourse recommendations like counterfactual explanations or causal recourse may not improve the underlying prediction but game the predictor instead. The problem stems from predictive, but non-causal relationships, which are abundant in machine learning applications.[17]

We tackled the problem in the explanation domain and introduced Improvement-Focused Causal Recourse (ICR), an explanation technique that guides towards improvement of the prediction target and demonstrated how to design post-recourse predictors such that improvement leads to acceptance. We confirm the theoretical results in experiments.

With ICR we hope to inspire a shift from acceptance- to improvement-focused recourse.

[16] E.g., due to a spurious association with the causal variable *type of vaccine*.

[17] For instance, in hiring, certain keywords in the CV may be associated with qualification, but adding them to the CV does not improve aptitude (Strong 2022).

Acknowledgements

This project is supported by the German Federal Ministry of Education and Research (BMBF), the Carl Zeiss Foundation (project on "Certification and Foundations of Safe Machine Learning Systems in Healthcare"), and by the Graduate School of Systemic Neurosciences (GSN) Munich. The authors of this work take full responsibility for its content. We thank the anonymous reviewers for their feedback, which guided us towards improvement (and acceptance).

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2017. Fairness in machine learning. *Nips tutorial*, 1: 2.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California law review*, 671–732.

Barocas, S.; Selbst, A. D.; and Raghavan, M. 2020. The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 8089. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Bechavod, Y.; Ligett, K.; Wu, Z. S.; and Ziani, J. 2020. Causal feature discovery through strategic modification. *arXiv preprint arXiv:2002.07024*.

Bongers, S.; Forré, P.; Peters, J.; and Mooij, J. M. 2021. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5): 2885–2915.

Breiman, L. 2001. Statistical Modeling: The Two Cultures. *Statistical Science*, 16(3): 199–231.

Chen, Y.; Wang, J.; and Liu, Y. 2020. Linear Classifiers that Encourage Constructive Adaptation. *arXiv preprint arXiv:2011.00355*.

Dandl, S.; Molnar, C.; Binder, M.; and Bischl, B. 2020. Multi-Objective Counterfactual Explanations. In Bäck, T.; Preuss, M.; Deutz, A.; Wang, H.; Doerr, C.; Emmerich, M.; and Trautmann, H., eds., *Parallel Problem Solving from Nature – PPSN XVI*, 448–469. Cham: Springer International Publishing. ISBN 978-3-030-58112-1.

Deb, K.; Pratap, A.; Agarwal, S.; and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2): 182–197.

Freiesleben, T. 2021. The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds and Machines*.

Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33: 1229–1239.

Glymour, C.; Zhang, K.; and Spirtes, P. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10: 524.

Glymour, M.; Pearl, J.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Hardt, M.; Megiddo, N.; Papadimitriou, C.; and Wootters, M. 2016. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Heinze-Deml, C.; Maathuis, M. H.; and Meinshausen, N. 2018. Causal structure learning. *Annual Review of Statistics and Its Application*, 5: 371–391.

Hernán MA, R. J. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Heskes, T.; Sijben, E.; Bucur, I. G.; and Claassen, T. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33: 4778–4789.

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396): 945–960.

Janzing, D.; Sgouritsa, E.; Stegle, O.; Peters, J.; and Schölkopf, B. 2012. Detecting low-complexity unobserved causes. *CoRR*, abs/1202.3737.

Jehi, L.; Ji, X.; Milinovich, A.; Erzurum, S.; Rubin, B. P.; Gordon, S.; Young, J. B.; and Kattan, M. W. 2020. Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11,672 patients. *Chest*, 158(4): 1364–1375.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2020a. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 353362. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2020b. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 265–277. virtual: Curran Associates, Inc.

Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2022. Towards Causal Algorithmic Recourse. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, 139–166. Springer.

Kilbertus, N.; Rojas Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Makhlouf, K.; Zhioua, S.; and Palamidessi, C. 2020. Survey on causal-based machine learning fairness notions. *arXiv preprint arXiv:2010.09553*.

Malinsky, D.; and Danks, D. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1): e12470.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

Miller, J.; Milli, S.; and Hardt, M. 2020. Strategic Classification is Causal Modeling in Disguise. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 6917–6926. Online: PMLR.

Neal, B. 2020. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*.

Obermeyer, Z.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm that guides health decisions for 70 million people. In *Proceedings of the conference on fairness, accountability, and transparency*, 89–89.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. On Counterfactual Explanations under Predictive Multiplicity. In Peters, J.; and Sontag, D., eds., *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, 809–818. Online: PMLR.

Pearl, J. 2009. *Causality*. Cambridge, UK: Cambridge University Press, 2 edition. ISBN 978-0-521-89560-6.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 469481. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Shavit, Y.; Edelman, B.; and Axelrod, B. 2020. Causal Strategic Linear Regression. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8676–8686. virtual: PMLR.

Spirtes, P.; and Zhang, K. 2016. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, 1–28. SpringerOpen.

Stepin, I.; Alonso, J. M.; Catala, A.; and Pereira-Fariña, M. 2021. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9: 11974–12001.

Strong, J. 2022. MIT Technology Review: Beating the AI hiring machines. https://www.technologyreview.com/2021/08/04/1030513/podcast-beating-the-ai-hiring-machines/. Accessed 2022-07-15.

Tsirtsis, S.; and Gomez Rodriguez, M. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems*, 33: 16749–16760.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 1019. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Venkatasubramanian, S.; and Alfano, M. 2020. The Philosophical Basis of Algorithmic Recourse. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 284293. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wang, J.; Wiens, J.; and Lundberg, S. 2021. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, 721–729. PMLR.

Wynants, L.; Van Calster, B.; Collins, G. S.; Riley, R. D.; Heinze, G.; Schuit, E.; Bonten, M. M.; Dahly, D. L.; Damen, J. A.; Debray, T. P.; et al. 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj*, 369.

Zeng, J.; Ustun, B.; and Rudin, C. 2017. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3): 689–722.

Zhang, J.; and Bareinboim, E. 2018. Fairness in decision-makingthe causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. Issue: 1.

Zhao, Q.; and Hastie, T. 2021. Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1): 272–281.

A Extended Background

As follows, we recapitulate well-known definitions in our notation, provide more detailed background on related work and recapitulate results that we use in the proofs. Readers who are already familiar with recourse terminology and d-separation (A.1 and A.2), and who are not interested in more detailed introductions of intervention stability (A.3, only required for the proof of Proposition 2) or causal recourse (A.4), may skip this section.

A.1 Overview of important terms

An overview of important terms is provided in Table 2.

A.2 d-separation

Two variable sets X, Y are called d-separated (Geiger, Verma, and Pearl 1990; Spirtes et al. 2000) by the variable set Z in a graph $\mathcal{G}$ (denoted as $X \perp_{\mathcal{G}} Y|Z$), if, and only if, for every path p it either holds that (i) p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ where $m \in Z$ or (ii) p contains a collider $i \rightarrow m \leftarrow j$ such that m and for all of its descendants n it holds that $m, n \notin Z$. Given the causal Markov property, d-separation in a causal graph implies (conditional) independence in the data (Peters, Janzing, and Schölkopf 2017).

A.3 Generalizability and intervention stability

For Proposition 2, we leverage necessary conditions for invariant conditional distributions as derived in (Pfister et al. 2021). The authors introduce a d-separation based intervention stability criterion that is applied to a modified version of $\mathcal{G}$. For every intervened upon variable X_l an auxiliary intervention variable, denoted as I_l, is added as direct cause of X_l, yielding $\mathcal{G}^*$. The intervention variable can be seen as a switch between different mechanisms. A set $S \subseteq \{1, \dots, d\}$ is called *intervention stable* regarding a set of actions if for all intervened upon variables X_l (where $l \in I^{\text{total}}$) the d-separation $I^l \perp_{\mathcal{G}^*} Y|X_S$ holds in $\mathcal{G}^*$. The authors show that intervention stability implies an invariant conditional distribution, i.e., for all actions $a, b \in \mathbb{A}$ with $I^a, I^b \subseteq I^{\text{total}}$ it holds that $p(y^a|x_S) = p(y^b|x_S)$ (Pfister et al. (2021), Appendix A).

A.4 Causal recourse

ICR is closely related to the CR framework (Karimi et al. 2020b; Karimi, Schölkopf, and Valera 2021), but differs substantially in its motivation and target. In order to allow for a direct comparison we briefly sketch the main ideas and the central CR definitions in our notation. Like ICR, CR aims to guide individuals to revert unfavorable algorithmic decisions (recourse). Therefore, they suggest to search for cost-efficient actions that lead to acceptance by the prediction model. Actions are modeled as structural interventions $a : \Pi \rightarrow \Pi$, which can be constructed as $a = do(\{X_i := \theta_i\}_{i \in I})$, where I is the index set of features to be intervened upon (Karimi, Schölkopf, and Valera 2021). The conservativeness of the suggested actions can be adjusted using the hyperparameter γ_{LCB}, that determines the adaptive threshold $\texttt{thresh}(a)$ and thereby how many standard deviations the expected prediction shall be away from the model's decision threshold t. In order to accommodate different levels of causal knowledge, two probabilistic versions of CR were introduced (Karimi et al. 2020b): While individualized recourse assumes knowledge of the SCM, subpopulation-based CR only assumes knowledge of the causal graph.

Individualized recourse Individualized recourse predicts the effect of actions using structural counterfactuals (Karimi, Schölkopf, and Valera 2021), which require a full specification of the SCM.

Given a function that evaluates the cost of actions ($\text{cost}(a, x^{pre})$), the optimization goal for individualized causal recourse is given below. The adaptive threshold $\texttt{thresh}$ bounds the prediction away from the decision threshold.[18]

$$
\begin{aligned}
a^* \in \underset{a \in \mathbb{A}}{\text{argmin}} \quad & \text{cost}(a, x^{pre}) \\
\text{s.t. } & \mathbb{E}[\hat{h}(x^{post})|do(a), x^{pre}] \geq \texttt{thresh}(a) \\
\text{with } & \texttt{thresh}(a) := 0.5 + \gamma_{LCB} \sqrt{\text{Var}[\hat{h}(x^{post,a})]}
\end{aligned}
$$

Subpopulation-based recourse: If no knowledge of the SCM is given, counterfactual distributions cannot be estimated and consequently individualized recourse recommendations cannot be computed. Subpopulation-based CR is based on the average treatment effect within a subgroup of similar individuals (Karimi et al. 2020b). More specifically individuals belong to the same group if the non-descendants $nd(I)$ of intervention variables (which ceteris paribus remain constant despite the intervention) take the same value. The subpopulation-based objective is given below.

$$
\begin{aligned}
a^* \in \underset{a \in \mathbb{A}}{\text{argmin}} \; & \text{cost}(a, x^{pre}) \text{ s.t.} \\
& \mathbb{E}_{X_{d(I)}|do(X_I=\theta), x^{pre}_{nd(I)}}[\hat{h}(x^{pre}_{nd(I)}, \theta, X_{d(I)})] \\
& \qquad \geq \texttt{thresh}(a).
\end{aligned}
$$

A.5 Robust algorithmic recourse

The robustness of CEs and CR has been investigated before (Rawal, Kamar, and Lakkaraju 2021; Pawelczyk, Broelemann, and Kasneci 2020; Upadhyay, Joshi, and Lakkaraju 2021; Dominguez-Olmedo, Karimi, and Schölkopf 2021; Pawelczyk et al. 2022), yet only with respect to generic shifts of model and data. Only (Pawelczyk, Broelemann, and Kasneci 2020) investigate the robustness regarding refits on the same data. They find that on-the-manifold CEs are more robust than standard CEs. In contrast, we empirically compare the robustness of CE, CR and ICR with respect to refits on the same data.

[18] Further constraints have been suggested, e.g., $x^{post,a} \in \mathcal{P}$lausible or $a \in \mathcal{F}$easible (Laugel et al. 2019; Ustun, Spangher, and Liu 2019; Mahajan, Tan, and Sharma 2020; Dandl et al. 2020; Karimi, Schölkopf, and Valera 2021).

Table 2: Overview of important terms and their meanings.

term	meaning
explainee	individual for whom the explanation is generated, e.g. loan applicant
model authority	decision-making entity, e.g. credit institute
recourse	action of the explainee that reverts unfavorable decision
acceptance	desirable model prediction ($\hat{Y} = 1$)
improvement	(yield) desirable state of the underlying target ($Y = 1$)
gaming	yield acceptance without improvement, e.g. treating the symptoms
pre-/post-recourse	before/after implementing recourse recommendation
contestability	the explainee's ability to contest an algorithmic decision
robustness of recourse	probability that recourse is accepted despite model/data shifts

B Estimation and Optimization

As follows we provide detailed explanations of the proposed estimation procedures. First, we explain how to sample from the individualized post-recourse distribution, which allows us to estimate the individualized improvement and acceptance rates (γ^{ind} and η^{ind}, B.1). Based on the same sampling mechanism we can also estimate the individualized post-recourse prediction $h^{*,ind}$ (B.2). Then we explain how to sample from the subpopulation-based post-recourse distribution, which allows us to estimate the subpopulation-based improvement and acceptance rates (γ^{sub} and η^{sub}, B.3). Furthermore, we provide details on optimization (B.4) and demonstrate that the optimal observational predictor h^* can also be estimated using the SCM (B.5).

B.1 Estimation of the individualized improvement confidence γ^{ind} and individualized acceptance rate η^{ind}

We recall that γ^{ind} is the counterfactual probability of the underlying target Y taking the favorable outcome, and η^{ind} the counterfactual probability of the prediction $\hat{Y}$ taking the favorable outcome. In order to estimate γ^{ind} and η^{ind} we first sample covariates and target from the counterfactual post-recourse distribution and then compute the proportion of favorable outcomes for Y and $\hat{Y}$ in the sample.

In general, sampling from counterfactual distributions based on a SCM is performed in three steps (Section 3, (Pearl 2009)).

1. *Abduction*: The exogenous noise variables are reconstructed from the observations, i.e., $p(u_{Y,D}|x^{pre})$ is estimated.
2. *Intervention*: The intervention $do(a)$ on the SCM $\mathcal{M}$ is performed by replacing the respective structural equations $f_{I_a} := \theta_{I_a}$, yielding $\mathcal{M}_{do(a)}$.
3. *Prediction*: The abducted noise variables are sampled from $p(u_{Y,D}|x^{pre})$ and passed through the model $\mathcal{M}_{do(a)}$ to sample from the counterfactual distribution $P(Y^{post}, X^{post}|x^{pre}, do(a))$.

Given knowledge of the SCM, the challenge is to sample the exogeneous variables from $p(u_{Y,D}|x^{pre})$ (abduction). As follows we explain the abduction in two steps. First, we explain how we can abduct u_j for variables for which both the node x_j and all parents $(x, y)_{pa(j)}$ are observed, which we refer to as the standard abduction case. Then we factorize the abduction of the joint $p(u_{Y,D}|x^{pre})$ into several components which can be reduced to said standard abduction case. The sampling procedure is summarized in Algorithm 1.

Recap: Standard abduction If for a node u_j both the node $(x, y)_j$ and the parents $(x, y)_p a(j)$ are observed, we can apply standard abduction. The standard abduction procedure depends on the type of structural equation and exogenous noise distribution.

Given invertible structural equations, observation of $x_j, x_{pa(j)}$ determines u_j. More specifically, u_j can be reconstructed using

$$u_j = f^{-1}(x_j; x_{pa(j)}).$$

For instance, for additive structural equations $f_j(u_j; x_{pa(j)}) = g(x_{pa(j)}) + u_j$, the inversion is given by $f_j^{-1}(x_j; x_{pa(j)}) = x_j - g(x_{pa(j)})$.

In our experiments we also included binomial variables with a sigmoidal (non-invertible) structural equation. More specifically, the structural equations are defined as $x_j = [\sigma(l(x_{pa(j)})) \leq u_j]$ with $U_j \sim Unif(0, 1)$. Here σ refers to the sigmoid function and l to some linear combination. $[cond]$ evaluates to 1 when the condition is true and otherwise to 0. Intuitively, $\sigma(l(x_{pa(j)}))$ can be seen as a nonlinear activation function which determines the probability of the node being activated ($x_j = 1$). u_j acts as a dice, where values $\leq \sigma(l(x_{pa(j)}))$ imply $x_j = 1$ and vice versa.

For those variables, if $x_j = 1$, we know that $u_j \leq \sigma(l(x_{pa(j)}))$ and vice versa, such that we can abduct U_j as follows (and can therefore sample u_j):

$$P(U_j|x_j; x_{pa(j)}) = \begin{cases} Unif(0, \sigma(l(x_{pa(j)}))), & \text{for } x_j = 1 \\ Unif(\sigma(l(x_{pa(j)})), 1), & \text{for } x_j = 0 \end{cases}$$

As we will see in the next section, our estimation procedure can be flexibly extended to SCMs with different types of structural equations, as long as a procedure to sample from the abducted exogneous noise variable for the standard case (where parents and the node itself are observed) is available.

Factorization of $p(u|x)$ We have demonstrated how to abduct individual nodes in the standard setting where the

Algorithm 1: Sampling from the individualized post-recourse distribution

Data: pre-recourse observation x^{pre}, action a (where $do(a) := do(X_{I_a} := \theta)$), sample size M, structural causal model $\mathcal{M}$ with structural equations f_j, observational predictor h
Result: sample from $p(y^{post}, x^{post}|x^{pre}, do(a))$
get $\mathcal{M}_{do(a)}$ by updating $f_i(x_{pa(i)}; u_i) := \theta_i$ for $i \in I_a$;
for m ***in*** $(0, ..., M-1)$ **do**
  sample y' from $Binomial(h(x^{pre}))$;
  for j ***in*** D **do**
    sample $u_j^{(m)}$ from $p(u_j|(x, y')_j, (x, y')_{pa(j)})$;
    ▷ comment: leveraging standard abduction;
  end
  sample $u_Y^{(m)}$ from $p(u_Y|y', x_{pa(Y)})$;
  compute $(x^{post}, y^{post})^{(m)} = f_{\mathcal{M}_{do(a)}}(u^{(m)})$;
end

corresponding endogenous variable and its parents are observed.

As follows we demonstrate how to sample from the joint distribution of the exogenous variables given an observation of X (and without observing Y). Therefore, we show that $p(u|x)$ can be seen as a mixture of two distributions, one for each possible state y' of Y. In order to sample from it, we (1) need to sample y' from the mixing distribution $p(y|x)$ and (2) given y', sample from the respective abducted noise variable $p(u|y', x)$.

$$p(u|x) \tag{2}$$

$$\overset{\text{law tot. prob.}}{=} \sum_{y' \in \{0,1\}} p(u, y'|x) \tag{3}$$

$$\overset{\text{cond. prob.}}{=} \sum_{y' \in \{0,1\}} p(u|y', x)p(y'|x) \tag{4}$$

The binomial mixing distribution $p(y|x)$ can be obtained and sampled from by leveraging the cross-entropy optimal predictor h^* (which can for instance be derived from the SCM, see B.5). In order to sample from $p(u|y', x)$ we leverage the Markov factorization, which allows us to sample each component independently using the standard abduction procedure described above.

$$\begin{aligned} p(u|x, y') \overset{\text{d-sep.}}{=} & P(u_Y|x_{pa(Y)}, y') \\ & \prod_{k \in ch(Y)} P(u_k|x_k, x_{pa(k)}, y') \\ & \prod_{k \notin ch(Y)} P(u_k|x_k, x_{pa(k)}). \end{aligned} \tag{5}$$

The overall procedure is summarized in Algorithm 1.

Estimation of γ^{ind} and η^{ind} Given the procedure to sample from the individualized post-recourse distribution we can estimate γ^{ind} by taking the mean over the samples taken for Y^{post}. Similarly, for each sample for X^{post} we can compute the prediction $\hat{y}^{post}$ using either $h \geq t$ or $h^{ind} \geq t$. By taking the mean over all sampled predictions $\hat{y}^{post}$ we can estimate the respective acceptance probability $\eta(t; x^{pre}, a, h)$ or $\eta(t; x^{pre}, a, h^{ind})$.

Algorithm 2: Estimating $h^{*,ind}$

Data: pre-recourse observation x^{pre}, action a, sample size M, structural causal model $\mathcal{M}$, observational predictor h, $m = 0$
Result: $\hat{h}^{ind}(x^{post}; x^{pre}, do(a))$
while $m < M$ **do**
  sample (x', y') using Alg. 1 and $x^{pre}, a, \mathcal{M}, h$;
  if $x' = x^{post}$ **then**
    $m = m + 1$; store y' as $y'^{(m)}$;
  end
end
$\hat{h}^{ind}(x^{post}) = \frac{1}{M} \sum_{m=1}^{M} y'^{(m)}$

B.2 Estimation of the individualized post-recourse prediction

We continue to show how the individualized post-recourse prediction can be estimated. We recall that $h^{*,ind}$ is

$$h^{*,ind}(x^{post}; x^{pre}, a) = P(Y^{post} = 1|x^{post}, x^{pre}, do(a)).$$

We can estimate $h^{*,ind}$ by leveraging the procedure to sample from the post-recourse covariate distribution (Algorithm 1). More specifically, we draw samples (y', x') from $P(Y^{post}, X^{post}|do(a), x^{pre})$ and keep those that conform with x^{post} (i.e., $x' = x^{post}$). Within the subsample, we compute the proportion of samples for which $y' = 1$ to estimate $p(y^{post}|x^{pre}, x^{post}, do(a))$. In more formal terms, we approximate Eq. 6 using rejection sampling and Monte Carlo integration (Koller and Friedman 2009).

If the structural equations are invertible[19] or the nodes are categorical the procedure is tractable, since many or all samples conform with x^{post}. Otherwise the estimation may become intractable. We see the application of likelihood weighting or MCMC as promising directions and refer interested readers to Koller and Friedman (2009).

In addition to the sampling-based procedure we also derive a closed-form solution for settings with invertible structural equations, which is provided in Proposition 5, Eq. 7.

Proposition 5. *In general, the individualized post-recourse predictor can be estimated as*

$$\begin{aligned} & p(y^{post}|x^{pre}, x^{post}, do(a)) \\ & = \frac{\int_{\mathcal{U}} p(y^{post}, x^{post}|u, do(a))p(u|x^{pre})du}{\sum_{y' \in \{0,1\}} \left(\int_{\mathcal{U}} p(y', x^{post}|u, do(a))p(u|x^{pre})du\right)} \end{aligned} \tag{6}$$

[19] Meaning that the abducted joint distribution has point mass probability for two configurations, one for each possible state of Y.

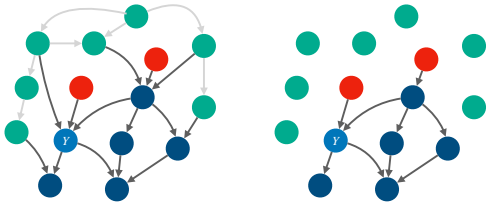


Figure 6: Left: Causal graph $\mathcal{G}_{\overline{I_a}}$ visualizing the subpopulation-based post-recourse setting, including the prediction target Y (light blue), intervened-upon variables I_a (red), the subgroup characteristics G_a (cyan) and the descendants Γ that shall be resampled (dark blue). $\overline{I_a}$ indicates that incoming edges to I_a were removed. Right: Causal graph $\mathcal{G}_{\overline{I_a}\underline{G_a}}$ where incoming edges to I_a and outgoing edges from G_a were removed. We observe that in this manipulated graph G_a is d-separated from Γ. Thus, according to the second rule of do-calculus, for G_a intervention and conditioning coincide.

Given invertible structural equations, the individualized post-recourse prediction function reduces to

$$p(y^{post}|x^{post}, x^{pre}, do(a)) = \frac{p(U_{-I} = f^{-1}_{do(a)}(y^{post}, x^{post})|x^{pre}, do(a))}{\sum_{y' \in \{0,1\}} p(U_{-I} = f^{-1}_{do(a)}(y', x^{post})|x^{pre}, do(a))}. \tag{7}$$

B.3 Estimation of the subpopulation-based improvement confidence γ^{sub} and the subpopulation-based acceptance rate η^{sub}

As follows we detail how to estimate γ^{sub} and η^{sub}. We focus on actions a that potentially affect Y, meaning that they intervene on causes of Y.[20]

In order to estimate γ^{sub} and η^{sub} we sample (x', y') from the subpopulation-based post-recourse distribution. Given a sample from the subpopulation-based post-recourse distribution we can estimate γ^{sub} and η^{sub} by taking the respective sample means.

We explain the sampling procedure in two steps: We first recall how causal graphs can be leveraged to sample interventional distributions, and then explain why we can apply the procedure to sample from the subpopulation-based post-recourse distribution.

Recap: Sampling interventional distributions leveraging a causally sufficient causal graph $\mathcal{G}$ Given a causal graph $\mathcal{G}$ (that fulfills the global Markov property), the joint distribution $P(X, Y)$ can be reformulated using the Markov factorization, which makes use of the d-separations in the graph.

$$p(x, y) = p(y|x_{pa(y)}) \prod_{j \in D} p(x_j|(x, y)_{pa(j)})$$

As a consequence, we can sample from the joint distribution by sampling each component given its respective parents. In order to ensure that the parents for each node have been sampled already, the graph is traversed in topological order, starting with the root node and ending with the sink nodes (Koller and Friedman 2009).

Given that causal sufficiency (no unobserved confounders) and the principle of independent mechanisms hold, the same procedure can also be applied when sampling from interventional distributions of the form $p(x, y|do(a))$ by leveraging the so-called truncated factorization. The intervened upon nodes are not sampled from their parents, but fixed to the values θ_a. The remaining nodes Γ are sampled as before:

$$p((x, y)_\Gamma|do(a)) = \prod_{j \in \Gamma} p((x, y)_j|(x, y)_{pa(j) \cap \Gamma}, \theta_{pa(j) \cap I_a})$$

with $\Gamma := D \backslash I_a$

Sampling from the subpopulation-based post-recourse distribution using $\mathcal{G}$ We recall that for actions a that potentially affect Y the subpopulation-based post-recourse distribution is defined as

$$P(Y^{post}, X^{post}|do(a), X^{post}_{G_a} = x^{pre}_{G_a}). \tag{8}$$

As we will see, the previously described sampling procedure can be applied. Therefore we apply the second rule of do-calculus to show that in Equation 8 conditioning on x_{G_a} is equal to intervening $do(X_{G_a} = x_{G_a})$. More specifically, if we remove all outgoing edges from X_{G_a} and all incoming edges to I_a, then X_{G_a} and X_Γ with $\Gamma := D \backslash I_a \cap G_a = d(I_a)$ are d-separated, meaning that conditioning and intervention

Algorithm 3: Sampling from the subpopulation-based post-recourse distribution

Data: pre-recourse observation x^{pre}, action a with $I_a \cap asc(Y) \neq \emptyset$ ($do(a) := do(X_{I_a} := \theta)$), sample size M, causal graph $\mathcal{G}$, conditional distributions $P(X_j|X_{pa(j)})$ for $j \in \Gamma$ with $\Gamma := \{r : r \in asc(Y) \wedge r \in d(I)\}$
Result: sample from $p(y, x_\Gamma|do(a), x_{G_a})$
for $m \leftarrow 0$ **to** M **do**
  $\Gamma^{sorted} \leftarrow$ topologicalsort($\Gamma; \mathcal{G}_{do(a)}$) ▷ sort such that causes precede effects ;
  for j ***in*** Γ^{sorted} **do**
    sample $(x, y)^{post,(m)}_j$ $\sim P((X, Y)_j|(X, Y)_{pa(j)} = (x, y)^{post}_{pa(j)})$;
  end
end

[20] Actions that do not affect Y trivially do not lead to improvement. The respective probability of $Y = 1$ can be estimated using the optimal observational predictor.

are equivalent (Figure 6).

$$P((Y,X)_{\Gamma}^{post}|do(a), X_{G_a}^{post} = x_{G_a}^{pre}) = P((Y,X)_{\Gamma}^{post}|do(a), do(X_{G_a}^{post} = x_{G_a}^{pre}))$$

As follows we can leverage the procedure to sample interventional distributions to sample from the subpopulation-based post-recourse distribution. The procedure is illustrated in Algorithm 3.

Learning the conditional distributions $P(X_j|x_{pa(j)})$ In this work we assume that we have prior knowledge that allows us to sample from the components of the factorization ($P(X_j|x_{pa(j)})$, e.g. available if we know the SCM).
If the conditional distributions are not known, they can be learned from observational data; depending on which assumptions about distribution and functional can be made, different techniques may be employed. For categorical variables the problem reduces to standard supervised learning with cross-entropy loss. For linear Gaussian data, the conditional distribution can be estimated analytically from the covariance matrix (Page Jr 1984). A variety of estimation techniques exist for continuous settings with nonlinearities (Bishop 1994; Bashtannyk and Hyndman 2001; Sohn, Lee, and Yan 2015; Trippe and Turner 2018; Winkler et al. 2019; Hothorn and Zeileis 2021).

B.4 Optimization

Like the optimization problems for CE (Wachter, Mittelstadt, and Russell 2017; Tsirtsis and Gomez Rodriguez 2020) or CR (Karimi et al. 2020b), the optimization problem for ICR is computationally challenging. It can be seen as a two-stage problem, where in the first stage the intervention targets I_a, and in the second stage the corresponding intervention values θ_a are optimized (Karimi et al. 2020b). For the selection of intervention targets I_a alone $2^{d'}$ combinations exist, with $d' \leq d$ being the number of causes of Y. We jointly optimize the intervention targets and the intervention values using a genetic algorithm called NSGA-II (Deb et al. 2002). For mixed categorical and continuous data, previous work in the field (Dandl et al. 2020) suggests to use NSGA-II in combination with *mixed integer evaluation strategies* (Li et al. 2013). The exact hyperparameter configurations are reported in C.3.

B.5 Estimation of the optimal observational predictor h^* using the SCM

Instead of leveraging supervised learning with cross-entropy loss, we can factorize the optimal observational predictor as shown in Proposition 6 and then leverage the SCM for the estimation.

Proposition 6. *The optimal observational predictor can be factorized into conditional distributions of nodes given their parents (using the Markov factorization). More specifically, we yield*

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\sum_{y' \in \{0,1\}} p(x,y)} \tag{9}$$

$$\overset{\text{M.f.}}{=} \frac{p(y|x_{pa(j)}) \prod_{j \in D} p(x_j|(x,y)_{pa(j)})}{\sum_{y' \in \{0,1\}} p(y'|x_{pa(j)}) \prod_{j \in D} p(x_j|(x,y')_{pa(j)})} \tag{10}$$

$$= \frac{p(y|x_{pa(j)}) \prod_{j \in ch(y)} p(x_j|x_{pa(j)}, y)}{\sum_{y' \in \{0,1\}} p(y'|x_{pa(j)}) \prod_{j \in ch(y)} p(x_j|x_{pa(j)}, y')}. \tag{11}$$

It remains to show how the conditional distribution $p(x_j|x_{pa(j)})$ of a node given its parents can be estimated. Generally it holds that

$$p(x_j|x_{pa(j)}) \tag{12}$$

$$\overset{\text{law tot. prob.}}{=} \int_{\mathcal{U}_j} p(x_j|x_{pa(j)}, u_j) p(u_j|x_{pa(j)}) du \tag{13}$$

$$\overset{\text{SCM, } u_j \perp x_{pa(j)}}{=} \int_{\mathcal{U}_j} [f(x_{pa(j)}, u_j) = x_j] p(u_j) du. \tag{14}$$

The integral can be approximated using Monte Carlo integration: we can sample from $p(u_j)$, compute the respective $\tilde{x}_j = f_j(x_{pa(j)}, \tilde{u}_j)$ and compute the proportion of cases where $x_j = \tilde{x}_j$. If X_j and U_j are continuous, this may require huge sample sizes to converge.
Furthermore, we may be able to leverage assumptions about f_j to derive a closed form solution. If f_j is invertible, the integral reduces to $p(x_j|x_{pa(j)}) = p(U_j = f_j^{-1}(x_j, x_{pa(j)}))$. For binary nodes with $x_j := [\sigma(l(x_{pa(j)})) \leq u_j]$ and $U_j \sim Unif(0,1)$, we directly see that $p(x_j|x_{pa(j)}) = \sigma(l(x_{pa(j)}))$.

C Details on Experiments

In this section we provide additional details on the experiments. More specifically, we explain which open-source libraries we use, how to access our code and how to reproduce the results in C.1. We formally introduce the synthetic and semi-synthetic datasets that we used in our experiments in C.2 and the corresponding figures. Details on hyperparameters, models as well as detailed results are reported in C.3 and the corresponding tables.

C.1 Implementation

The code relies of efficient tensor calculations with `numpy` (Harris et al. 2020), `pytorch` (Paszke et al. 2019) and `jax` (Bradbury et al. 2018). For named dataframes we use `pandas` (pandas development team 2020). For plotting we rely on `matplotlib` (Hunter 2007) and `seaborn` (Waskom 2021). We use the evolutionary optimization library deap (Fortin et al. 2012) and NSGA-II (Deb et al. 2002) to solve the combinatorial optimization problem.[21] In order to speed up the computation, we cache

[21] We also implemented abduction based on probabilistic inference. Thereby we rely on on `pyro` (Bingham et al. 2018) for

queries and results for the improvement confidence using `functools.cache`. For continuous variables the intervention can be rounded to a specified number of digits to increase the probability of reusing a cached result (with neglectable loss of precision).[22]

All code is publicly available via https://anonymous.4open.science/r/icr-aaai/README.md. The repository contains the user-friendly python package `icr`, which we use in our experiments to generate and evaluate recourse. Furthermore, the scripts for the experiments, the scripts for the visualization of the results as well as a `README.md` with instructions for the installation of all dependencies are contained in the repository, such that the experiments are reproducible.

C.2 Synthetic and Semi-Synthetic Datasets

3var-causal and *3var-noncausal* are abstract, synthetic settings. *5var-skill* is inspired by Montandon, Valente, and Silva (2021), who use GitHub profiles to detect the role of a developer. In our SCM we model *senior-level skill* as a binary variable which is caused by *programming experience* and the education *degree*. The skill is causal for GitHub metrics such as the number of *commits*, the number of programming *languages* and the number of *stars*. The *7var-covid* dataset is inspired by Jehi et al. (2020). The following variables are introduced: population density D, flu vaccination V_I, number of covid vaccination shots V_C, deviation from average BMI B, whether someone is free of covid disease C, whether the individual has influence I, appetite loss S_A, fever S_{Fe} and fatigue S_{Fa}. The corresponding structural equations, noise distributions and causal graphs are provided in Figure 7 (*3var-causal*), 8 (*3var-noncausal*), 9 (*5var-skill*) and 10 (*7var-covid*). A pairplot for each dataset is presented in Figure 11. In our notation σ is the sigmoid function, N the Gaussian distribution, Cat a categorical distribution, $Unif$ the uniform distribution, $Bern$ a Bernoulli distribution and GaP a Gamma-Poisson mixture. $[cond]$ is 1 when the condition is met and 0 if not. As a consequence variables with $[Z \leq U]$ and $U \sim Unif(0,1)$ are bernoulli distributed with $Bern(Z)$.

C.3 Detailed Results

In this section we report all experimental results in tabular form. More specifically, the results for *3var-causal* are reported in Table 3, for *3var-noncausal* in Table 4, for *5var-skill* in Table 5 and for *7var-covid* in Table 6. For each experiment we report the specified confidence γ (or η for CR), as well as the observed improvement rate γ_{obs}, the observed acceptance rate η_{obs}, the observed acceptance rate by the individualized post-recourse predictor $\eta_{obs}^{\text{indiv.}}$, the observed acceptance rate on refits $\eta_{obs}^{\text{refit}}$ and the average recourse cost for individuals who were rejected and whom were provided with a recourse recommendation. A visual summary of the results is provided in Section 8.

In order to enable a more direct comparison of the CR and ICR targets, we equalize the optimization thresholds for ICR and CR. More specifically, for CR we require the (individualized or subpopulation-based) acceptance probability to be $\geq \eta$, and for ICR we require the (individualized or subpopulation-based) improvement probability to be $\geq \overline{\gamma}$, where $\overline{\gamma} = \overline{\eta}$.[23] Furthermore, in order to be able to estimate the effects of recourse actions, CR assumes causal sufficiency, meaning that there are no two endogeneous variables that share an unobserved cause. If the target variable Y is exogeneous then any causal model with more than one endogeneous direct effect of Y violates the assumptions. In order to enable an application of CR on datasets with more than one effect variable we assume knowledge of the SCM including Y for CR as well and draw ground-truth interventional samples from the SCM instead of identifying the interventional distribution from observational data.

For *3var-causal* and *3var-noncausal* we configured NSGA-II to optimize over 600 generations with a population size of 300, for *5var-skill* and *7var-covid* 1000 generations with 500 individuals were used. For all experiments the crossover probability was 0.3 and the mutation probability 0.05. For all settings continuous variables were rounded to 1 decimal point. For the 3 variable settings a standard `sklearn LogisticRegression` was used, for the refits without penality. For the nonlinear dataset a `RandomForestClassifier` with max depth 30, 50 estimators and balanced subsampling was applied. The experimental results were computed on a Quad core Intel Core i7-7700 Kaby Lake processor. For each setting, the experiments took between 24 to 48 hours.

D Proofs

As follows we provide the full proofs for Propositions 1 - 5.

D.1 Linking individualized prediction with γ^{ind}, Proof of Proposition 1

Proposition 1. *The expected individualized post-recourse score is equal to the individualized improvement probability* $\gamma^{ind}(x^{pre}, a) := P(Y^{post} = 1 | x^{pre}, do(a))$*, i.e.*

$$E[\hat{h}^{*,ind}(x^{post}) | x^{pre}, do(a)] = \gamma^{ind}(a).$$

Proof: It holds that

$$\begin{aligned} & E[h^{*,ind}(x^{post}) | x^{pre}, do(a)] \\ & = E[E[Y | x^{pre}, x^{post}] | x^{pre}, do(a)] \\ & \overset{\text{total exp.}}{=} E[Y | x^{pre}, do(a)] \\ & = \gamma^{ind}(a). \end{aligned}$$

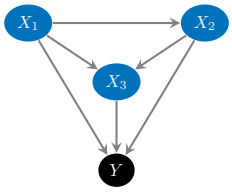
D.2 Intervention stability w.r.t. ICR actions, Proposition 2

Proposition 2. *Given nonzero cost for all interventions, ICR exclusively suggests actions on causes of* Y*. Assuming*

discrete inference and `numpyro` (Phan, Pradhan, and Jankowiak 2019) for MCMC inference of continuous variables. For our experiments we used the analytical formulas presented in B

[22] All packages are open source. For detailed license information we refer to the respective package websites.

[23] A short comment on the choice of a non-adaptive threshold can be found in E.2.

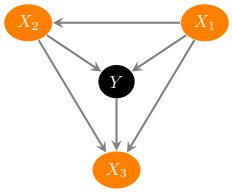


(a) Causal graph

$$
\begin{aligned}
X_1 &:= U_1, & U_1 &\sim N(0,1)\\
X_2 &:= X_1 + U_2, & U_2 &\sim N(0,1)\\
X_3 &:= X_1 + X_2 + U_3, & U_3 &\sim N(0,1)\\
Y &\sim [\sigma(X_1 + X_2 + X_3) \leq U_Y], & U_Y &\sim Unif(0,1)
\end{aligned}
$$

(b) Structural Equations

Figure 7: SCM for *3var-causal*. The cost function is given as $cost(a) = \delta_1 + \delta_2 + \delta_3$, where δ is the vector of absolute changes to the intervened upon variables. E.g., for $do(a) = do(X_1 = x_1')$, $\delta_1 = |x_1' - x_1|$ and $\delta_2 = \delta_3 = 0$

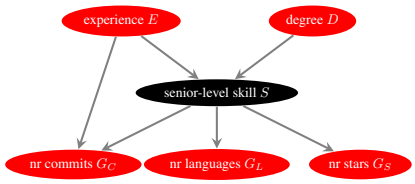


(a) Causal graph

$$
\begin{aligned}
X_1 &:= U_1, & U_1 &\sim N(0,1)\\
X_2 &:= X_1 + U_1, & U_1 &\sim N(0,1)\\
Y &:= [\sigma(X_1 + X_2) \leq U_Y], & U_Y &\sim Unif(0,1)\\
X_3 &:= X_1 + X_2 + Y + U_3, & U_3 &\sim N(0,0.1)
\end{aligned}
$$

(b) Structural Equations

Figure 8: SCM for *3var-noncausal* with $cost(a) = \delta_1 + \delta_2 + \delta_3$.

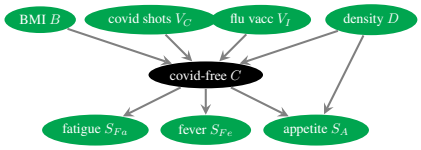


(a) Causal graph

$$
\begin{aligned}
E &:= U_E; U_E \sim GaP(8, 8/3)\\
D &:= U_D; U_D \sim Cat(0.4, 0.2, 0.3, 0.1)\\
S &:= [\sigma(-10 + 3E + 4D)) \leq U_S]; U_S \sim Unif(0,1)\\
G_C &:= 10E(11 + 100D) + U_{G_C}; U_{G_C} \sim GaP(40, 40/4)\\
G_L &:= \sigma(10S) + U_{G_L}; U_{G_L} \sim GaP(2, 2/4)\\
G_S &:= 10S + U_{G_S}; U_{G_S} \sim GaP(5, 5/4)
\end{aligned}
$$

(b) Structural Equations

Figure 9: SCM for *5var-skill* with $cost(a) = 5\delta_E + 5\delta_D + 0.0001\delta_{G_C} + 0.01\delta_{G_L} + 0.1\delta G_S$.

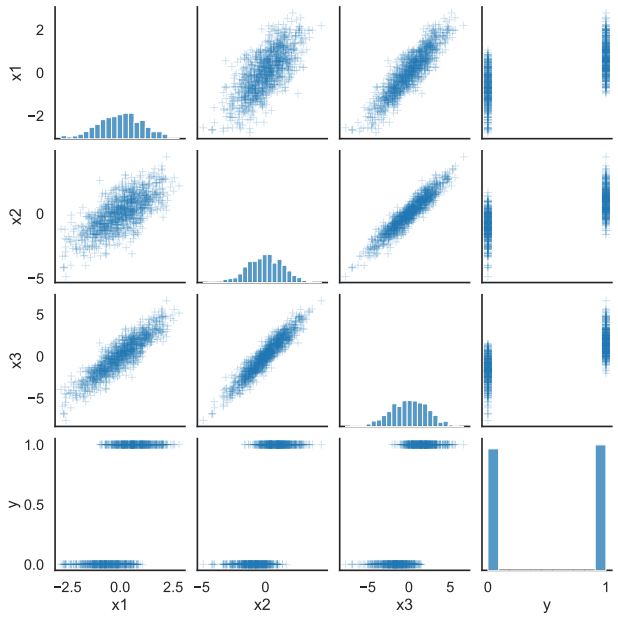


(a) Causal graph

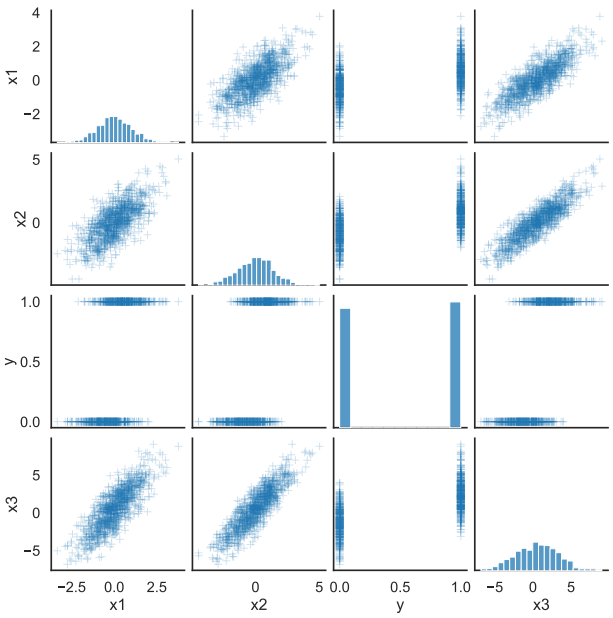
$$
\begin{aligned}
D &:= U_D; U_D \sim \Gamma(4, 4/3)\\
V_I &:= U_{V_I}; U_{V_I} \sim Bern(0.39)\\
V_C &:= U_{V_C}; U_{V_C} \sim Cat(0.24, 0.02, 0.15, 0.59)\\
B &:= U_B; U_B \sim N(0,1)\\
C &:= \left[\sigma(-(-3 + D - V_I - 2.5V_C + 0.2B^2)) \leq U_C\right];\\
U_C &\sim Unif(0,1)\\
S_A &:= [\sigma(-2C) \leq U_{S_A}]; U_{S_A} \sim Unif(0,1)\\
S_{Fe} &:= [\sigma(5 - 9C) \leq U_{S_{Fe}}]; U_{S_{Fe}} \sim Unif(0,1)\\
S_{Fa} &:= \left[\sigma(-1 + B^2 - 2C) \leq U_{S_{Fa}}\right];\\
U_{S_{Fa}} &\sim Unif(0,1)
\end{aligned}
$$

(b) Structural Equations

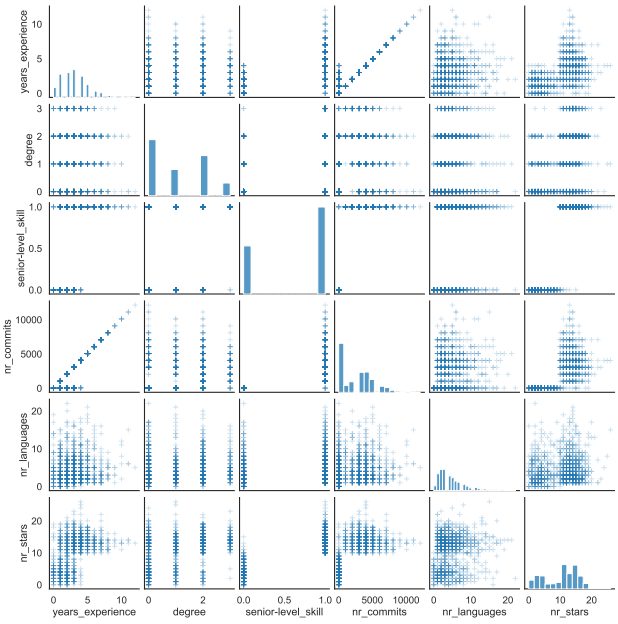
Figure 10: SCM for *7var-covid* with cost function $cost(a) = \delta_D + \delta_{V_I} + \delta_{V_C} + \delta_B + \delta_{S_A} + \delta_{S_{Fe}} + \delta_{S_{Fa}}$.



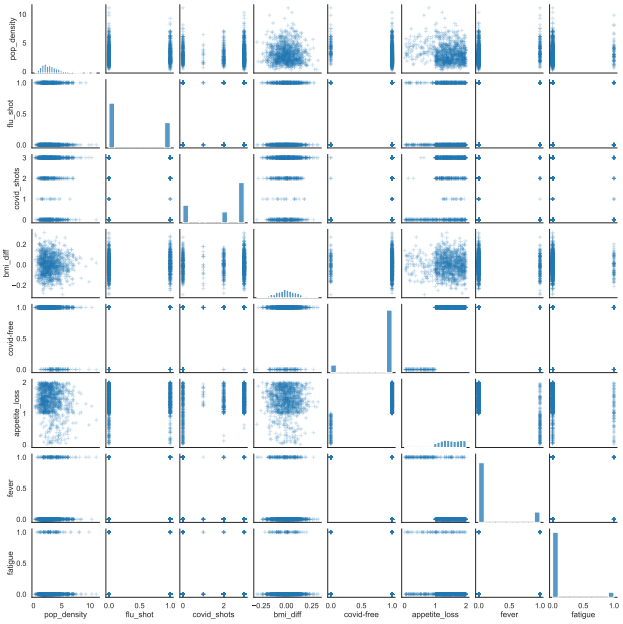
(a) Pairplot for *3var-causal*.



(b) Pairplot for *3var-noncausal*.



(c) Pairplot for *5var-skill*.



(d) Pairplot for *7var-covid*.

Figure 11: Pairplots for the SCMs.

Table 3: Results for 3var-causal.

3var-causal	$\overline{\gamma}$ / $\overline{\eta}$	$\gamma_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}^{individ.}$	$\pm$	$\eta_{\text{obs.}}^{\text{refit}}$	$\pm$	$\emptyset$ cost	$\pm$
CE	-	0.41	0.09	1.00	0.00	-	-	0.60	0.20	3.08	0.41
ind. CR	0.75	0.47	0.10	1.00	0.00	-	-	0.70	0.10	2.46	0.37
ind. CR	0.85	0.44	0.08	1.00	0.00	-	-	0.72	0.12	2.39	0.25
ind. CR	0.90	0.47	0.09	1.00	0.00	-	-	0.72	0.14	2.36	0.35
ind. CR	0.95	0.49	0.07	1.00	0.00	-	-	0.67	0.10	2.44	0.31
subp. CR	0.75	0.46	0.11	0.86	0.04	-	-	0.64	0.14	2.66	0.41
subp. CR	0.85	0.43	0.08	0.93	0.02	-	-	0.69	0.14	2.64	0.32
subp. CR	0.90	0.45	0.09	0.96	0.02	-	-	0.70	0.15	2.73	0.42
subp. CR	0.95	0.48	0.09	0.98	0.01	-	-	0.64	0.14	2.86	0.41
ind. ICR	0.75	0.79	0.06	0.98	0.02	1.0	0.0	0.96	0.03	3.27	0.50
ind. ICR	0.85	0.86	0.03	1.00	0.01	1.0	0.0	0.97	0.02	3.82	0.30
ind. ICR	0.90	0.90	0.02	1.00	0.01	1.0	0.0	0.98	0.03	3.70	0.31
ind. ICR	0.95	0.95	0.01	1.00	0.00	1.0	0.0	0.99	0.01	4.08	0.24
subp. ICR	0.75	0.75	0.04	0.93	0.04	-	-	0.90	0.04	3.34	0.49
subp. ICR	0.85	0.87	0.03	0.98	0.01	-	-	0.96	0.02	4.05	0.29
subp. ICR	0.90	0.89	0.02	0.99	0.01	-	-	0.97	0.02	3.87	0.25
subp. ICR	0.95	0.94	0.02	1.00	0.00	-	-	0.99	0.01	4.22	0.28

Table 4: Results for 3var-noncausal

3var-noncausal	$\overline{\gamma}$ / $\overline{\eta}$	$\gamma_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}^{individ.}$	$\pm$	$\eta_{\text{obs.}}^{\text{refit}}$	$\pm$	$\emptyset$ cost	$\pm$
CE	-	0.17	0.03	0.98	0.04	-	-	0.67	0.15	2.28	0.26
ind. CR	0.75	0.25	0.03	1.00	0.00	-	-	0.70	0.13	2.28	0.21
ind. CR	0.85	0.24	0.02	1.00	0.00	-	-	0.73	0.13	2.29	0.17
ind. CR	0.90	0.24	0.04	1.00	0.00	-	-	0.71	0.11	2.24	0.16
ind. CR	0.95	0.23	0.04	1.00	0.00	-	-	0.73	0.12	2.18	0.32
subp. CR	0.75	0.22	0.03	0.91	0.03	-	-	0.63	0.15	2.18	0.12
subp. CR	0.85	0.19	0.03	0.95	0.02	-	-	0.67	0.15	2.33	0.21
subp. CR	0.90	0.19	0.03	0.97	0.01	-	-	0.65	0.14	2.42	0.19
subp. CR	0.95	0.19	0.03	0.99	0.01	-	-	0.69	0.14	2.26	0.32
ind. ICR	0.75	0.77	0.03	0.93	0.02	0.79	0.03	0.93	0.02	2.16	0.11
ind. ICR	0.85	0.86	0.02	0.99	0.01	0.90	0.02	0.99	0.01	2.51	0.08
ind. ICR	0.90	0.91	0.03	1.00	0.00	0.94	0.01	1.00	0.00	3.00	0.08
ind. ICR	0.95	0.96	0.02	0.98	0.07	0.98	0.01	0.98	0.08	3.32	0.16
subp. ICR	0.75	0.69	0.03	0.77	0.05	-	-	0.76	0.05	2.11	0.20
subp. ICR	0.85	0.82	0.03	0.93	0.02	-	-	0.92	0.02	2.42	0.11
subp. ICR	0.90	0.89	0.03	0.98	0.01	-	-	0.97	0.01	2.86	0.13
subp. ICR	0.95	0.94	0.02	0.97	0.10	-	-	0.96	0.12	3.19	0.15

Table 5: Results for 5var-skill

5var-skill	$\overline{\gamma}$ / $\overline{\eta}$	$\gamma_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}^{individ.}$	$\pm$	$\eta_{\text{obs.}}^{\text{refit}}$	$\pm$	$\emptyset$ cost	$\pm$
CE	-	0.00	0.00	1.00	0.00	-	-	0.76	0.14	1.34	1.28
ind. CR	0.75	0.00	0.00	1.00	0.00	-	-	0.86	0.11	0.27	0.28
ind. CR	0.85	0.00	0.00	1.00	0.00	-	-	0.81	0.14	0.24	0.20
ind. CR	0.90	0.00	0.01	1.00	0.00	-	-	0.70	0.15	0.10	0.00
ind. CR	0.95	0.00	0.00	1.00	0.00	-	-	0.66	0.16	0.11	0.03
subp. CR	0.75	0.00	0.00	1.00	0.00	-	-	0.85	0.11	4.06	4.97
subp. CR	0.85	0.00	0.00	1.00	0.00	-	-	0.80	0.15	0.24	0.19
subp. CR	0.90	0.00	0.01	1.00	0.00	-	-	0.70	0.15	0.10	0.01
subp. CR	0.95	0.00	0.00	1.00	0.00	-	-	0.66	0.15	0.12	0.04
ind. ICR	0.75	0.94	0.02	0.94	0.02	0.94	0.02	0.94	0.02	4.95	5.32
ind. ICR	0.85	0.94	0.01	0.93	0.02	0.94	0.01	0.93	0.02	9.80	0.27
ind. ICR	0.90	0.96	0.02	0.96	0.02	0.96	0.02	0.96	0.02	10.38	0.23
ind. ICR	0.95	0.98	0.01	0.98	0.01	0.98	0.01	0.98	0.01	11.23	0.21
subp. ICR	0.75	0.93	0.01	0.93	0.02	-	-	0.93	0.01	4.72	5.08
subp. ICR	0.85	0.94	0.01	0.94	0.01	-	-	0.94	0.02	9.74	0.17
subp. ICR	0.90	0.96	0.01	0.96	0.01	-	-	0.96	0.01	10.46	0.53
subp. ICR	0.95	0.97	0.01	0.97	0.01	-	-	0.97	0.01	10.88	0.21

Table 6: Results for 7var-covid

7var-covid	$\overline{\gamma}$ / $\overline{\eta}$	$\gamma_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}$	$\pm$	$\eta_{\text{obs.}}^{individ.}$	$\pm$	$\eta_{\text{obs.}}^{\text{refit}}$	$\pm$	$\emptyset$ cost	$\pm$
CE	-	0.00	0.00	1.00	0.00	-	-	1.00	0.00	0.60	0.12
ind. CR	0.75	0.01	0.00	1.00	0.00	-	-	0.99	0.01	0.56	0.02
ind. CR	0.85	0.00	0.00	1.00	0.00	-	-	0.99	0.00	0.55	0.02
ind. CR	0.90	0.00	0.00	1.00	0.00	-	-	1.00	0.00	0.55	0.03
ind. CR	0.95	0.00	0.00	1.00	0.00	-	-	0.99	0.01	0.54	0.07
subp. CR	0.75	0.01	0.01	0.92	0.02	-	-	0.91	0.02	0.52	0.03
subp. CR	0.85	0.00	0.01	0.97	0.01	-	-	0.96	0.01	0.75	0.40
subp. CR	0.90	0.00	0.00	0.98	0.01	-	-	0.98	0.01	0.55	0.03
subp. CR	0.95	0.00	0.00	0.99	0.01	-	-	0.98	0.01	0.51	0.07
ind. ICR	0.75	0.81	0.03	0.81	0.03	0.82	0.04	0.81	0.03	1.26	0.02
ind. ICR	0.85	0.85	0.03	0.85	0.03	0.86	0.03	0.85	0.03	1.14	0.44
ind. ICR	0.90	0.89	0.03	0.89	0.03	0.90	0.02	0.89	0.03	1.61	0.02
ind. ICR	0.95	0.95	0.01	0.95	0.01	0.95	0.01	0.95	0.01	1.97	0.06
subp. ICR	0.75	0.61	0.04	0.61	0.04	-	-	0.61	0.04	1.06	0.03
subp. ICR	0.85	0.73	0.03	0.73	0.03	-	-	0.73	0.03	1.09	0.34
subp. ICR	0.90	0.81	0.04	0.81	0.04	-	-	0.81	0.04	1.42	0.05
subp. ICR	0.95	0.90	0.03	0.90	0.03	-	-	0.90	0.03	1.73	0.06

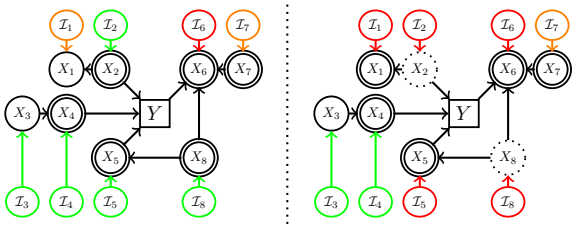


Figure 12: A schematic drawing illustrating under which interventions $I_1, \ldots, I_8$ the Markov blanket (double circle) is intervention stable. In this setting, we consider the intervention variables to be independent treatment variables: We would like to know how the different actions influence the conditional distribution, irrespective of how likely they are to be applied. Therefore, they are modeled as parentless variables. Green indicates intervention stability, red indicates no intervention stability. Orange indicates intervention stability of non-causal variables. Dotted variables are not observed. *Left:* Since all endogenous variables are observed, $MB_O(Y)$ is stable w.r.t. interventions on every endogenous cause of Y (Proposition 3). *Right:* Unobserved variables (X_2, X_8) open paths between interventions on causes and Y.

causal sufficiency, for any optimal predictor the conditional distribution of Y given the variables that the model uses X_S (i.e. $P(Y|X_S)$) is stable w.r.t interventions on causes. Therefore, optimal predictors are intervention stable w.r.t. ICR actions.

Proof: We prove the statement in six steps.

ICR only intervenes on causes: The goal of meaningful recourse is to improve Y with minimal cost. Only interventions on causes alter Y. Consequently, actions on non-causes of Y would not be suggested by meaningful recourse.

Given causal sufficiency, a graph $\mathcal{G}$ and an endogenous Y, the set of endogeneous direct parents, direct effects and direct parents of effects are the minimal d-separating set $S_\mathcal{G}$: Standard result, see e.g. Peters, Janzing, and Schölkopf (2017), Proposition 6.27.

The set $S_{\mathcal{G}^*}$ in the augmented graph $\mathcal{G}^*$ coincides with $S_\mathcal{G}$: The minimal d-separating set contains direct causes, direct effects and direct parents of direct effects. I_l is never a direct cause of X_l. Also, since I_l has no endogenous causes, it cannot be a direct effect. Furthermore, since we restrict interventions to be performed on causes, I_l cannot be a direct parent of a direct effect.

$S_\mathcal{G}$ is intervention stable: As follows, all intervention variables are d-separated from Y in $\mathcal{G}^*$ by $S_\mathcal{G}$. Therefore $S_\mathcal{G}$ is intervention stable. An example is given in Figure 12.

Then also the markov blanket is intervention stable: Since d-separation implies independence $MB(Y) \subseteq S_\mathcal{G}$. Therefore, if $X_T \perp Y | X_{MB(Y)}$ then also $X_T \perp Y | S_\mathcal{G}$. If any element $s \in S_\mathcal{G}$ it holds that $s \notin MB(Y)$, then it must hold that $X_s \perp Y | X_{MB(Y)}$. Therefore, if $X_T \perp Y | X_{MB(Y)}, X_s$ then also $X_T \perp Y | X_{MB(Y)}$ and therefore any independence entailed by $S_\mathcal{G}$ also holds for $MB(Y)$. Since Pfister et al. (2021) only require the independence that is implied by d-separation in their invariant conditional proof, the same implication holds for the $MB(Y)$. As follows, $P(Y|X_{MB(Y)})$ is invariant with respect to interventions on any set of endogenous causes.

Then any superset of the markov blanket is intervention stable: We prove the statement by contradiction. The markov blanket d-separates the target variable Y from any other set of variables. If adding a set of variables S_1 to the markov blanket would open a path to any other set of variables S_2, then it would hold that $S := S_1 \cup S_2$ is not d-separated from Y ($P(Y|MB(Y)) = P(Y|MB(Y), S_1, S_2) \neq P(Y|MB(Y), S_1) = P(Y|MB(Y))$)

D.3 Linking observational prediction and γ^{sub}, Proposition 3

Proposition 3. *Given causal sufficiency and positivity*[24]*, for interventions on causes the expected subgroup-wide optimal score h^* is equal to the subgroup-wide improvement probability $\gamma^{sub}(a) := P(Y^{post} = 1 | do(a), x^{pre}_{G_a})$, i.e.*

$$E[\hat{h}^*(x^{post}) | x^{pre}_{G_a}, do(a)] = \gamma^{sub}(a).$$

Proof: The proposition follows from Proposition 2. More specifically

$$\begin{aligned} E[h^*(x^{post,a}) | x^{pre}_G, a] & && (15)\\ &= E[E[Y|x^{post,a}] | x^{pre}_G, a] && (16)\\ &\overset{\text{total exp.}}{=} E[Y | x^{pre}_G, a] && (17)\\ &\overset{\text{def. } \gamma^{sub}}{=} \gamma^{sub}(a). && (18) \end{aligned}$$

D.4 Acceptance Bound, Proof of Proposition 4

Proposition 4. *Let g be a predictor with $E[g(x^{post}) | x^{pre}_S, do(a)] = \gamma(x^{pre}_S, a)$. Then for a decision threshold t the post-recourse acceptance probability $\eta(t; x^{pre}_S, a) := P(g(x^{post}) > t | x^{pre}_S, do(a))$ is lower bounded:*

$$\eta(t; x^{pre}_S, a) \geq \frac{\gamma(x^{pre}_S, a) - t}{1 - t}.$$

Proof: Positivity ($p^{pre}(x^{post}) > 0$) is necessary for subpopulation-based ICR since only then we can assume that the model is actually optimal for any input that it receives. The problem is discussed in more detail in Hernán MA (2020); Neal (2020).

As follows we denote $\hat{h}^*$ as the random variable indicating the predictions of the post-recourse predictors described in Section 5.

From Propositions 1 and 3, for both individualized and subpopulation-based post-recourse predictors we know that

$$E[\hat{h}(x^{post,a})^* | x^{pre}_S, do(a)] = \gamma(x^{pre}_S, a).$$

We decompose the expected prediction

[24] Positivity ensures that the post-recourse observation lies within the observational support , where the model was trained (i.e., $p^{pre}(x^{post}) > 0$), (Neal 2020)).

$$\gamma(x_S^{pre}, a) \tag{19}$$
$$= E[\hat{h}^*|x_S^{pre}, a] \tag{20}$$
$$= \begin{array}{l} E[\hat{h}^*|\hat{h}^* > t]P(\hat{h}^* > t) \\ +E[\hat{h}^*|\hat{h}^* \leq t]P(\hat{h}^* \leq t) \end{array} \Big|_{x_S^{pre},a} \tag{21}$$
$$= \begin{array}{l} E[\hat{h}^*|\hat{h}^* > t]P(\hat{h}^* > t) \\ +E[\hat{h}^*|\hat{h}^* \leq t](1 - P(\hat{h}^* > t)) \end{array} \Big|_{x_S^{pre},a} \tag{22}$$
$$= \begin{array}{l} E[\hat{h}^*|\hat{h}^* > t]P(\hat{h}^* > t) \\ +E[\hat{h}^*|\hat{h}^* \leq t] - P(\hat{h}^* > t)E[\hat{h}^*|\hat{h}^* \leq t] \end{array} \Big|_{x_S^{pre},a} \tag{23}$$
$$= \begin{array}{l} E[\hat{h}^*|\hat{h}^* \leq t] \\ +P(\hat{h}^* > t)\Big(E[\hat{h}^*|\hat{h}^* > t] - E[\hat{h}^*|\hat{h}^* \leq t]\Big) \end{array} \Big|_{x_S^{pre},a} \tag{24}$$

which can be reformulated to yield the acceptance rate η:

$$\frac{\gamma - E[\hat{h}^*|\hat{h}^* \leq t]}{E[\hat{h}^*|\hat{h}^* > t] - E[\hat{h}^*|\hat{h}^* \leq t]}\Big|_{x_S^{pre},a} \tag{25}$$
$$= P(\hat{h}^* > t|x_S^{pre}, a) = \eta(x_S^{pre}, a). \tag{26}$$

It holds that $E[\hat{h}^{*,ind}|\hat{h}^* \leq t] = FNR(t)$ and $E[\hat{h}^*|\hat{h}^* > t] = TPR(t)$.

We can show that $E[\hat{h}^*|\hat{h}^* \leq t] \leq t$:

$$0 \leq FNR(t|x_S^{pre}, a) \tag{27}$$
$$= P(Y^{a,post} = 1|h^* \leq t, x_S^{pre}, a) \tag{28}$$
$$= E[Y^{a,post}|h^* \leq t, x_S^{pre}, a] \tag{29}$$
$$= E[E[Y^{a,post}|x^{post,a}]|h^* \leq t, x_S^{pre}, a] \tag{30}$$
$$= E[h^*|h^* \leq t, x_S^{pre}, a] \tag{31}$$
$$\leq t \tag{32}$$

and analog that $1 \geq TPR(t) \geq t$. Therefore

$$\eta(t, x_S^{pre}, a) \tag{33}$$
$$= \frac{\gamma - FNR(t)}{TPR(t) - FNR(t)}\Big|_{x_S^{pre},a} \tag{34}$$
$$\geq \frac{\gamma(x_S^{pre}, a) - FNR(t)}{1 - FNR(t)} \geq \frac{\gamma(x_S^{pre}, a) - t}{1 - t}. \tag{35}$$

D.5 Individualized post-recourse prediction, proof of Proposition 5

Proposition 5. *In general, the individualized post-recourse predictor can be estimated as*

$$p(y^{post}|x^{pre}, x^{post}, do(a)) \tag{36}$$
$$= \frac{\int_{\mathcal{U}} p(y^{post}, x^{post}|u, do(a))p(u|x^{pre})du}{\sum_{y' \in \{0,1\}} \left(\int_{\mathcal{U}} p(y', x^{post}|u, do(a))p(u|x^{pre})du\right)} \tag{37}$$

Given binary decision problems with invertible structural equations, the individualized post-recourse prediction function reduces to

$$p(y^{post}|x^{post}, x^{pre}, do(a)) \tag{38}$$
$$= \frac{p(U_{-I} = f_{do(a)}^{-1}(y^{post}, x^{post})|x^{pre}, do(a))}{\sum_{y' \in \{0,1\}} p(U_{-I} = f_{do(a)}^{-1}(y', x^{post})|x^{pre}, do(a))}. \tag{39}$$

Proof: It holds that

$$p(y^{post}|x^{pre}, x^{post}, do(a)) \tag{40}$$
$$\overset{\text{def. cond.}}{=} \frac{p(y^{post}, x^{post}|x^{pre}, do(a))}{p(x^{post}|x^{pre}, do(a))} \tag{41}$$
$$\tag{42}$$

We can reformulate the conditional distribution $p(y^{post}, x^{post}|x^{pre}, do(a))$ as two parts, one that describes the probability of a state of the context given x^{pre}, and one that describes the probability of a post-recourse state x^{post}, y^{post} given a certain noise state u and $do(a)$.

$$p(y^{post}, x^{post}|x^{pre}, do(a)) \tag{43}$$
$$\overset{\text{marginal.}}{=} \int_{\mathcal{U}} p(y^{post}, x^{post}, u|x^{pre}, do(a))du \tag{44}$$
$$\overset{\text{chain rule}}{=} \int_{\mathcal{U}} p(y^{post}, x^{post}|u, x^{pre}, do(a))p(u|x^{pre})du \tag{45}$$
$$(y, x)^{post} \overset{}{=} \perp x^{pre}|u \int_{\mathcal{U}} p(y^{post}, x^{post}|u, do(a))p(u|x^{pre})du. \tag{46}$$

In combination we yield

$$p(y^{post}|x^{pre}, x^{post}, do(a)) \tag{47}$$
$$= \frac{\int_{\mathcal{U}} p(y^{post}, x^{post}|u, do(a))p(u|x^{pre})du}{\int_{\mathcal{Y}} \left(\int_{\mathcal{U}} p(y', x^{post}|u, do(a))p(u|x^{pre})du\right) dy'} \tag{48}$$
$$= \frac{\int_{\mathcal{U}} p(y^{post}, x^{post}|u, do(a))p(u|x^{pre})du}{\sum_{y' \in 0,1} \left(\int_{\mathcal{U}} p(y', x^{post}|u, do(a))p(u|x^{pre})du\right)} \tag{49}$$

For a setting with invertible structural equations this reduces to

$$p(y^{post}|x^{post}, x^{pre}, do(a)) \tag{50}$$
$$= \frac{p(y^{post}, x^{post}|x^{pre}, do(a))}{p(x^{post}|x^{pre}, do(a))} \tag{51}$$
$$= \frac{p(U_{-I} = f^{-1}(y^{post}, x^{post})|x^{pre}, do(a))}{\sum_{y' \in \{0,1\}} p(U_{-I} = f^{-1}(y^{post}, x^{post})|x^{pre}, do(a))}. \tag{52}$$

where $-I$ is the index set for variables that have not been intervened on (since the noise terms for the intervened upon variables are isolated variables in the interventional graph).

E Misc

E.1 Negative Result: Algorithmic recourse is neither meaningful nor robust

In the introduction we claimed that CR recommendations (Karimi et al. 2020b; Karimi, Schölkopf, and Valera 2021) may not lead to improvement. Now, we formally demonstrate the case on the Covid hospital admission example (Figure 1) which we extend with the full structural causal model (Example 1). Furthermore, we show that CR is not robust to refits of the model on mixed pre- and post-recourse data. All code is publicly available via https://anonymous.4open.science/r/icr-aaai/README.md.

Example 1. *Let V indicate whether someone is fully vaccinated, Y indicate whether someone is free of Covid and S whether someone is asymptomatic. The data is generated by the following structural causal model (SCM) entailing the causal graph depicted in Figure 1:*

$$V := U_V, \qquad U_V \sim Bern(0.5) \tag{53}$$
$$Y := V + U_Y \mod 2, \qquad U_Y \sim Bern(0.09) \tag{54}$$
$$S := Y + U_S \mod 2, \qquad U_S \sim Bern(0.05) \tag{55}$$

For prediction, a `sklearn` *logistic regression model is fit on* 2000 *samples, yielding* $\hat{h}$ *with* $\beta_v \approx 3.7$, $\beta_s \approx 5.1$, $\beta_0 \approx -4.3$. *Visitors are allowed to enter the hospital if* $\hat{h} < 0.5$. *Intervening on (flipping)* V *and* S *costs* 0.5 *and* 0.1 *respectively.*

Lack of improvement: Given a decision threshold of 0.5, the model admits everyone without symptoms ($S = 1$), irrespective of their vaccination status V. Therefore, in order to revert rejections ($S = 0$), both individualized and subpopulation-based CR suggest removing the symptoms S ($do(S = 1)$, for instance by taking cough drops). However, since they only treat the symptoms S, the actual Covid risk Y is unaffected: none of the recourse-implementing individuals actually improve. We say the predictor is *gamed*.

Lack of robustness: For individuals who implement recourse the association between symptom state S and Covid risk Y is broken. Thus, the predictive power of the model for recourse-seeking individual drops from ≈ 95 percent pre-recourse to ≈ 5 percent post-recourse.[25] A refit of the model on a mix pre- and post-recourse data (2000 samples each) yields $\hat{h}$ with $\beta_V \approx 4.1, \beta_S \approx 3.3, \beta_0 \approx -4.8$. Since the association between symptom state and disease status is broken post-recourse, the new model rejects individuals if they are not vaccinated, irrespective of their symptom state. For that reason, recourse recommendations that were designed for the original model only lead to acceptance by the refitted model for those individuals who happened to be vaccinated anyway.

The example demonstrates that CR recommendations are prone to gaming the predictor and therefore may neither lead to improvement nor be robust to model refits.

[25] The previously wrongly-rejected individuals are correctly classified after implementing recourse.

E.2 Interpretability of improvement confidence γ

Counterfactuals are concerned with changing the inputs to the model such that the model prediction changes in the desired way. Since the prediction function is deterministic and accessible, the post-recourse prediction can be determined exactly.

In contrast CR and ICR deal with the effects of real-world interventions on real-world variables. As such, the effects of recourse actions on the covariates (and the underlying prediction target) cannot be determined exactly. Therefore both CR and ICR have to deal with uncertainty.

CR deals with this uncertainty by phrasing the optimization objective for CR in terms of an expectation over the prediction distribution and by using an action-adaptive confidence threshold. This threshold `thresh` bounds the expected prediction away from the model's decision threshold (e.g. $t = 0.5$). Using the conservativeness parameters, the user can roughly steer how far the expected prediction shall be away from the decision boundary.

In contrast, ICR deals with the uncertainty by letting the user specify the confidence γ, which can be intuitively interpreted as improvement probability (whereas the expected prediction cannot be interpreted as acceptance probability). A lower-bound on the acceptance probability for a combination of γ and t is given in Proposition 4. Furthermore, we can estimate the individualized and subpopulation-based acceptance rates for a specific situation (a, x^{pre}) as detailed in B.1 and B.3. The human-interpretable improvement and acceptance confidences are vital for the explainee to make an informed decision.

In order to allow a direct comparison of the methods, we rephrase the CR objective to optimize the acceptance probability η in our experiments.

E.3 Imbalance between standard predictors and individualized ICR recommendations

In Section 6 we argued that there is an imbalance in predictive capability between (optimal) observational predictors and the pre-recourse SCM (which used to predict γ^{ind}). We illustrate the problem on a simple example.

Example 2. *Let there be a three variable chain $X_1 \to Y \to X_2$ where at every step the value is incremented by one with 50% chance and the maximum value is set to 2 ($X_1 := U_1$, $Y := X_1 + U_Y$, $X_2 := min(2, Y + U_2)$ where $U_1, U_2, U_Y \sim Bern(0.5)$). Let us assume a factual observation $x^{pre} = (0, 2)$ and action $a = do(X_1 = 1)$ yielding $x^{post} = (1, 2)$. For the observation $x^{pre} = (0, 2)$ we can infer that U_Y must have been 1, since two increments are needed to get from 0 to 2. However, from the post-intervention observation $x^{post} = (1, 2)$ we cannot infer where the increment happened (U_Y or U_2). As a consequence, an optimal predictive model that only has access to x^{post} would predict that y^{post} for $x^{post} = (1, 2)$ could be 1 or 2 with equal likelihood. In contrast, with access to x^{pre} and the SCM we can infer that $y^{post} = 2$ since $U_Y = 1$.*

In the above example, given knowledge of the SCM, the pre-intervention observation x^{pre} and the performed action a we can already abduct U_Y perfectly and therefore correctly

determine the post-intervention state of Y (even without access to the post-intervention observation x^{post}). In contrast, with the post-recourse observation alone it is impossible to reconstruct U_Y and therefore impossible to determine the post-intervention state of Y.[26] In the context of ICR this means that the observational predictor's post-recourse predictions are not directly linked with γ: they may not honor the implementation of actions with $\gamma^{ind} = 1$. As a consequence, we suggested to use the SCM for post-recourse prediction in Section 6.

Supplementary References

Bashtannyk, D. M.; and Hyndman, R. J. 2001. Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36(3): 279–298.

Bingham, E.; Chen, J. P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; and Goodman, N. D. 2018. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*.

Bishop, C. M. 1994. Mixture density networks. Technical report, Aston University.

Bradbury, J.; Frostig, R.; Hawkins, P.; Johnson, M. J.; Leary, C.; Maclaurin, D.; Necula, G.; Paszke, A.; VanderPlas, J.; Wanderman-Milne, S.; and Zhang, Q. 2018. JAX: composable transformations of Python+NumPy programs.

Dandl, S.; Molnar, C.; Binder, M.; and Bischl, B. 2020. Multi-Objective Counterfactual Explanations. In Bäck, T.; Preuss, M.; Deutz, A.; Wang, H.; Doerr, C.; Emmerich, M.; and Trautmann, H., eds., *Parallel Problem Solving from Nature – PPSN XVI*, 448–469. Cham: Springer International Publishing. ISBN 978-3-030-58112-1.

Deb, K.; Pratap, A.; Agarwal, S.; and Meyarivan, T. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2): 182–197.

Dominguez-Olmedo, R.; Karimi, A.-H.; and Schölkopf, B. 2021. On the Adversarial Robustness of Causal Algorithmic Recourse. *arXiv preprint arXiv:2112.11313*.

Fortin, F.-A.; De Rainville, F.-M.; Gardner, M.-A.; Parizeau, M.; and Gagné, C. 2012. DEAP: Evolutionary Algorithms Made Easy. *Journal of Machine Learning Research*, 13: 2171–2175.

Geiger, D.; Verma, T.; and Pearl, J. 1990. Identifying independence in Bayesian networks. *Networks*, 20(5): 507–534.

Harris, C. R.; Millman, K. J.; van der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; Kern, R.; Picus, M.; Hoyer, S.; van Kerkwijk, M. H.; Brett, M.; Haldane, A.; del Río, J. F.; Wiebe, M.; Peterson, P.; Gérard-Marchant, P.; Sheppard, K.; Reddy, T.; Weckesser, W.; Abbasi, H.; Gohlke, C.; and Oliphant, T. E. 2020. Array programming with NumPy. *Nature*, 585(7825): 357–362.

[26]The optimal pre-recourse predictor $\hat{h}^*(x^{post})$ predicts 0.5 for both $y = 1$ and $y = 2$.

Hernán MA, R. J. 2020. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC.

Hothorn, T.; and Zeileis, A. 2021. Predictive distribution modeling using transformation forests. *Journal of Computational and Graphical Statistics*, 30(4): 1181–1196.

Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3): 90–95.

Jehi, L.; Ji, X.; Milinovich, A.; Erzurum, S.; Rubin, B. P.; Gordon, S.; Young, J. B.; and Kattan, M. W. 2020. Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11,672 patients. *Chest*, 158(4): 1364–1375.

Karimi, A.-H.; Schölkopf, B.; and Valera, I. 2021. Algorithmic Recourse: From Counterfactual Explanations to Interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 353362. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 265–277. virtual: Curran Associates, Inc.

Koller, D.; and Friedman, N. 2009. *Probabilistic graphical models: principles and techniques*. MIT press.

Laugel, T.; Lesot, M.-J.; Marsala, C.; Renard, X.; and Detyniecki, M. 2019. The Dangers of Post-Hoc Interpretability: Unjustified Counterfactual Explanations. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, IJCAI'19, 28012807. Macao, China: AAAI Press. ISBN 9780999241141.

Li, R.; Emmerich, M. T.; Eggermont, J.; Bäck, T.; Schütz, M.; Dijkstra, J.; and Reiber, J. H. 2013. Mixed integer evolution strategies for parameter optimization. *Evolutionary computation*, 21(1): 29–64.

Mahajan, D.; Tan, C.; and Sharma, A. 2020. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. arXiv:1912.03277.

Montandon, J. E.; Valente, M. T.; and Silva, L. L. 2021. Mining the technical roles of github users. *Information and Software Technology*, 131: 106485.

Neal, B. 2020. Introduction to causal inference from a machine learning perspective. *Course Lecture Notes (draft)*.

Page Jr, T. J. 1984. Multivariate statistics: A vector space approach. *JMR, Journal of Marketing Research (pre-1986)*, 21(000002): 236.

pandas development team, T. 2020. pandas-dev/pandas: Pandas.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.;

and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Pawelczyk, M.; Broelemann, K.; and Kasneci, G. 2020. On Counterfactual Explanations under Predictive Multiplicity. In Peters, J.; and Sontag, D., eds., *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, 809–818. Online: PMLR.

Pawelczyk, M.; Datta, T.; van-den Heuvel, J.; Kasneci, G.; and Lakkaraju, H. 2022. Algorithmic Recourse in the Face of Noisy Human Responses. *arXiv preprint arXiv:2203.06768*.

Pearl, J. 2009. *Causality*. Cambridge, UK: Cambridge University Press, 2 edition. ISBN 978-0-521-89560-6.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Pfister, N.; Williams, E. G.; Peters, J.; Aebersold, R.; and Bühlmann, P. 2021. Stabilizing variable selection and regression. *The Annals of Applied Statistics*, 15(3): 1220 – 1246.

Phan, D.; Pradhan, N.; and Jankowiak, M. 2019. Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

Rawal, K.; Kamar, E.; and Lakkaraju, H. 2021. Algorithmic Recourse in the Wild: Understanding the Impact of Data and Model Shifts. arXiv:2012.11788.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Trippe, B. L.; and Turner, R. E. 2018. Conditional density estimation with bayesian normalising flows. *arXiv preprint arXiv:1802.04908*.

Tsirtsis, S.; and Gomez Rodriguez, M. 2020. Decisions, counterfactual explanations and strategic behavior. *Advances in Neural Information Processing Systems*, 33: 16749–16760.

Upadhyay, S.; Joshi, S.; and Lakkaraju, H. 2021. Towards robust and reliable algorithmic recourse. *Advances in Neural Information Processing Systems*, 34.

Ustun, B.; Spangher, A.; and Liu, Y. 2019. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 1019. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Waskom, M. L. 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60): 3021.

Winkler, C.; Worrall, D.; Hoogeboom, E.; and Welling, M. 2019. Learning likelihoods with conditional normalizing flows. *arXiv preprint arXiv:1912.00042*.

Chapter 6

Paper V: General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models

Molnar, C., König, G., Herbinger, J., **Freiesleben, T.**, Dandl, S., Scholbeck, C., Casalicchio, C., Grosse-Wentrup, M. and Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. *In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science, vol 13200. Springer*, 39–68. doi: https://doi.org/10.1007/978-3-031-04083-2_4.

Author contributions:

C.M. initiated and coordinated the project. **T.F.** wrote Section 5.2 and partially wrote Section 4. and 11. **T.F.** and S.D. contributed paragraphs within each pitfall on local IML methods. **All authors** proofread and revised the paper. The other pitfalls were written by the co-authors.

General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models

Christoph Molnar[1,7( )], Gunnar König[1,4], Julia Herbinger[1], Timo Freiesleben[2,3], Susanne Dandl[1], Christian A. Scholbeck[1], Giuseppe Casalicchio[1], Moritz Grosse-Wentrup[4,5,6], and Bernd Bischl[1]

[1] Department of Statistics, LMU Munich, Munich, Germany
christoph.molnar.ai@gmail.com
[2] Munich Center for Mathematical Philosophy, LMU Munich, Munich, Germany
[3] Graduate School of Systemic Neurosciences, LMU Munich, Munich, Germany
[4] Research Group Neuroinformatics, Faculty for Computer Science, University of Vienna, Vienna, Austria
[5] Research Platform Data Science @ Uni Vienna, Vienna, Austria
[6] Vienna Cognitive Science Hub, Vienna, Austria
[7] Leibniz Institute for Prevention Research and Epidemiology - BIPS GmbH, Bremen, Germany

Abstract. An increasing number of model-agnostic interpretation techniques for machine learning (ML) models such as partial dependence plots (PDP), permutation feature importance (PFI) and Shapley values provide insightful model interpretations, but can lead to wrong conclusions if applied incorrectly. We highlight many general pitfalls of ML model interpretation, such as using interpretation techniques in the wrong context, interpreting models that do not generalize well, ignoring feature dependencies, interactions, uncertainty estimates and issues in high-dimensional settings, or making unjustified causal interpretations, and illustrate them with examples. We focus on pitfalls for global methods that describe the average model behavior, but many pitfalls also apply to local methods that explain individual predictions. Our paper addresses ML practitioners by raising awareness of pitfalls and identifying solutions for correct model interpretation, but also addresses ML researchers by discussing open issues for further research.

Keywords: Interpretable machine learning · Explainable AI

This work is funded by the Bavarian State Ministry of Science and the Arts (coordinated by the Bavarian Research Institute for Digital Transformation (bidt)), by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A, by the German Research Foundation (DFG), Emmy Noether Grant 437611051, and by the Graduate School of Systemic Neurosciences (GSN) Munich. The authors of this work take full responsibilities for its content.

A. Holzinger et al. (Eds.): xxAI 2020, LNAI 13200, pp. 39–68, 2022.
https://doi.org/10.1007/978-3-031-04083-2_4

1 Introduction

In recent years, both industry and academia have increasingly shifted away from parametric models, such as generalized linear models, and towards non-parametric and non-linear machine learning (ML) models such as random forests, gradient boosting, or neural networks. The major driving force behind this development has been a considerable outperformance of ML over traditional models on many prediction tasks [32]. In part, this is because most ML models handle interactions and non-linear effects automatically. While classical statistical models – such as generalized additive models (GAMs) – also support the inclusion of interactions and non-linear effects, they come with the increased cost of having to (manually) specify and evaluate these modeling options. The benefits of many ML models are partly offset by their lack of interpretability, which is of major importance in many applications. For certain model classes (e.g. linear models), feature effects or importance scores can be directly inferred from the learned parameters and the model structure. In contrast, it is more difficult to extract such information from complex non-linear ML models that, for instance, do not have intelligible parameters and are hence often considered black boxes. However, model-agnostic interpretation methods allow us to harness the predictive power of ML models while gaining insights into the black-box model. These interpretation methods are already applied in many different fields. Applications of interpretable machine learning (IML) include understanding pre-evacuation decision-making [124] with partial dependence plots [36], inferring behavior from smartphone usage [105,106] with the help of permutation feature importance [107] and accumulated local effect plots [3], or understanding the relation between critical illness and health records [70] using Shapley additive explanations (SHAP) [78]. Given the widespread application of interpretable machine learning, it is crucial to highlight potential pitfalls, that, in the worst case, can produce incorrect conclusions.

This paper focuses on pitfalls for model-agnostic IML methods, i.e. methods that can be applied to any predictive model. Model-specific methods, in contrast, are tied to a certain model class (e.g. saliency maps [57] for gradient-based models, such as neural networks), and are mainly considered out-of-scope for this work. We focus on pitfalls for global interpretation methods, which describe the expected behavior of the entire model with respect to the whole data distribution. However, many of the pitfalls also apply to local explanation methods, which explain individual predictions or classifications. Global methods include the partial dependence plot (PDP) [36], partial importance (PI) [19], accumulated local affects (ALE) [3], or the permutation feature importance (PFI) [12,19,33]. Local methods include the individual conditional expectation (ICE) curves [38], individual conditional importance (ICI) [19], local interpretable model-agnostic explanations (LIME) [94], Shapley values [108] and SHapley Additive exPlanations (SHAP) [77,78] or counterfactual explanations [26,115]. Furthermore, we distinguish between feature effect and feature importance methods. A feature effect indicates the direction and magnitude of a change in predicted outcome due to changes in feature values. Effect methods include

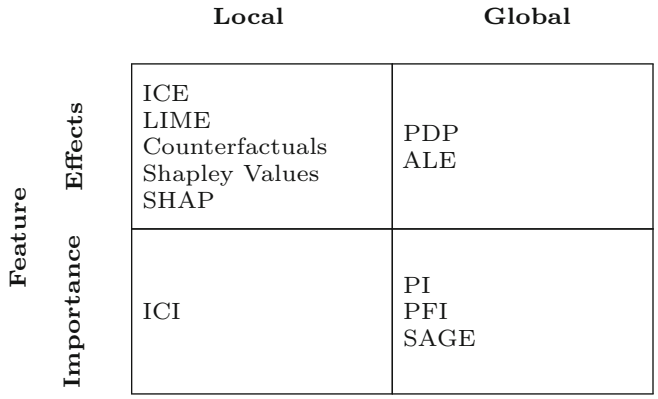


Fig. 1. Selection of popular model-agnostic interpretation techniques, classified as local or global, and as effect or importance methods.

Shapley values, SHAP, LIME, ICE, PDP, or ALE. Feature importance methods quantify the contribution of a feature to the model performance (e.g. via a loss function) or to the variance of the prediction function. Importance methods include the PFI, ICI, PI, or SAGE. See Fig. 1 for a visual summary.

The interpretation of ML models can have subtle pitfalls. Since many of the interpretation methods work by similar principles of manipulating data and "probing" the model [100], they also share many pitfalls. The sources of these pitfalls can be broadly divided into three categories: (1) application of an unsuitable ML model which does not reflect the underlying data generating process very well, (2) inherent limitations of the applied IML method, and (3) wrong application of an IML method. Typical pitfalls for (1) are bad model generalization or the unnecessary use of complex ML models. Applying an IML method in a wrong way (3) often results from the users' lack of knowledge of the inherent limitations of the chosen IML method (2). For example, if feature dependencies and interactions are present, potential extrapolations might lead to misleading interpretations for perturbation-based IML methods (inherent limitation). In such cases, methods like PFI might be a wrong choice to quantify feature importance.

Table 1. Categorization of the pitfalls by source.

Sources of pitfall	Sections
Unsuitable ML model	3, 4
Limitation of IML method	5.1, 6.1, 6.2, 9.1, 9.2
Wrong application of IML method	2, 5.2, 5.3, 7, 8, 9.3, 10

Contributions: We uncover and review general pitfalls of model-agnostic interpretation techniques. The categorization of these pitfalls into different sources is provided in Table 1. Each section describes and illustrates a pitfall, reviews possible solutions for practitioners to circumvent the pitfall, and discusses open issues that require further research. The pitfalls are accompanied by illustrative

examples for which the code can be found in this repository: https://github.com/compstat-lmu/code_pitfalls_iml.git. In addition to reproducing our examples, we invite readers to use this code as a starting point for their own experiments and explorations.

Related Work: Rudin et al. [96] present principles for interpretability and discuss challenges for model interpretation with a focus on inherently interpretable models. Das et al. [27] survey methods for explainable AI and discuss challenges with a focus on saliency maps for neural networks. A general warning about using and explaining ML models for high stakes decisions has been brought forward by Rudin [95], in which the author argues against model-agnostic techniques in favor of inherently interpretable models. Krishnan [64] criticizes the general conceptual foundation of interpretability, but does not dispute the usefulness of available methods. Likewise, Lipton [73] criticizes interpretable ML for its lack of causal conclusions, trust, and insights, but the author does not discuss any pitfalls in detail. Specific pitfalls due to dependent features are discussed by Hooker [54] for PDPs and functional ANOVA as well as by Hooker and Mentch [55] for feature importance computations. Hall [47] discusses recommendations for the application of particular interpretation methods but does not address general pitfalls.

2 Assuming One-Fits-All Interpretability

Pitfall: Assuming that a single IML method fits in all interpretation contexts can lead to dangerous misinterpretation. IML methods condense the complexity of ML models into human-intelligible descriptions that only provide insight into specific aspects of the model and data. The vast number of interpretation methods make it difficult for practitioners to choose an interpretation method that can answer their question. Due to the wide range of goals that are pursued under the umbrella term "interpretability", the methods differ in which aspects of the model and data they describe.

For example, there are several ways to quantify or rank the features according to their relevance. The relevance measured by PFI can be very different from the relevance measured by the SHAP importance. If a practitioner aims to gain insight into the relevance of a feature regarding the model's generalization error, a loss-based method (on unseen test data) such as PFI should be used. If we aim to expose which features the model relies on for its prediction or classification – irrespective of whether they aid the model's generalization performance – PFI on test data is misleading. In such scenarios, one should quantify the relevance of a feature regarding the model's prediction (and not the model's generalization error) using methods like the SHAP importance [76].

We illustrate the difference in Fig. 2. We simulated a data-generating process where the target is completely independent of all features. Hence, the features are just noise and should not contribute to the model's generalization error. Consequently, the features are not considered relevant by PFI on test data.

However, the model mechanistically relies on a number of spuriously correlated features. This reliance is exposed by marginal global SHAP importance.

As the example demonstrates, it would be misleading to view the PFI computed on test data or global SHAP as one-fits-all feature importance techniques. Like any IML method, they can only provide insight into certain aspects of model and data.

Many pitfalls in this paper arise from situations where an IML method that was designed for one purpose is applied in an unsuitable context. For example, extrapolation (Sect. 5.1) can be problematic when we aim to study how the model behaves under realistic data but simultaneously can be the correct choice if we want to study the sensitivity to a feature outside the data distribution.

For some IML techniques – especially local methods – even the same method can provide very different explanations, depending on the choice of hyperparameters: For counterfactuals, explanation goals are encoded in their optimization metrics [26,34] such as sparsity and data faithfulness; The scope and meaning of LIME explanations depend on the kernel width and the notion of complexity [8,37].

Solution: The suitability of an IML method cannot be evaluated with respect to one-fits-all interpretability but must be motivated and assessed with respect to well-defined interpretation goals. Similarly, practitioners must tailor the choice of the IML method and its respective hyperparameters to the interpretation context. This implies that these goals need to be clearly stated in a detailed manner *before* any analysis – which is still often not the case.

Open Issues: Since IML methods themselves are subject to interpretation, practitioners must be informed about which conclusions can or cannot be drawn given different choices of IML technique. In general, there are three aspects to be considered: (a) an intuitively understandable and plausible algorithmic construction of the IML method to achieve an explanation; (b) a clear mathematical axiomatization of interpretation goals and properties, which are linked by proofs and theoretical considerations to IML methods, and properties of models and data characteristics; (c) a practical translation for practitioners of the axioms from (b) in terms of what an IML method provides and what not, ideally with implementable guidelines and diagnostic checks for violated assumptions to guarantee correct interpretations. While (a) is nearly always given for any published method, much work remains for (b) and (c).

3 Bad Model Generalization

Pitfall: Under- or overfitting models can result in misleading interpretations with respect to the true feature effects and importance scores, as the model does not match the underlying data-generating process well [39]. Formally, most IML methods are designed to interpret the model instead of drawing inferences about

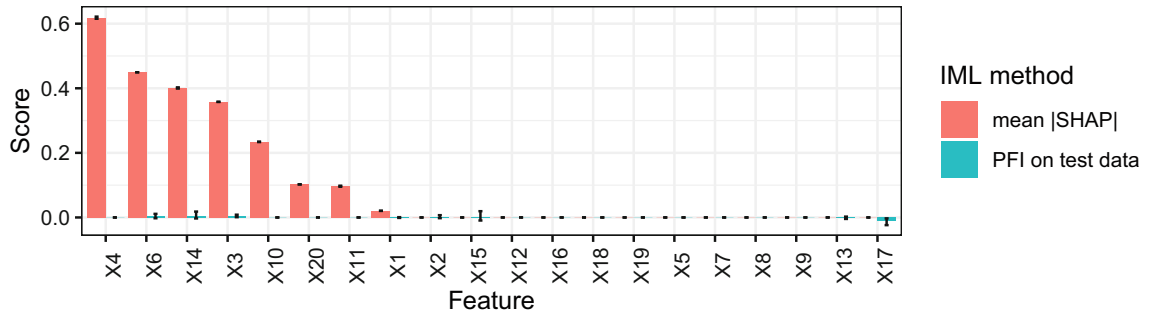


Fig. 2. Assuming one-fits-all interpretability. A default `xgboost` regression model that minimizes the mean squared error (MSE) was fitted on 20 independently and uniformly distributed features to predict another independent, uniformly sampled target. In this setting, predicting the (unconditional) mean $\mathbb{E}[Y]$ in a constant model is optimal. The learner overfits due to a small training data size. Mean marginal SHAP (red, error bars indicate 0.05 and 0.95 quantiles) exposes all mechanistically used features. In contrast, PFI on test data (blue, error bars indicate 0.05 and 0.95 quantiles) considers all features to be irrelevant, since no feature contributes to the generalization performance.

the data-generating process. In practice, however, the latter is often the goal of the analysis, and then an interpretation can only be as good as its underlying model. If a model approximates the data-generating process well enough, its interpretation should reveal insights into the underlying process.

Solution: In-sample evaluation (i.e. on training data) should not be used to assess the performance of ML models due to the risk of overfitting on the training data, which will lead to overly optimistic performance estimates. We must resort to out-of-sample validation based on resampling procedures such as hold-out for larger datasets or cross-validation, or even repeated cross-validation for small sample size scenarios. These resampling procedures are readily available in software [67,89], and well-studied in theory as well as practice [4,11,104], although rigorous analysis of cross-validation is still considered an open problem [103]. Nested resampling is necessary, when computational model selection and hyperparameter tuning are involved [10]. This is important, as the Bayes error for most practical situations is unknown, and we cannot make absolute statements about whether a model already optimally fits the data.

Figure 3 shows the mean squared errors for a simulated example on both training and test data for a support vector machine (SVM), a random forest, and a linear model. Additionally, PDPs for all models are displayed, which show to what extent each model's effect estimates deviate from the ground truth. The linear model is unable to represent the non-linear relationship, which is reflected in a high error on both test and training data and the linear PDPs. In contrast, the random forest has a low training error but a much higher test error, which indicates overfitting. Also, the PDPs for the random forest display overfitting behavior, as the curves are quite noisy, especially at the lower and upper value

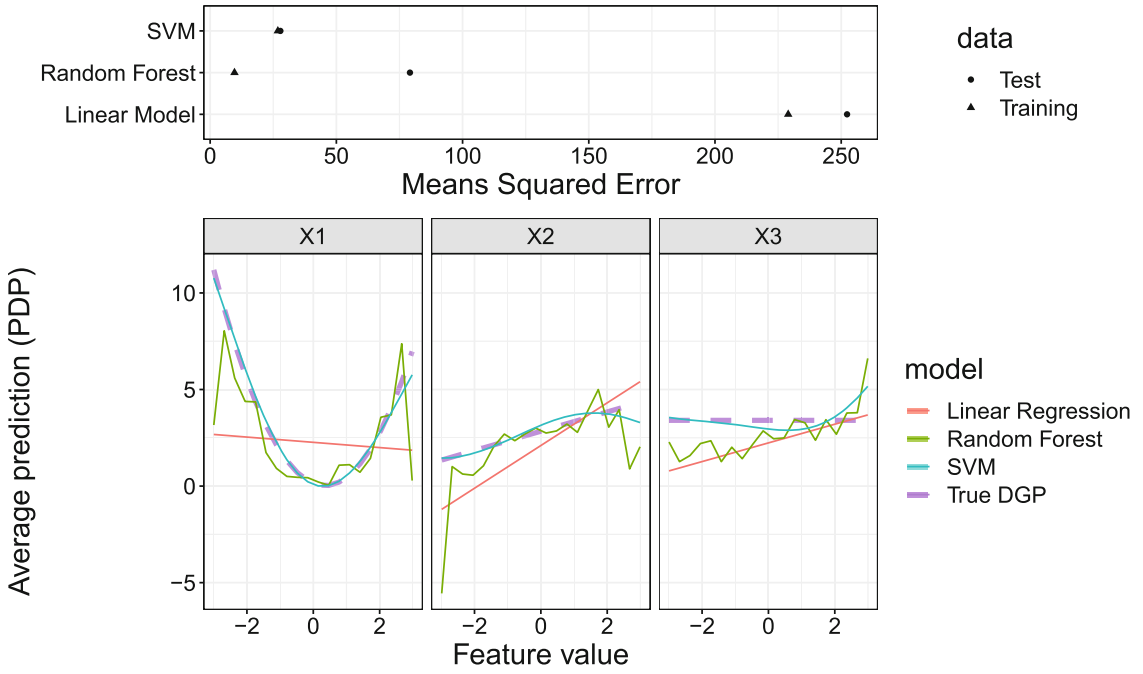


Fig. 3. Bad model generalization. **Top:** Performance estimates on training and test data for a linear regression model (underfitting), a random forest (overfitting) and a support vector machine with radial basis kernel (good fit). The three features are drawn from a uniform distribution, and the target was generated as $Y = X_1^2 + X_2 - 5X_1X_2 + \epsilon$, with $\epsilon \sim N(0, 5)$.**Bottom:** PDPs for the data-generating process (DGP) – which is the ground truth – and for the three models.

ranges of each feature. The SVM with both low training and test error comes closest to the true PDPs.

4 Unnecessary Use of Complex Models

Pitfall: A common mistake is to use an opaque, complex ML model when an interpretable model would have been sufficient, i.e. when the performance of interpretable models is only negligibly worse – or maybe the same or even better – than that of the ML model. Although model-agnostic methods can shed light on the behavior of complex ML models, inherently interpretable models still offer a higher degree of transparency [95] and considering them increases the chance of discovering the true data-generating function [23]. What constitutes an interpretable model is highly dependent on the situation and target audience, as even a linear model might be difficult to interpret when many features and interactions are involved.

It is commonly believed that complex ML models always outperform more interpretable models in terms of accuracy and should thus be preferred. However, there are several examples where interpretable models have proven to be serious competitors: More than 15 years ago, Hand [49] demonstrated that simple models often achieve more than 90% of the predictive power of potentially highly complex models across the UCI benchmark data repository and concluded that such

models often should be preferred due to their inherent interpretability; Makridakis et al. [79] systematically compared various ML models (including long-short-term-memory models and multi-layer neural networks) to statistical models (e.g. damped exponential smoothing and the Theta method) in time series forecasting tasks and found that the latter consistently show greater predictive accuracy; Kuhle et al. [65] found that random forests, gradient boosting and neural networks did not outperform logistic regression in predicting fetal growth abnormalities; Similarly, Wu et al. [120] have shown that a logistic regression model performs as well as AdaBoost and even better than an SVM in predicting heart disease from electronic health record data; Baesens et al. [7] showed that simple interpretable classifiers perform competitively for credit scoring, and in an update to the study the authors note that "the complexity and/or recency of a classifier are misleading indicators of its prediction performance" [71].

Solution: We recommend starting with simple, interpretable models such as linear regression models and decision trees. Generalized additive models (GAM) [50] can serve as a gradual transition between simple linear models and more complex machine learning models. GAMs have the desirable property that they can additively model smooth, non-linear effects and provide PDPs out-of-the-box, but without the potential pitfall of masking interactions (see Sect. 6). The additive model structure of a GAM is specified before fitting the model so that only the pre-specified feature or interaction effects are estimated. Interactions between features can be added manually or algorithmically (e.g. via a forward greedy search) [18]. GAMs can be fitted with component-wise boosting [99]. The boosting approach allows to smoothly increase model complexity, from sparse linear models to more complex GAMs with non-linear effects and interactions. This smooth transition provides insight into the tradeoffs between model simplicity and performance gains. Furthermore, component-wise boosting has an in-built feature selection mechanism as the model is build incrementally, which is especially useful in high-dimensional settings (see Sect. 9.1). The predictive performance of models of different complexity should be carefully measured and compared. Complex models should only be favored if the additional performance gain is both significant and relevant – a judgment call that the practitioner must ultimately make. Starting with simple models is considered best practice in data science, independent of the question of interpretability [23]. The comparison of predictive performance between model classes of different complexity can add further insights for interpretation.

Open Issues: Measures of model complexity allow quantifying the trade-off between complexity and performance and to automatically optimize for multiple objectives beyond performance. Some steps have been made towards quantifying model complexity, such as using functional decomposition and quantifying the complexity of the components [82] or measuring the stability of predictions [92]. However, further research is required, as there is no single perfect definition of interpretability, but rather multiple depending on the context [30,95].

5 Ignoring Feature Dependence

5.1 Interpretation with Extrapolation

Pitfall: When features are dependent, perturbation-based IML methods such as PFI, PDP, LIME, and Shapley values extrapolate in areas where the model was trained with little or no training data, which can cause misleading interpretations [55]. This is especially true if the ML model relies on feature interactions [45] – which is often the case. Perturbations produce artificial data points that are used for model predictions, which in turn are aggregated to produce global or local interpretations [100]. Feature values can be perturbed by replacing original values with values from an equidistant grid of that feature, with permuted or randomly subsampled values [19], or with quantiles. We highlight two major issues: First, if features are dependent, all three perturbation approaches produce unrealistic data points, i.e. the new data points are located outside of the multivariate joint distribution of the data (see Fig. 4). Second, even if features are independent, using an equidistant grid can produce unrealistic values for the feature of interest. Consider a feature that follows a skewed distribution with outliers. An equidistant grid would generate many values between outliers and non-outliers. In contrast to the grid-based approach, the other two approaches maintain the marginal distribution of the feature of interest.

Both issues can result in misleading interpretations (illustrative examples are given in [55,84]), since the model is evaluated in areas of the feature space with few or no observed real data points, where model uncertainty can be expected to be very high. This issue is aggravated if interpretation methods integrate over such points with the same weight and confidence as for much more realistic samples with high model confidence.

Solution: Before applying interpretation methods, practitioners should check for dependencies between features in the data, e.g. via descriptive statistics or measures of dependence (see Sect. 5.2). When it is unavoidable to include dependent features in the model (which is usually the case in ML scenarios), additional information regarding the strength and shape of the dependence structure should be provided. Sometimes, alternative interpretation methods can be used as a workaround or to provide additional information. Accumulated local effect plots (ALE) [3] can be applied when features are dependent, but can produce non-intuitive effect plots for simple linear models with interactions [45]. For other methods such as the PFI, conditional variants exist [17,84,107]. In the case of LIME, it was suggested to focus in sampling on realistic (i.e. close to the data manifold) [97] and relevant areas (e.g. close to the decision boundary) [69]. Note, however, that conditional interpretations are often different and should not be used as a substitute for unconditional interpretations (see Sect. 5.3). Furthermore, dependent features should not be interpreted separately but rather jointly. This can be achieved by visualizing e.g. a 2-dimensional ALE plot of two dependent features, which, admittedly, only works for very low-dimensional combinations. Especially in high-dimensional settings where dependent features

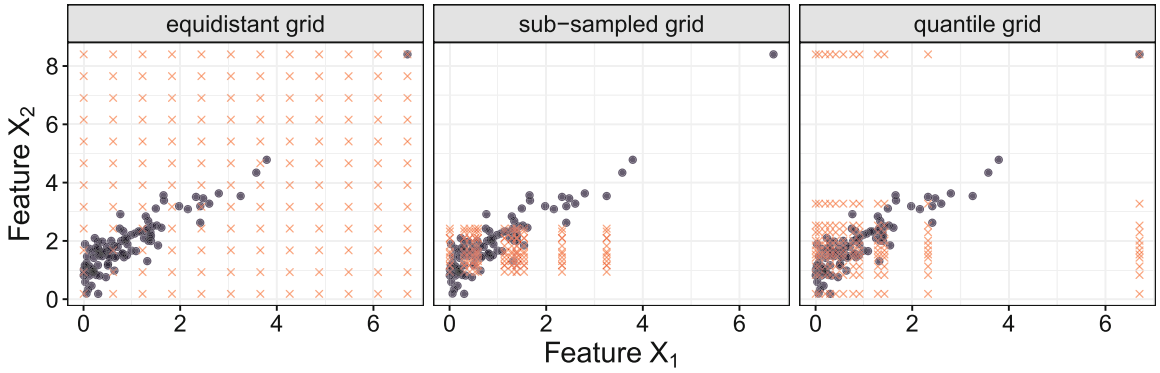


Fig. 4. Interpretation with extrapolation. Illustration of artificial data points generated by three different perturbation approaches. The black dots refer to observed data points and the red crosses to the artificial data points.

can be grouped in a meaningful way, grouped interpretation methods might be more reasonable (see Sect. 9.1).

We recommend using quantiles or randomly subsampled values over equidistant grids. By default, many implementations of interpretability methods use an equidistant grid to perturb feature values [41,81,89], although some also allow using user-defined values.

Open Issues: A comprehensive comparison of strategies addressing extrapolation and how they affect an interpretation method is currently missing. This also includes studying interpretation methods and their conditional variants when they are applied to data with different dependence structures.

5.2 Confusing Linear Correlation with General Dependence

Pitfall: Features with a Pearson correlation coefficient (PCC) close to zero can still be dependent and cause misleading model interpretations (see Fig. 5). While independence between two features implies that the PCC is zero, the converse is generally false. The PCC, which is often used to analyze dependence, only tracks linear correlations and has other shortcomings such as sensitivity to outliers [113]. Any type of dependence between features can have a strong impact on the interpretation of the results of IML methods (see Sect. 5.1). Thus, knowledge about the (possibly non-linear) dependencies between features is crucial for an informed use of IML methods.

Solution: Low-dimensional data can be visualized to detect dependence (e.g. scatter plots) [80]. For high-dimensional data, several other measures of dependence in addition to PCC can be used. If dependence is monotonic, Spearman's rank correlation coefficient [72] can be a simple, robust alternative to PCC. For categorical or mixed features, separate dependence measures have been proposed, such as Kendall's rank correlation coefficient for ordinal features, or the phi coefficient and Goodman & Kruskal's lambda for nominal features [59].

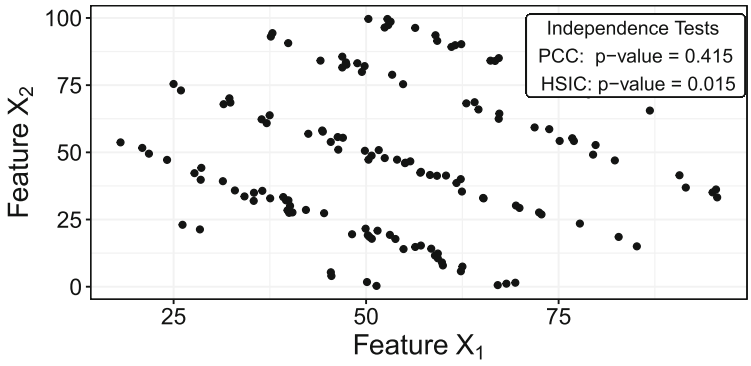


Fig. 5. Confusing linear correlation with dependence. Highly dependent features X_1 and X_2 that have a correlation close to zero. A test (H_0: Features are independent) using Pearson correlation is not significant, but for HSIC, the H_0-hypothesis gets rejected. Data from [80].

Studying non-linear dependencies is more difficult since a vast variety of possible associations have to be checked. Nevertheless, several non-linear association measures with sound statistical properties exist. Kernel-based measures, such as kernel canonical correlation analysis (KCCA) [6] or the Hilbert-Schmidt independence criterion (HSIC) [44], are commonly used. They have a solid theoretical foundation, are computationally feasible, and robust [113]. In addition, there are information-theoretical measures, such as (conditional) mutual information [24] or the maximal information coefficient (MIC) [93], that can however be difficult to estimate [9,116]. Other important measures are e.g. the distance correlation [111], the randomized dependence coefficient (RDC) [74], or the alternating conditional expectations (ACE) algorithm [14]. In addition to using PCC, we recommend using at least one measure that detects non-linear dependencies (e.g. HSIC).

5.3 Misunderstanding Conditional Interpretation

Pitfall: Conditional variants of interpretation techniques avoid extrapolation but require a different interpretation. Interpretation methods that perturb features independently of others will extrapolate under dependent features but provide insight into the model's mechanism [56,61]. Therefore, these methods are said to be true to the model but not true to the data [21].

For feature effect methods such as the PDP, the plot can be interpreted as the isolated, average effect the feature has on the prediction. For the PFI, the importance can be interpreted as the drop in performance when the feature's information is "destroyed" (by perturbing it). Marginal SHAP value functions [78] quantify a feature's contribution to a specific prediction, and marginal SAGE value functions [25] quantify a feature's contribution to the overall prediction performance. All the aforementioned methods extrapolate under dependent features (see also Sect. 5.1), but satisfy sensitivity, i.e. are zero if a feature is not used by the model [25,56,61,110].

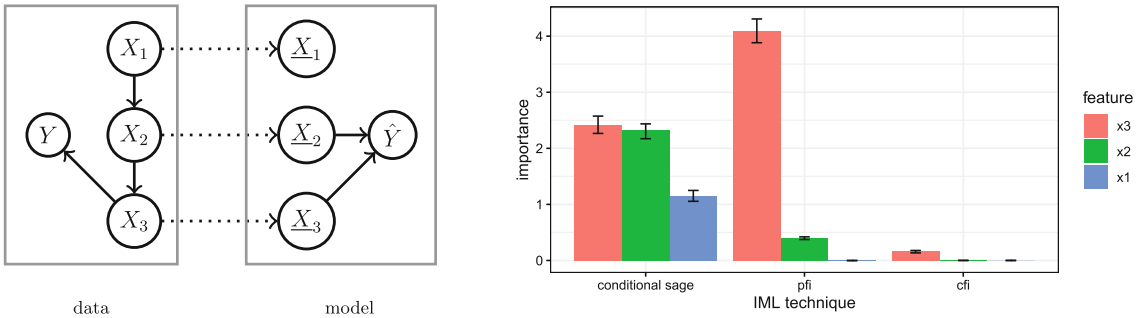


Fig. 6. Misunderstanding conditional interpretation. A linear model was fitted on the data-generating process modeled using a linear Gaussian structural causal model. The entailed directed acyclic graph is depicted on the left. For illustrative purposes, the original model coefficients were updated such that not only feature X_3, but also feature X_2 is used by the model. PFI on test data considers both X_3 and X_2 to be relevant. In contrast, conditional feature importance variants either only consider X_3 to be relevant (CFI) or consider all features to be relevant (conditional SAGE value function).

Conditional variants of these interpretation methods do not replace feature values independently of other features, but in such a way that they conform to the conditional distribution. This changes the interpretation as the effects of all dependent features become entangled. Depending on the method, conditional sampling leads to a more or less restrictive notion of relevance.

For example, for dependent features, the Conditional Feature Importance (CFI) [17,84,107,117] answers the question: "How much does the model performance drop if we permute a feature, *but given that we know the values of the other features?*" [63,84,107].[1] Two highly dependent features might be individually important (based on the unconditional PFI), but have a very low conditional importance score because the information of one feature is contained in the other and vice versa.

In contrast, the conditional variant of PDP, called marginal plot or M-plot [3], violates sensitivity, i.e. may even show an effect for features that are not used by the model. This is because for M-plots, the feature of interest is not sampled conditionally on the remaining features, but rather the remaining features are sampled conditionally on the feature of interest. As a consequence, the distribution of dependent covariates varies with the value of the feature of interest. Similarly, conditional SAGE and conditional SHAP value functions sample the remaining features conditional on the feature of interest and therefore violate sensitivity [25,56,61,109].

We demonstrate the difference between PFI, CFI, and conditional SAGE value functions on a simulated example (Fig. 6) where the data-generating mech-

[1] While for CFI the conditional independence of the feature of interest X_j with the target Y given the remaining features X_{-j} ($Y \perp X_j | X_{-j}$) is already a sufficient condition for zero importance, the corresponding PFI may still be nonzero [63].

anism is known. While PFI only considers features to be relevant if they are actually used by the model, SAGE value functions may also consider a feature to be important that is not directly used by the model if it contains information that the model exploits. CFI only considers a feature to be relevant if it is both mechanistically used by the model and contributes unique information about Y.

Solution: When features are highly dependent and conditional effects and importance scores are used, the practitioner must be aware of the distinct interpretation. Recent work formalizes the implications of marginal and conditional interpretation techniques [21,25,56,61,63]. While marginal methods provide insight into the model's mechanism but are not true to the data, their conditional variants are not true to the model but provide insight into the associations in the data.

If joint insight into model and data is required, designated methods must be used. ALE plots [3] provide interval-wise unconditional interpretations that are true to the data. They have been criticized to produce non-intuitive results for certain data-generating mechanisms [45]. Molnar et al. [84] propose a subgroup-based conditional sampling technique that allows for group-wise marginal interpretations that are true to model and data and that can be applied to feature importance and feature effects methods such as conditional PDPs and CFI. For feature importance, the DEDACT framework [61] allows to decompose conditional importance measures such as SAGE value functions into their marginal contributions and vice versa, thereby allowing global insight into both: the sources of prediction-relevant information in the data as well as into the feature pathways by which the information enters the model.

Open Issues: The quality of conditional IML techniques depends on the goodness of the conditional sampler. Especially in continuous, high-dimensional settings, conditional sampling is challenging. More research on the robustness of interpretation techniques regarding the quality of the sample is required.

6 Misleading Interpretations Due to Feature Interactions

6.1 Misleading Feature Effects Due to Aggregation

Pitfall: Global interpretation methods, such as PDP or ALE plots, visualize the average effect of a feature on a model's prediction. However, they can produce misleading interpretations when features interact. Figure 7 A and B show the marginal effect of features X_1 and X_2 of the below-stated simulation example. While the PDP of the non-interacting feature X_1 seems to capture the true underlying effect of X_1 on the target quite well (A), the global aggregated effect of the interacting feature X_2 (B) shows almost no influence on the target, although an effect is clearly there by construction.

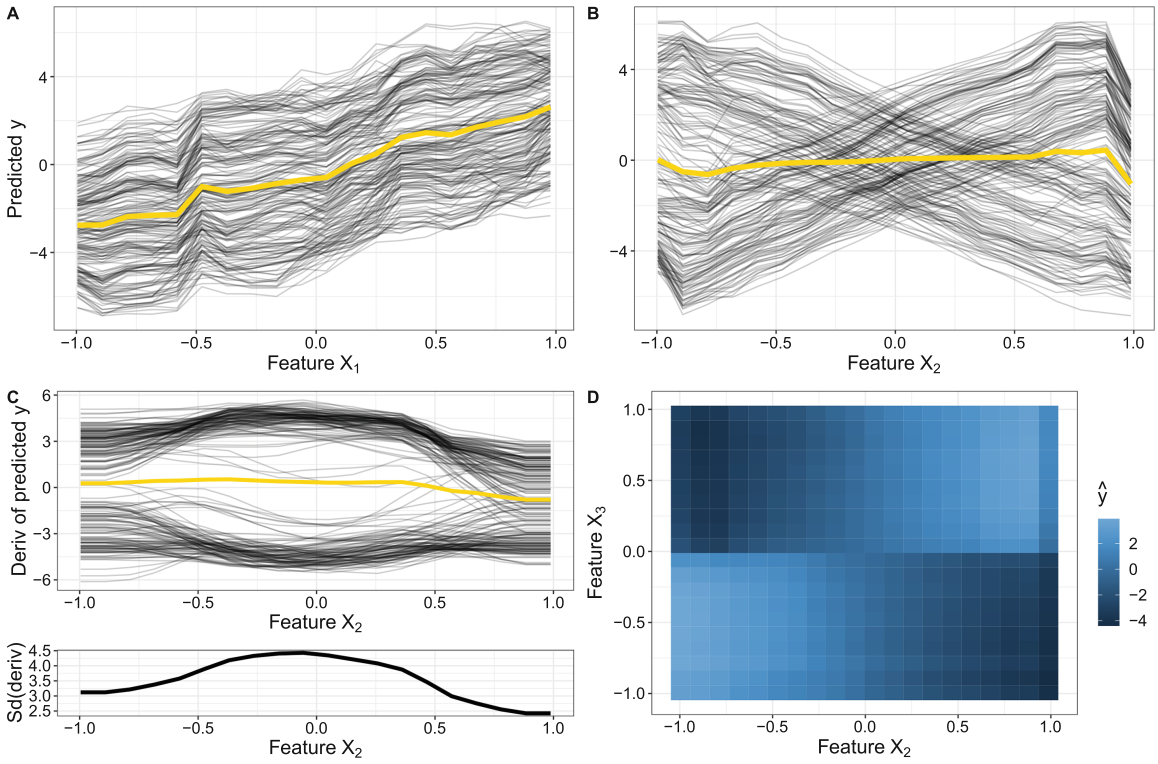


Fig. 7. Misleading effect due to interactions. Simulation example with interactions: $Y = 3X_1 - 6X_2 + 12X_2\mathbb{1}_{(X_3 \geq 0)} + \epsilon$ with $X_1, X_2, X_3 \overset{i.i.d.}{\sim} U[-1, 1]$ and $\epsilon \overset{i.i.d.}{\sim} N(0, 0.3)$. A random forest with 500 trees is fitted on 1000 observations. Effects are calculated on 200 randomly sampled (training) observations. **A, B:** PDP (yellow) and ICE curves of X_1 and X_2; **C:** Derivative ICE curves and their standard deviation of X_2; **D:** 2-dimensional PDP of X_2 and X_3.

Solution: For the PDP, we recommend to additionally consider the corresponding ICE curves [38]. While PDP and ALE average out interaction effects, ICE curves directly show the heterogeneity between individual predictions. Figure 7 A illustrates that the individual marginal effect curves all follow an upward trend with only small variations. Hence, by aggregating these ICE curves to a global marginal effect curve such as the PDP, we do not lose much information. However, when the regarded feature interacts with other features, such as feature X_2 with feature X_3 in this example, then marginal effect curves of different observations might not show similar effects on the target. Hence, ICE curves become very heterogeneous, as shown in Fig. 7 B. In this case, the influence of feature X_2 is not well represented by the global average marginal effect. Particularly for continuous interactions where ICE curves start at different intercepts, we recommend the use of derivative or centered ICE curves, which eliminate differences in intercepts and leave only differences due to interactions [38]. Derivative ICE curves also point out the regions of highest interaction with other features. For example, Fig. 7 C indicates that predictions for X_2 taking values close to 0 strongly depend on other features' values. While these methods show that interactions are present with regards to the feature of interest but do not reveal other

features with which it interacts, the 2-dimensional PDP or ALE plot are options to visualize 2-way interaction effects. The 2-dimensional PDP in Fig. 7 D shows that predictions with regards to feature X_2 highly depend on the feature values of feature X_3.

Other methods that aim to gain more insights into these visualizations are based on clustering homogeneous ICE curves, such as visual interaction effects (VINE) [16] or [122]. As an example, in Fig. 7 B, it would be more meaningful to average over the upward and downward proceeding ICE curves separately and hence show that the average influence of feature X_2 on the target depends on an interacting feature (here: X_3). Work by Zon et al. [125] followed a similar idea by proposing an interactive visualization tool to group Shapley values with regards to interacting features that need to be defined by the user.

Open Issues: The introduced visualization methods are not able to illustrate the type of the underlying interaction and most of them are also not applicable to higher-order interactions.

6.2 Failing to Separate Main from Interaction Effects

Pitfall: Many interpretation methods that quantify a feature's importance or effect cannot separate an interaction from main effects. The PFI, for example, includes both the importance of a feature and the importance of all its interactions with other features [19]. Also local explanation methods such as LIME and Shapley values only provide additive explanations without separation of main effects and interactions [40].

Solution: Functional ANOVA introduced by [53] is probably the most popular approach to decompose the joint distribution into main and interaction effects. Using the same idea, the H-Statistic [35] quantifies the interaction strength between two features or between one feature and all others by decomposing the 2-dimensional PDP into its univariate components. The H-Statistic is based on the fact that, in the case of non-interacting features, the 2-dimensional partial dependence function equals the sum of the two underlying univariate partial dependence functions. Another similar interaction score based on partial dependencies is defined by [42]. Instead of decomposing the partial dependence function, [87] uses the predictive performance to measure interaction strength. Based on Shapley values, Lundberg et al. [77] proposed SHAP interaction values, and Casalicchio et al. [19] proposed a fair attribution of the importance of interactions to the individual features.

Furthermore, Hooker [54] considers dependent features and decomposes the predictions in main and interaction effects. A way to identify higher-order interactions is shown in [53].

Open Issues: Most methods that quantify interactions are not able to identify higher-order interactions and interactions of dependent features. Furthermore,

the presented solutions usually lack automatic detection and ranking of all interactions of a model. Identifying a suitable shape or form of the modeled interaction is not straightforward as interactions can be very different and complex, e.g., they can be a simple product of features (multiplicative interaction) or can have a complex joint non-linear effect such as smooth spline surface.

7 Ignoring Model and Approximation Uncertainty

Pitfall: Many interpretation methods only provide a mean estimate but do not quantify uncertainty. Both the model training and the computation of interpretation are subject to uncertainty. The model is trained on (random) data, and therefore should be regarded as a random variable. Similarly, LIME's surrogate model relies on perturbed and reweighted samples of the data to approximate the prediction function locally [94]. Other interpretation methods are often defined in terms of expectations over the data (PFI, PDP, Shapley values, ...), but are approximated using Monte Carlo integration. Ignoring uncertainty can result in the interpretation of noise and non-robust results. The true effect of a feature may be flat, but – purely by chance, especially on smaller datasets – the Shapley value might show an effect. This effect could cancel out once averaged over multiple model fits.

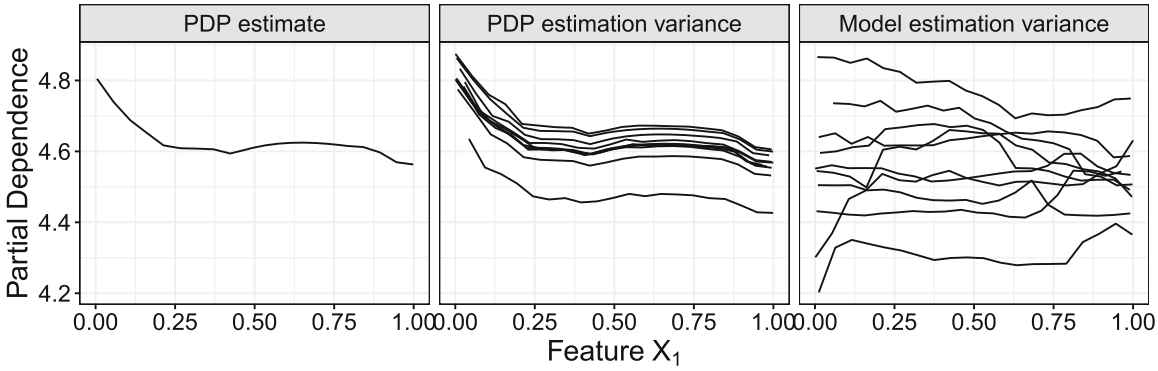


Fig. 8. Ignoring model and approximation uncertainty. PDP for X_1 with $Y = 0 \cdot X_1 + \sum_{j=2}^{10} X_j + \epsilon_i$ with $X_1, \ldots, X_{10} \sim U[0,1]$ and $\epsilon_i \sim N(0, 0.9)$. **Left:** PDP for X_1 of a random forest trained on 100 data points. **Middle:** Multiple PDPs (10x) for the model from left plots, but with different samples (each n=100) for PDP estimation. **Right:** Repeated (10x) data samples of n=100 and newly fitted random forest.

Figure 8 shows that a single PDP (first plot) can be misleading because it does not show the variance due to PDP estimation (second plot) and model fitting (third plot). If we are not interested in learning about a specific model, but rather about the relationship between feature X_1 and the target (in this case), we should consider the model variance.

Solution: By repeatedly computing PDP and PFI with a given model, but with different permutations or bootstrap samples, the uncertainty of the estimate can be quantified, for example in the form of confidence intervals. For PFI, frameworks for confidence intervals and hypothesis tests exist [2,117], but they assume a fixed model. If the practitioner wants to condition the analysis on the modeling process and capture the process' variance instead of conditioning on a fixed model, PDP and PFI should be computed on multiple model fits [83].

Open Issues: While Moosbauer et al. [85] derived confidence bands for PDPs for probabilistic ML models that cover the model's uncertainty, a general model-agnostic uncertainty measure for feature effect methods such as ALE [3] and PDP [36] has (to the best of our knowledge) not been introduced yet.

8 Ignoring the Rashomon Effect

Pitfall: Sometimes different models explain the data-generating process equally well, but contradict each other. This phenomenon is called the Rashomon effect, named after the movie "Rashomon" from the year 1950. Breiman formalized it for predictive models in 2001 [13]: Different prediction models might perform equally well (Rashomon set), but construct the prediction function in a different way (e.g. relying on different features). This can result in conflicting interpretations and conclusions about the data. Even small differences in the training data can cause one model to be preferred over another.

For example, Dong and Rudin [29] identified a Rashomon set of equally well performing models for the COMPAS dataset. They showed that the models differed greatly in the importance they put on certain features. Specifically, if criminal history was identified as less important, race was more important and vice versa. Cherry-picking one model and its underlying explanation might not be sufficient to draw conclusions about the data-generating process. As Hancox-Li [48] states "just because race happens to be an unimportant variable in that one explanation does not mean that it is objectively an unimportant variable".

The Rashomon effect can also occur at the level of the interpretation method itself. Differing hyperparameters or interpretation goals can be one reason (see Sect. 2). But even if the hyperparameters are fixed, we could still obtain contradicting explanations by an interpretation method, e.g., due to a different data sample or initial seed.

A concrete example of the Rashomon effect is counterfactual explanations. Different counterfactuals may all alter the prediction in the desired way, but point to different feature changes required for that change. If a person is deemed uncreditworthy, one corresponding counterfactual explaining this decision may point to a scenario in which the person had asked for a shorter loan duration and amount, while another counterfactual may point to a scenario in which the person had a higher income and more stable job. Focusing on only one counterfactual explanation in such cases strongly limits the possible epistemic access.

Solution: If multiple, equally good models exist, their interpretations should be compared. Variable importance clouds [29] is a method for exploring variable importance scores for equally good models within one model class. If the interpretations are in conflict, conclusions must be drawn carefully. Domain experts or further constraints (e.g. fairness or sparsity) could help to pick a suitable model. Semenova et al. [102] also hypothesized that a large Rashomon set could contain simpler or more interpretable models, which should be preferred according to Sect. 4.

In the case of counterfactual explanations, multiple, equally good explanations exist. Here, methods that return a set of explanations rather than a single one should be used – for example, the method by Dandl et al. [26] or Mothilal et al. [86].

Open Issues: Numerous very different counterfactual explanations are overwhelming for users. Methods for aggregating or combining explanations are still a matter of future research.

9 Failure to Scale to High-Dimensional Settings

9.1 Human-Intelligibility of High-Dimensional IML Output

Pitfall: Applying IML methods naively to high-dimensional datasets (e.g. visualizing feature effects or computing importance scores on feature level) leads to an overwhelming and high-dimensional IML output, which impedes human analysis. Especially interpretation methods that are based on visualizations make it difficult for practitioners in high-dimensional settings to focus on the most important insights.

Solution: A natural approach is to reduce the dimensionality before applying any IML methods. Whether this facilitates understanding or not depends on the possible semantic interpretability of the resulting, reduced feature space – as features can either be selected or dimensionality can be reduced by linear or non-linear transformations. Assuming that users would like to interpret in the original feature space, many feature selection techniques can be used [46], resulting in much sparser and consequently easier to interpret models. Wrapper selection approaches are model-agnostic and algorithms like greedy forward selection or subset selection procedures [5,60], which start from an empty model and iteratively add relevant (subsets of) features if needed, even allow to measure the relevance of features for predictive performance. An alternative is to directly use models that implicitly perform feature selection such as LASSO [112] or component-wise boosting [99] as they can produce sparse models with fewer features. In the case of LIME or other interpretation methods based on surrogate models, the aforementioned techniques could be applied to the surrogate model.

When features can be meaningfully grouped in a data-driven or knowledge-driven way [51], applying IML methods directly to grouped features instead of

single features is usually more time-efficient to compute and often leads to more appropriate interpretations. Examples where features can naturally be grouped include the grouping of sensor data [20], time-lagged features [75], or one-hot-encoded categorical features and interaction terms [43]. Before a model is fitted, groupings could already be exploited for dimensionality reduction, for example by selecting groups of features by the group LASSO [121].

For model interpretation, various papers extended feature importance methods from single features to groups of features [5,43,114,119]. In the case of grouped PFI, this means that we perturb the entire group of features at once and measure the performance drop compared to the unperturbed dataset. Compared to standard PFI, the grouped PFI does not break the association to the other features of the group, but to features of other groups and the target. This is especially useful when features within the same group are highly correlated (e.g. time-lagged features), but between-group dependencies are rather low. Hence, this might also be a possible solution for the extrapolation pitfall described in Sect. 5.1.

We consider the PhoneStudy in [106] as an illustration. The PhoneStudy dataset contains 1821 features to analyze the link between human behavior based on smartphone data and participants' personalities. Interpreting the results in this use case seems to be challenging since features were dependent and single feature effects were either small or non-linear [106]. The features have been grouped in behavior-specific categories such as app-usage, music consumption, or overall phone usage. Au et al. [5] calculated various grouped importance scores on the feature groups to measure their influence on a specific personality trait (e.g. conscientiousness). Furthermore, the authors applied a greedy forward subset selection procedure via repeated subsampling on the feature groups and showed that combining app-usage features and overall phone usage features were most of the times sufficient for the given prediction task.

Open Issues: The quality of a grouping-based interpretation strongly depends on the human intelligibility and meaningfulness of the grouping. If the grouping structure is not naturally given, then data-driven methods can be used. However, if feature groups are not meaningful (e.g. if they cannot be described by a super-feature such as app-usage), then subsequent interpretations of these groups are purposeless. One solution could be to combine feature selection strategies with interpretation methods. For example, LIME's surrogate model could be a LASSO model. However, beyond surrogate models, the integration of feature selection strategies remains an open issue that requires further research.

Existing research on grouped interpretation methods mainly focused on quantifying grouped feature importance, but the question of "how a group of features influences a model's prediction" remains almost unanswered. Only recently, [5,15,101] attempted to answer this question by using dimension-reduction techniques (such as PCA) before applying the interpretation method. However, this is also a matter of further research.

9.2 Computational Effort

Pitfall: Some interpretation methods do not scale linearly with the number of features. For example, for the computation of exact Shapley values the number of possible coalitions [25,78], or for a (full) functional ANOVA decomposition the number of components (main effects plus all interactions) scales with $\mathcal{O}(2^p)$ [54].[2]

Solution: For the functional ANOVA, a common solution is to keep the analysis to the main effects and selected 2-way interactions (similar for PDP and ALE). Interesting 2-way interactions can be selected by another method such as the H-statistic [35]. However, the selection of 2-way interactions requires additional computational effort. Interaction strength usually decreases quickly with increasing interaction size, and one should only consider d-way interactions when all their $(d-1)$-way interactions were significant [53]. For Shapley-based methods, an efficient approximation exists that is based on randomly sampling and evaluating feature orderings until the estimates converge. The variance of the estimates reduces in $\mathcal{O}(\frac{1}{m})$, where m is the number of evaluated orderings [25,78].

9.3 Ignoring Multiple Comparison Problem

Pitfall: Simultaneously testing the importance of multiple features will result in false-positive interpretations if the multiple comparisons problem (MCP) is ignored. The MCP is well known in significance tests for linear models and exists similarly in testing for feature importance in ML. For example, suppose we simultaneously test the importance of 50 features (with the H_0-hypothesis of zero importance) at the significance level $\alpha = 0.05$. Even if all features are unimportant, the probability of observing that at least one feature is significantly important is $1 - \mathbb{P}(\text{'no feature important'}) = 1 - (1 - 0.05)^{50} \approx 0.923$. Multiple comparisons become even more problematic the higher the dimension of the dataset.

Solution: Methods such as Model-X knockoffs [17] directly control for the false discovery rate (FDR). For all other methods that provide p-values or confidence intervals, such as PIMP (Permutation IMPortance) [2], which is a testing approach for PFI, MCP is often ignored in practice to the best of our knowledge, with some exceptions[105,117]. One of the most popular MCP adjustment methods is the Bonferroni correction [31], which rejects a null hypothesis if its p-value is smaller than α/p, with p as the number of tests. It has the disadvantage that it increases the probability of false negatives [90]. Since MCP is well known in statistics, we refer the practitioner to [28] for an overview and discussion of alternative adjustment methods, such as the Bonferroni-Holm method [52].

[2] Similar to the PDP or ALE plots, the functional ANOVA components describe individual feature effects and interactions.

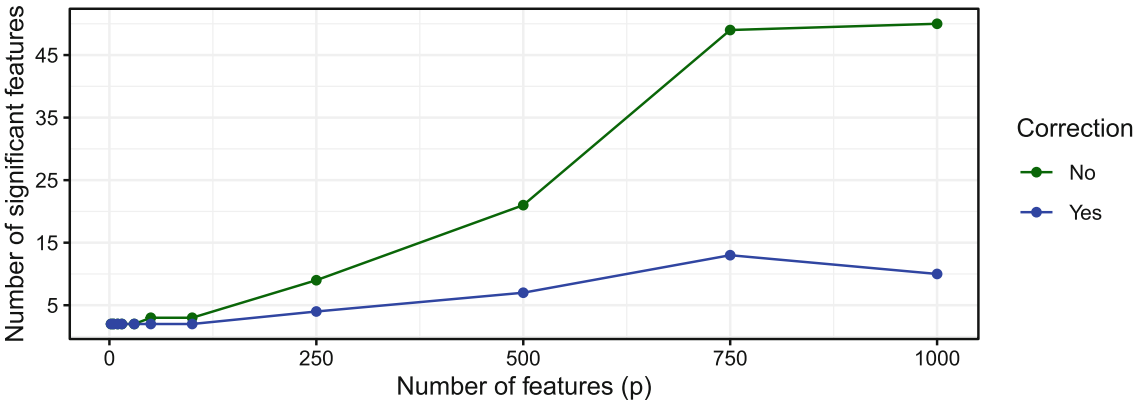


Fig. 9. Failure to scale to high-dimensional settings. Comparison of the number of features with significant importance - once with and once without Bonferroni-corrected significance levels for a varying number of added noise variables. Datasets were sampled from $Y = 2X_1 + 2X_2^2 + \epsilon$ with $X_1, X_2, \epsilon \sim N(0, 1)$. $X_3, X_4, ..., X_p \sim N(0, 1)$ are additional noise variables with p ranging between 2 and 1000. For each p, we sampled two datasets from this data-generating process – one to train a random forest with 500 trees on and one to test whether feature importances differed from 0 using PIMP. In all experiments, X_1 and X_2 were correctly identified as important.

As an example, in Fig. 9 we compare the number of features with significant importance measured by PIMP once with and once without Bonferroni-adjusted significance levels ($\alpha = 0.05$ vs. $\alpha = 0.05/p$). Without correcting for multi-comparisons, the number of features mistakenly evaluated as important grows considerably with increasing dimension, whereas Bonferroni correction results in only a modest increase.

10 Unjustified Causal Interpretation

Pitfall: Practitioners are often interested in causal insights into the underlying data-generating mechanisms, which IML methods do not generally provide. Common causal questions include the identification of causes and effects, predicting the effects of interventions, and answering counterfactual questions [88]. For example, a medical researcher might want to identify risk factors or predict average and individual treatment effects [66]. In search of answers, a researcher can therefore be tempted to interpret the result of IML methods from a causal perspective.

However, a causal interpretation of predictive models is often not possible. Standard supervised ML models are not designed to model causal relationships but to merely exploit associations. A model may therefore rely on causes and effects of the target variable as well as on variables that help to reconstruct unobserved influences on Y, e.g. causes of effects [118]. Consequently, the question of whether a variable is relevant to a predictive model (indicated e.g. by PFI > 0) does not directly indicate whether a variable is a cause, an effect, or does not stand in any causal relation to the target variable. Furthermore,

even if a model would rely solely on direct causes for the prediction, the causal structure between features must be taken into account. Intervening on a variable in the real world may affect not only Y but also other variables in the feature set. Without assumptions about the underlying causal structure, IML methods cannot account for these adaptions and guide action [58,62].

As an example, we constructed a dataset by sampling from a structural causal model (SCM), for which the corresponding causal graph is depicted in Fig. 10. All relationships are linear Gaussian with variance 1 and coefficients 1. For a linear model fitted on the dataset, all features were considered to be relevant based on the model coefficients ($\hat{y} = 0.329x_1 + 0.323x_2 - 0.327x_3 + 0.342x_4 + 0.334x_5$, $R^2 = 0.943$), although x_3, x_4 and x_5 do not cause Y.

Solution: The practitioner must carefully assess whether sufficient assumptions can be made about the underlying data-generating process, the learned model, and the interpretation technique. If these assumptions are met, a causal interpretation may be possible. The PDP between a feature and the target can be interpreted as the respective average causal effect if the model performs well and the set of remaining variables is a valid adjustment set [123]. When it is known whether a model is deployed in a causal or anti-causal setting – i.e. whether the model attempts to predict an effect from its causes or the other way round – a partial identification of the causal roles based on feature relevance is possible (under strong and non-testable assumptions) [118]. Designated tools and approaches are available for causal discovery and inference [91].

Open Issues: The challenge of causal discovery and inference remains an open key issue in the field of ML. Careful research is required to make explicit under which assumptions what insight about the underlying data-generating mechanism can be gained by interpreting an ML model.

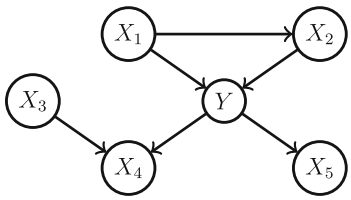


Fig. 10. Causal graph

11 Discussion

In this paper, we have reviewed numerous pitfalls of local and global model-agnostic interpretation techniques, e.g. in the case of bad model generalization, dependent features, interactions between features, or causal interpretations. We have not attempted to provide an exhaustive list of all potential pitfalls in ML

model interpretation, but have instead focused on common pitfalls that apply to various model-agnostic IML methods and pose a particularly high risk.

We have omitted pitfalls that are more specific to one IML method type: For local methods, the vague notions of neighborhood and distance can lead to misinterpretations [68,69], and common distance metrics (such as the Euclidean distance) are prone to the curse of dimensionality [1]; Surrogate methods such as LIME may not be entirely faithful to the original model they replace in interpretation. Moreover, we have not addressed pitfalls associated with certain data types (like the definition of superpixels in image data [98]), nor those related to human cognitive biases (e.g. the illusion of model understanding [22]).

Many pitfalls in the paper are strongly linked with axioms that encode desiderata of model interpretation. For example, pitfall Sect. 5.3 (misunderstanding conditional interpretations) is related to violations of sensitivity [56,110]. As such, axioms can help to make the strengths and limitations of methods explicit. Therefore, we encourage an axiomatic evaluation of interpretation methods.

We hope to promote a more cautious approach when interpreting ML models in practice, to point practitioners to already (partially) available solutions, and to stimulate further research on these issues. The stakes are high: ML algorithms are increasingly used for socially relevant decisions, and model interpretations play an important role in every empirical science. Therefore, we believe that users can benefit from concrete guidance on properties, dangers, and problems of IML techniques – especially as the field is advancing at high speed. We need to strive towards a recommended, well-understood set of tools, which will in turn require much more careful research. This especially concerns the meta-issues of comparisons of IML techniques, IML diagnostic tools to warn against misleading interpretations, and tools for analyzing multiple dependent or interacting features.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, pp. 420–434. Springer, Heidelberg (2001). https://doi.org/10.1007/3-540-44503-X_27
2. Altmann, A., Toloşi, L., Sander, O., Lengauer, T.: Permutation importance: a corrected feature importance measure. Bioinformatics **26**(10), 1340–1347 (2010). https://doi.org/10.1093/bioinformatics/btq134
3. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **82**(4), 1059–1086 (2020). https://doi.org/10.1111/rssb.12377
4. Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Statist. Surv. **4**, 40–79 (2010). https://doi.org/10.1214/09-SS054
5. Au, Q., Herbinger, J., Stachl, C., Bischl, B., Casalicchio, G.: Grouped feature importance and combined features effect plot. arXiv preprint arXiv:2104.11688 (2021)
6. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. J. Mach. Learn. Res. **3**(Jul), 1–48 (2002)

7. Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J.: Benchmarking state-of-the-art classification algorithms for credit scoring. J. Oper. Res. Soc. **54**(6), 627–635 (2003). https://doi.org/10.1057/palgrave.jors.2601545
8. Bansal, N., Agarwal, C., Nguyen, A.: SAM: the sensitivity of attribution methods to hyperparameters. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8673–8683 (2020)
9. Belghazi, M.I., et al.: Mutual information neural estimation. In: International Conference on Machine Learning, pp. 531–540 (2018)
10. Bischl, B., et al.: Hyperparameter optimization: foundations, algorithms, best practices and open challenges. arXiv preprint arXiv:2107.05847 (2021)
11. Bischl, B., Mersmann, O., Trautmann, H., Weihs, C.: Resampling methods for meta-model validation with recommendations for evolutionary computation. Evol. Comput. **20**(2), 249–275 (2012). https://doi.org/10.1162/EVCO_a_00069
12. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001). https://doi.org/10.1023/A:1010933404324
13. Breiman, L.: Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. **16**(3), 199–231 (2001). https://doi.org/10.1214/ss/1009213726
14. Breiman, L., Friedman, J.H.: Estimating optimal transformations for multiple regression and correlation. J. Am. Stat. Assoc. **80**(391), 580–598 (1985). https://doi.org/10.1080/01621459.1985.10478157
15. Brenning, A.: Transforming feature space to interpret machine learning models. arXiv:2104.04295 (2021)
16. Britton, M.: Vine: visualizing statistical interactions in black box models. arXiv preprint arXiv:1904.00561 (2019)
17. Candes, E., Fan, Y., Janson, L., Lv, J.: Panning for gold:'model-x'knockoffs for high dimensional controlled variable selection. J. R. Stat. Soc.: Ser. B (Stat. Methodol.) **80**(3), 551–577 (2018). https://doi.org/10.1111/rssb.12265
18. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730 (2015). https://doi.org/10.1145/2783258.2788613
19. Casalicchio, G., Molnar, C., Bischl, B.: Visualizing the feature importance for black box models. In: Berlingerio, M., Bonchi, F., Gärtner, T., Hurley, N., Ifrim, G. (eds.) ECML PKDD 2018. LNCS (LNAI), vol. 11051, pp. 655–670. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-10925-7_40
20. Chakraborty, D., Pal, N.R.: Selecting useful groups of features in a connectionist framework. IEEE Trans. Neural Netw. **19**(3), 381–396 (2008). https://doi.org/10.1109/TNN.2007.910730
21. Chen, H., Janizek, J.D., Lundberg, S., Lee, S.I.: True to the model or true to the data? arXiv preprint arXiv:2006.16234 (2020)
22. Chromik, M., Eiband, M., Buchner, F., Krüger, A., Butz, A.: I think I get your point, AI! the illusion of explanatory depth in explainable AI. In: 26th International Conference on Intelligent User Interfaces, IUI 2021, pp. 307–317. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3397481.3450644
23. Claeskens, G., Hjort, N.L., et al.: Model Selection and Model Averaging. Cambridge Books (2008). https://doi.org/10.1017/CBO9780511790485

24. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley (2012). https://doi.org/10.1002/047174882X
25. Covert, I., Lundberg, S.M., Lee, S.I.: Understanding global feature contributions with additive importance measures. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 17212–17223. Curran Associates, Inc. (2020)
26. Dandl, S., Molnar, C., Binder, M., Bischl, B.: Multi-objective counterfactual explanations. In: Bäck, T., et al. (eds.) PPSN 2020. LNCS, vol. 12269, pp. 448–469. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58112-1_31
27. Das, A., Rad, P.: Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv preprint arXiv:2006.11371 (2020)
28. Dickhaus, T.: Simultaneous Statistical Inference. Springer, Heidelberg (2014). https://doi.org/10.1007/978-3-642-45182-9
29. Dong, J., Rudin, C.: Exploring the cloud of variable importance for the set of all good models. Nat. Mach. Intell. **2**(12), 810–824 (2020). https://doi.org/10.1038/s42256-020-00264-0
30. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
31. Dunn, O.J.: Multiple comparisons among means. J. Am. Stat. Assoc. **56**(293), 52–64 (1961). https://doi.org/10.1080/01621459.1961.10482090
32. Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D.: Do we need hundreds of classifiers to solve real world classification problems. J. Mach. Learn. Res. **15**(1), 3133–3181 (2014). https://doi.org/10.5555/2627435.2697065
33. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**(177), 1–81 (2019)
34. Freiesleben, T.: Counterfactual explanations & adversarial examples-common grounds, essential differences, and potential transfers. arXiv preprint arXiv:2009.05487 (2020)
35. Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles. Ann. Appl. Stat. **2**(3), 916–954 (2008). https://doi.org/10.1214/07-AOAS148
36. Friedman, J.H., et al.: Multivariate adaptive regression splines. Ann. Stat. **19**(1), 1–67 (1991). https://doi.org/10.1214/aos/1176347963
37. Garreau, D., von Luxburg, U.: Looking deeper into tabular lime. arXiv preprint arXiv:2008.11092 (2020)
38. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Stat. **24**(1), 44–65 (2015). https://doi.org/10.1080/10618600.2014.907095
39. Good, P.I., Hardin, J.W.: Common Errors in Statistics (and How to Avoid Them). Wiley (2012). https://doi.org/10.1002/9781118360125
40. Gosiewska, A., Biecek, P.: Do not trust additive explanations. arXiv preprint arXiv:1903.11420 (2019)
41. Greenwell, B.M.: PDP: an R package for constructing partial dependence plots. R J. **9**(1), 421–436 (2017). https://doi.org/10.32614/RJ-2017-016
42. Greenwell, B.M., Boehmke, B.C., McCarthy, A.J.: A simple and effective model-based variable importance measure. arXiv:1805.04755 (2018)
43. Gregorutti, B., Michel, B., Saint-Pierre, P.: Grouped variable importance with random forests and application to multiple functional data analysis. Comput. Stat. Data Anal. **90**, 15–35 (2015). https://doi.org/10.1016/j.csda.2015.04.002

44. Gretton, A., Bousquet, O., Smola, A., Schölkopf, B.: Measuring statistical dependence with Hilbert-Schmidt norms. In: Jain, S., Simon, H.U., Tomita, E. (eds.) ALT 2005. LNCS (LNAI), vol. 3734, pp. 63–77. Springer, Heidelberg (2005). https://doi.org/10.1007/11564089_7
45. Grömping, U.: Model-agnostic effects plots for interpreting machine learning models. Reports in Mathematics, Physics and Chemistry Report 1/2020 (2020)
46. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**(Mar), 1157–1182 (2003)
47. Hall, P.: On the art and science of machine learning explanations. arXiv preprint arXiv:1810.02909 (2018)
48. Hancox-Li, L.: Robustness in machine learning explanations: does it matter? In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* 2020, pp. 640–647. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3351095.3372836
49. Hand, D.J.: Classifier technology and the illusion of progress. Stat. Sci. **21**(1), 1–14 (2006). https://doi.org/10.1214/088342306000000060
50. Hastie, T., Tibshirani, R.: Generalized additive models. Stat. Sci. **1**(3), 297–310 (1986). https://doi.org/10.1214/ss/1177013604
51. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Comput. Biol. Chem. **34**(4), 215–225 (2010). https://doi.org/10.1016/j.compbiolchem.2010.07.002
52. Holm, S.: A simple sequentially rejective multiple test procedure. Scand. J. Stat. **6**(2), 65–70 (1979)
53. Hooker, G.: Discovering additive structure in black box functions. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2004, pp. 575–580. Association for Computing Machinery, New York (2004). https://doi.org/10.1145/1014052.1014122
54. Hooker, G.: Generalized functional ANOVA diagnostics for high-dimensional functions of dependent variables. J. Comput. Graph. Stat. **16**(3), 709–732 (2007). https://doi.org/10.1198/106186007X237892
55. Hooker, G., Mentch, L.: Please stop permuting features: an explanation and alternatives. arXiv preprint arXiv:1905.03151 (2019)
56. Janzing, D., Minorics, L., Blöbaum, P.: Feature relevance quantification in explainable AI: a causality problem. arXiv preprint arXiv:1910.13413 (2019)
57. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vis. **45**(2), 83–105 (2001). https://doi.org/10.1023/A:1012460413855
58. Karimi, A.H., Schölkopf, B., Valera, I.: Algorithmic recourse: from counterfactual explanations to interventions. arXiv:2002.06278 (2020)
59. Khamis, H.: Measures of association: how to choose? J. Diagn. Med. Sonography **24**(3), 155–162 (2008). https://doi.org/10.1177/8756479308317006
60. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**(1–2), 273–324 (1997)
61. König, G., Freiesleben, T., Bischl, B., Casalicchio, G., Grosse-Wentrup, M.: Decomposition of global feature importance into direct and associative components (DEDACT). arXiv preprint arXiv:2106.08086 (2021)
62. König, G., Freiesleben, T., Grosse-Wentrup, M.: A causal perspective on meaningful and robust algorithmic recourse. arXiv preprint arXiv:2107.07853 (2021)
63. König, G., Molnar, C., Bischl, B., Grosse-Wentrup, M.: Relative feature importance. In: 2020 25th International Conference on Pattern Recognition (ICPR), pp. 9318–9325. IEEE (2021). https://doi.org/10.1109/ICPR48806.2021.9413090

64. Krishnan, M.: Against interpretability: a critical examination of the interpretability problem in machine learning. Philos. Technol. **33**(3), 487–502 (2019). https://doi.org/10.1007/s13347-019-00372-9
65. Kuhle, S., et al.: Comparison of logistic regression with machine learning methods for the prediction of fetal growth abnormalities: a retrospective cohort study. BMC Pregnancy Childbirth **18**(1), 1–9 (2018). https://doi.org/10.1186/s12884-018-1971-2
66. König, G., Grosse-Wentrup, M.: A Causal Perspective on Challenges for AI in Precision Medicine (2019)
67. Lang, M., et al.: MLR3: a modern object-oriented machine learning framework in R. J. Open Source Softw. (2019). https://doi.org/10.21105/joss.01903
68. Laugel, T., Lesot, M.J., Marsala, C., Renard, X., Detyniecki, M.: The dangers of post-hoc interpretability: unjustified counterfactual explanations. In: Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, pp. 2801–2807. International Joint Conferences on Artificial Intelligence Organization (2019)
69. Laugel, T., Renard, X., Lesot, M.J., Marsala, C., Detyniecki, M.: Defining locality for surrogates in post-hoc interpretablity. arXiv preprint arXiv:1806.07498 (2018)
70. Lauritsen, S.M., et al.: Explainable artificial intelligence model to predict acute critical illness from electronic health records. Nat. Commun. **11**(1), 1–11 (2020). https://doi.org/10.1038/s41467-020-17431-x
71. Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C.: Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. Eur. J. Oper. Res. **247**(1), 124–136 (2015). https://doi.org/10.1016/j.ejor.2015.05.030
72. Liebetrau, A.: Measures of Association. No. Bd. 32; Bd. 1983 in 07, SAGE Publications (1983)
73. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 31–57 (2018). https://doi.org/10.1145/3236386.3241340
74. Lopez-Paz, D., Hennig, P., Schölkopf, B.: The randomized dependence coefficient. In: Advances in Neural Information Processing Systems, pp. 1–9 (2013). https://doi.org/10.5555/2999611.2999612
75. Lozano, A.C., Abe, N., Liu, Y., Rosset, S.: Grouped graphical granger modeling for gene expression regulatory networks discovery. Bioinformatics **25**(12), i110–i118 (2009). https://doi.org/10.1093/bioinformatics/btp199
76. Lundberg, S.M., et al.: From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. **2**(1), 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9
77. Lundberg, S.M., Erion, G.G., Lee, S.I.: Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888 (2018)
78. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: NIPS, vol. 30, pp. 4765–4774. Curran Associates, Inc. (2017). https://doi.org/10.5555/3295222.3295230
79. Makridakis, S., Spiliotis, E., Assimakopoulos, V.: Statistical and machine learning forecasting methods: concerns and ways forward. PloS One **13**(3) (2018). https://doi.org/10.1371/journal.pone.0194889
80. Matejka, J., Fitzmaurice, G.: Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 1290–1294 (2017). https://doi.org/10.1145/3025453.3025912

81. Molnar, C., Casalicchio, G., Bischl, B.: IML: an R package for interpretable machine learning. J. Open Source Softw. **3**(26), 786 (2018). https://doi.org/10.21105/joss.00786
82. Molnar, C., Casalicchio, G., Bischl, B.: Quantifying model complexity via functional decomposition for better post-hoc interpretability. In: Cellier, P., Driessens, K. (eds.) ECML PKDD 2019. CCIS, vol. 1167, pp. 193–204. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43823-4_17
83. Molnar, C., Freiesleben, T., König, G., Casalicchio, G., Wright, M.N., Bischl, B.: Relating the partial dependence plot and permutation feature importance to the data generating process. arXiv preprint arXiv:2109.01433 (2021)
84. Molnar, C., König, G., Bischl, B., Casalicchio, G.: Model-agnostic feature importance and effects with dependent features-a conditional subgroup approach. arXiv preprint arXiv:2006.04628 (2020)
85. Moosbauer, J., Herbinger, J., Casalicchio, G., Lindauer, M., Bischl, B.: Towards explaining hyperparameter optimization via partial dependence plots. In: 8th ICML Workshop on Automated Machine Learning (AutoML) (2020)
86. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. CoRR abs/1905.07697 (2019). http://arxiv.org/abs/1905.07697
87. Oh, S.: Feature interaction in terms of prediction performance. Appl. Sci. **9**(23) (2019). https://doi.org/10.3390/app9235191
88. Pearl, J., Mackenzie, D.: The Ladder of Causation. The Book of Why: The New Science of Cause and Effect, pp. 23–52. Basic Books, New York (2018). https://doi.org/10.1080/14697688.2019.1655928
89. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). https://doi.org/10.5555/1953048.2078195
90. Perneger, T.V.: What's wrong with Bonferroni adjustments. BMJ **316**(7139), 1236–1238 (1998). https://doi.org/10.1136/bmj.316.7139.1236
91. Peters, J., Janzing, D., Scholkopf, B.: Elements of Causal Inference - Foundations and Learning Algorithms. The MIT Press (2017). https://doi.org/10.5555/3202377
92. Philipp, M., Rusch, T., Hornik, K., Strobl, C.: Measuring the stability of results from supervised statistical learning. J. Comput. Graph. Stat. **27**(4), 685–700 (2018). https://doi.org/10.1080/10618600.2018.1473779
93. Reshef, D.N., et al.: Detecting novel associations in large data sets. Science **334**(6062), 1518–1524 (2011). https://doi.org/10.1126/science.1205438
94. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should I trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016). https://doi.org/10.1145/2939672.2939778
95. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. **1**(5), 206–215 (2019). https://doi.org/10.1038/s42256-019-0048-x
96. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: fundamental principles and 10 grand challenges. arXiv preprint arXiv:2103.11251 (2021)
97. Saito, S., Chua, E., Capel, N., Hu, R.: Improving lime robustness with smarter locality sampling. arXiv preprint arXiv:2006.12302 (2020)
98. Schallner, L., Rabold, J., Scholz, O., Schmid, U.: Effect of superpixel aggregation on explanations in lime-a case study with biological data. arXiv preprint arXiv:1910.07856 (2019)

99. Schmid, M., Hothorn, T.: Boosting additive models using component-wise p-splines. Comput. Stat. Data Anal. **53**(2), 298–311 (2008). https://doi.org/10.1016/j.csda.2008.09.009
100. Scholbeck, C.A., Molnar, C., Heumann, C., Bischl, B., Casalicchio, G.: Sampling, intervention, prediction, aggregation: a generalized framework for model-agnostic interpretations. In: Cellier, P., Driessens, K. (eds.) ECML PKDD 2019. CCIS, vol. 1167, pp. 205–216. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-43823-4_18
101. Seedorff, N., Brown, G.: Totalvis: a principal components approach to visualizing total effects in black box models. SN Comput. Sci. **2**(3), 1–12 (2021). https://doi.org/10.1007/s42979-021-00560-5
102. Semenova, L., Rudin, C., Parr, R.: A study in Rashomon curves and volumes: a new perspective on generalization and model simplicity in machine learning. arXiv preprint arXiv:1908.01755 (2021)
103. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, Cambridge (2014)
104. Simon, R.: Resampling strategies for model assessment and selection. In: Dubitzky, W., Granzow, M., Berrar, D. (eds.) Fundamentals of Data Mining in Genomics and Proteomics, pp. 173–186. Springer, Cham (2007). https://doi.org/10.1007/978-0-387-47509-7_8
105. Stachl, C., et al.: Behavioral patterns in smartphone usage predict big five personality traits. PsyArXiv (2019). https://doi.org/10.31234/osf.io/ks4vd
106. Stachl, C., et al.: Predicting personality from patterns of behavior collected with smartphones. Proc. Natl. Acad. Sci. (2020). https://doi.org/10.1073/pnas.1920484117
107. Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A.: Conditional variable importance for random forests. BMC Bioinform. **9**(1), 307 (2008). https://doi.org/10.1186/1471-2105-9-307
108. Štrumbelj, E., Kononenko, I.: Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. **41**(3), 647–665 (2013). https://doi.org/10.1007/s10115-013-0679-x
109. Sundararajan, M., Najmi, A.: The many Shapley values for model explanation. arXiv preprint arXiv:1908.08474 (2019)
110. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning, pp. 3319–3328. PMLR (2017)
111. Székely, G.J., Rizzo, M.L., Bakirov, N.K., et al.: Measuring and testing dependence by correlation of distances. Ann. Stat. **35**(6), 2769–2794 (2007). https://doi.org/10.1214/009053607000000505
112. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) **58**(1), 267–288 (1996). https://doi.org/10.1111/j.1467-9868.2011.00771.x
113. Tjøstheim, D., Otneim, H., Støve, B.: Statistical dependence: beyond pearson's ρ. arXiv preprint arXiv:1809.10455 (2018)
114. Valentin, S., Harkotte, M., Popov, T.: Interpreting neural decoding models using grouped model reliance. PLoS Comput. Biol. **16**(1), e1007148 (2020). https://doi.org/10.1371/journal.pcbi.1007148
115. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. Harv. JL Tech. **31**, 841 (2017). https://doi.org/10.2139/ssrn.3063289

116. Walters-Williams, J., Li, Y.: Estimation of mutual information: a survey. In: Wen, P., Li, Y., Polkowski, L., Yao, Y., Tsumoto, S., Wang, G. (eds.) RSKT 2009. LNCS (LNAI), vol. 5589, pp. 389–396. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02962-2_49
117. Watson, D.S., Wright, M.N.: Testing conditional independence in supervised learning algorithms. arXiv preprint arXiv:1901.09917 (2019)
118. Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., Grosse-Wentrup, M.: Causal interpretation rules for encoding and decoding models in neuroimaging. Neuroimage **110**, 48–59 (2015). https://doi.org/10.1016/j.neuroimage.2015.01.036
119. Williamson, B.D., Gilbert, P.B., Simon, N.R., Carone, M.: A unified approach for inference on algorithm-agnostic variable importance. arXiv:2004.03683 (2020)
120. Wu, J., Roy, J., Stewart, W.F.: Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. Med. Care S106–S113 (2010). https://doi.org/10.1097/MLR.0b013e3181de9e17
121. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc.: Ser. B (Statistical Methodology) **68**(1), 49–67 (2006). https://doi.org/10.1111/j.1467-9868.2005.00532.x
122. Zhang, X., Wang, Y., Li, Z.: Interpreting the black box of supervised learning models: visualizing the impacts of features on prediction. Appl. Intell. **51**(10), 7151–7165 (2021). https://doi.org/10.1007/s10489-021-02255-z
123. Zhao, Q., Hastie, T.: Causal interpretations of black-box models. J. Bus. Econ. Stat. 1–10 (2019). https://doi.org/10.1080/07350015.2019.1624293
124. Zhao, X., Lovreglio, R., Nilsson, D.: Modelling and interpreting pre-evacuation decision-making using machine learning. Autom. Constr. **113**, 103140 (2020). https://doi.org/10.1016/j.autcon.2020.103140
125. van der Zon, S.B., Duivesteijn, W., van Ipenburg, W., Veldsink, J., Pechenizkiy, M.: ICIE 1.0: a novel tool for interactive contextual interaction explanations. In: Alzate, C., et al. (eds.) MIDAS/PAP -2018. LNCS (LNAI), vol. 11054, pp. 81–94. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-13463-1_6

Chapter 7

Discussion

> "The end of a melody is not its goal, and yet if a melody has not reached its end, it has not reached its goal."
>
> — Friedrich Nietzsche

This thesis included five different research projects, all focused on different sub-problems within the field of XAI. However, all projects can be understood as attempts to gain more clarity about the different "whats" we want to explain in different contexts using XAI tools. As described in the introduction, four different levels can be distinguished: the associative model level, the causal model level, the associative phenomenon level and the causal phenomenon level.

I want to use this final Chapter to summarize the five papers with respect to the four levels view (Section 1), orient my work within the literature and display its significance (Section 2), discuss limitations (Section 3), and provide an outlook (Section 4).

1 Summary: The Five Papers and the Four "Whats"

In Paper I and Paper II, my co-authors and I argued that scientists are ultimately interested in the phenomenon rather than the ML model. Thus, XAI methods for scientists should aim at the associative or ideally even causal phenomenon level rather than the causal model level. Both papers shown how associative properties of the phenomenon can be inferred by analyzing the model: Paper I focused on which properties of the conditional distribution $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})$ can potentially be described with XAI and discussed whether we can

reach the causal phenomenon level; Paper II focused on quantifying the uncertainty for two specific properties that can be described with the PDP and PFI, respectively, and experimentally/formally demonstrating the quality of this uncertainty quantification.

Paper III and Paper IV discussed the different levels in the context of counterfactual explanations (CEs). In both papers, (my co-authors and) I argued that standard CEs for model audit only concern the causal model level and are agnostic with respect to the associative/causal phenomenon level. However, if we aim to attack the model or provide recourse recommendations to end-users, the target changes: attacking the model requires to search for adversarial examples (AEs) i.e. counterfactuals for which $\hat{\boldsymbol{Y}}$ differs from $\boldsymbol{Y}$; providing recourse recommendation requires the discussion of interventions that allow both $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$ to change in the same way.

In Paper V, the four levels view is explicit in pitfalls 1, 5 and 10. While papers I, II, III, and IV were purpose-centric – that is, they started with a purpose and developed from there – Paper V takes a methods-centric perspective.[1] Due to the shared purpose-centric perspective, the line of reasoning in Papers I, II, III, and IV is very similar: 1. We start with a fixed purpose, such as learning about the world in Paper I and Paper II, attacking an ML model in Paper III, or algorithmic recourse in Paper IV. 2. Next, we argue how this purpose decides which of the four levels to aim for. 3. Next, we show that current research on the topic lacks clarity about the different levels. 4. Finally, (my co-authors and) I offer a formal solution that is aware of the different levels and purpose, and we show how it is applied on an example.

2 Orientation and Significance of the Five Papers

In this section, I will discuss my work in the context of:

- the central criticisms of XAI,
- sub-fields in XAI, and
- (philosophy of) science.

In each context I will display the significance of my work and, where applicable, its societal, epistemic, and ethical implications.

[1] Purpose dependence is also briefly discussed in Paper V, pitfall 1. In my opinion, the purpose-centric perspective should generally be preferred to the the method-centered perspective because it ensures that XAI contributes to solving real problems.

2.1 The Central Criticisms of XAI

XAI as a field is pre-paradigmatic in a Kuhnian sense (Kuhn, 1970) – its foundations are contested within the field and there are competing paradigm candidates. My work focused primarily on one of these paradigms, which is probably the most mature competitor on the market – the so-called *model-agnostic XAI* (Molnar, 2020). In model-agnostic XAI, the ML model is reduced to an input-output mapping (Scholbeck et al., 2019). The other major competitor, model-specific XAI, exploits the specific model-structure and model-properties. In Paper I, my co-authors and I argue why parts of model-specific XAI, namely those that analyze individual model elements, face fundamental problems because they assume that ML models learn humanly-accessible features (Olah et al., 2020). My primary focus on the model-agnostic paradigm stems from its universal applicability and these arguments from Paper I against interpretations of individual model elements.

As discussed in Section 1.4, current XAI[2] faces several fundamental criticisms: 1. XAI conflates or lacks goals; 2. XAI provides unsatisfactory explanation evaluation; 3. XAI provides misleading, non-robust explanations that lack uncertainty quantification. These criticisms have all also been discussed in more detail in Paper V, where we pointed out pitfalls in using XAI methods. I believe that all of the criticisms listed are valid, they describe key problems within the field XAI, and if we want to move forward as a field, these criticisms must be addressed – this was one major motivation behind my thesis.

The Problem XAI Tries to Solve is Ill-defined Lipton (2018) most prominently argues that interpretability is not a monolithic concept, but encompasses various endeavors such as increasing trust, improving model robustness, or providing legally required explanations for algorithmic decisions. He therefore criticizes the field for lacking a proper problem formulation and states:

> "When we have solid problem formulations, flaws in methodology can be addressed by articulating new methods. But when the problem formulation itself is flawed, neither algorithms nor experiments are sufficient to address the underlying problem."
> (Lipton, 2018)

My work has been strongly inspired and can be seen as a direct response to his fundamental criticism. Particularly in Paper I and Paper IV, my co-authors

[2] this applies to both model-agnostic and model-specific XAI

and I show how in specific contexts the problem we want to solve with XAI can be well defined and consequently approached. This is not to say that my work has solved the central problem of XAI, which is to provide a universal definition of explainability. Rather, my co-authors and I narrowed down the problem to a more fine-grained aspect, and specific purpose, and shown that XAI methods can(not) help to solve it.

I see the critique of a lacking problem formulation (fundamental criticism 1) as the most fundamental, it is one cause of the problem of unsatisfactory evaluation (fundamental criticism 2) and the problem of misleading explanations (fundamental criticism 3). The goal determines the evaluation strategy and the success conditions. Explanations themselves cannot mislead if they satisfy a prespecified goal, but the goal itself might be morally questionable. Moreover, uncertainty can be quantified if we know which estimate is uncertain with respect to which goal.

XAI Explanations are Hard to Evaluate Doshi-Velez & Kim (2017) provide a framework to discuss the evaluation of interpretability along three levels, application-grounded, human-grounded, and functionally grounded evaluation, with the first two evaluation levels requiring cumbersome and expensive experiments with humans. I agree that the gold standard we should aim for is testing whether the explanation does, when given to a human, improve her performance on the task at hand. However, the focus in my work on the different explananda allows to establish the sanity of XAI methods at an earlier state – does the method provide any explanation at all for the phenomenon it is supposed to explain? As argued in Section 2, the right explanandum must be established prior to the choice of the best possible explanation, and thus before psychological experiments are conducted. Moreover, we show in Paper I that for the case of scientific inference, where the goal is to learn an aspect of $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})$, a purely formal evaluation approach is sufficient and no human experiments are required.

XAI Explanations are Misleading and do not Display Uncertainties Rudin (2019) argues in her seminal work against the use of model-agnostic XAI (at least for high-stakes decisions) and recommends instead the use of interpretable models[3]. While I do not defend some of the methods she attacks, such as LIME or saliency maps, my work shows that the criticism is wrong in its generality. For instance, global model-agnostic techniques

[3] In Paper I, Paper II, and Paper V, my co-authors and I highlighted that her proposed solution to solely rely on interpretable models is a strong constraint for solving practical problems.

such as conditional PDP and conditional PFI (see Paper I and Paper II) can be epistemically well-grounded and, under certain conditions, allow us to learn something about the underlying phenomenon. In addition, in Papers III and IV, my co-authors and I have described ways in which counterfactual explanations can be less misleading by tailoring them to a particular purpose such as recourse or contestability.

The lacking uncertainty quantification of XAI explanations has been widely noted (Rudin et al., 2022; Doshi-Velez & Kim, 2017; Watson, 2022; Molnar et al., 2020). My co-authors and I address the problem for specific methods and purposes in Paper I, Paper II, and Paper IV. In Paper I and Paper II, we analyze the uncertainty arising from the randomness of the data, the learning process, and Monte-Carlo integration. In Paper IV, we focus on the uncertainty with respect to recourse acceptance guarantees under certain actions of the end-user. In addition, in Paper V, we discuss the problem of quantifying uncertainty and identify possible solutions.

2.2 Counterfactual Explanations, Algorithmic Recourse, and Adversarial Examples

Paper III and Paper IV both deal with counterfactual explanations. Paper III and Paper IV both deal with counterfactual explanations. In this section, we will discuss these two papers and their relationship in more detail.

Counterfactual Explanations and Algorithmic Recourse Wachter et al. (2017) proposed counterfactual explanations as a method that can serve various purposes, such as understanding individual ML predictions, contesting algorithmic decisions, and providing end-users with action recommendations to reverse unfavorable decisions. Especially the latter purpose – recommending recourse – gained increasing attention in the literature. Recommending recourse requires to: put additional constraints, such as actionability (Ustun et al., 2019); specify actions rather than alternative scenarios (Karimi et al., 2020); incorporate the causal dependencies between input features in the real world (Karimi et al., 2021).

In Paper IV, we build on these ideas on algorithmic recourse, but attack their central premise, namely we argue that it is more important in recourse to change the target rather than just the prediction; that is, we should target a different explanandum with our recourse recommendations. Our proposal has strong epistemic, ethical, and societal implications. On the epistemic side,

if we focus on the target rather than the prediction, our recourse recommendation becomes independent of the ML model mechanism and instead must be generated from causal knowledge about the world. Thus, we cannot simply apply XAI methods to solve this problem. On the ethical side, we argue that reversing only the prediction, but not the target, poses a problem not only for the model-authority, but potentially also the individual. In the stroke risk prediction from chapter 1, imagine a case in which the individual reduced her predicted stroke risk by lowering her blood-pressure using medication, but this medication has the side-effect of increasing her stroke risk. Our view here is consistent with recent arguments from moral philosophy promoting causal recourse explanations (Vredenburgh, 2022). Finally, on the societal side, unlike previous proposals, our proposal does not provide an incentive for data subjects to game the predictor, i.e., the functioning of ML models remains intact. If ML models serve important societal functions, e.g. correctly diagnosing patients or giving loans only to people who can repay them, our recourse formalization should be favored over the that of Karimi et al. (2020).

Counterfactual Explanations and Adversarial Examples In Paper III, counterfactual explanations are related to adversarial examples. This also happened in the original proposal of Wachter et al. (2017), however they incorrectly identify counterfactuals with adversarials, ignoring their relationship to the prediction target, as argued in Paper III. Instead, my paper shows that adversarial examples can be viewed as a special kind of counterfactuals, namely those that change the prediction but not the underlying target. It therefore allows for the integration of these two hitherto quite independent fields of research (Pawelczyk et al., 2022). Moreover, as I highlight the possible connections and introduce the main concepts, the paper also serves as a dictionary for the various communities to translate between the fields.

Adversarial Examples and Recourse Recommendations Taking the results of Paper III and Paper IV together has the potential to illuminate the discussion strongly. Together, the papers show that adversarial examples and recourse recommendations are formally opposite objects[4]: adversarials revert the prediction but maintain the target, while for recourse we want to revert both target and prediction in agreement; adversarials make widespread imperceptible changes, while for recourse recommendations sparse changes better allow subjects to act; in adversarials we break causal dependencies,

[4]This relationship is also captured in the papers, but with a focus on the notion of contesting algorithmic decisions.

while in recourse causal dependencies must be respected. Whenever two concepts that are their opposite, they can profit from each other. Adversarial examples research, in my view, would profit from reverting the causal approach that recourse research takes. On the other side, recourse research can strongly profit from adversarials research with respect to less structured data domains such as audio or image data.

2.3 (Philosophy of) Science

Paper I and Paper II focus on the use of XAI for scientific inference, i.e. for learning about the real world phenomenon represented by the ML model. In this section, we discuss the potential implications of these two papers for scientific practice and on debates within philosophy of science.

Significance for Science Paper I and Paper II have a very ambitious goal – to show that ML in alliance with XAI has the potential to complement and even partially replace classical (statistical) modeling for scientific purposes. While it is widely acknowledged that ML is superior when it comes to mere prediction, it was held that for other scientific goals, such gaining knowledge or providing causal explanations, classical (in Paper I, we call it elementwise representational (ER)) modeling remains superior (Shmueli et al., 2010). However, in Paper I, we showed that questions concerning the conditional distribution $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{X})$ can be addressed with the help of ML model analysis. Moreover, we can quantify the epistemic uncertainty with respect to these answers and provide confidence estimates. Paper II goes here more into depth and fleshes out this uncertainty quantification in full detail for the PDP and PFI.

Many questions previously explored with statistical modeling can now better be addressed because, as we argue in the papers, ML relies less on strong assumptions about the nature of the modeled phenomenon. Instead, ML models automatically capture complex relationships and interactions between features. However, Paper I and Paper II highlight not only the strengths of supervised ML for drawing inference but also its weaknesses. In general, the inferences that XAI allows to draw about the phenomenon should not be interpreted causally at the phenomenon level. ML models capture only the associational relationship in the data, as the effects of the target can be just as predictive as the causes. We point this out as a general limitation of XAI methods applied to supervised learning models in Paper I, Paper II, and Paper V.

As scientists are already starting to use XAI techniques to learn about their phenomenon (sometimes in the wrong ways), our work comes at the right

time. In particular, for fields where the data we obtain from measurement is high-dimensional and hard to interpret – think of neuroscience, molecular biology, or particle physics – ML combined with XAI could open new paths to study phenomena (Cichy & Kaiser, 2019; Hasson et al., 2020).

Significance for Philosophy of Science Sullivan (2020) and Cichy & Kaiser (2019) have famously argued that it is possible to learn about real-world phenomena via ML. However, the connection between the ML model and the phenomenon – Sullivan calls this the *link uncertainty* – must be small. Unfortunately, both remain very vague about what this link uncertainty consists of and how it can be reduced. In Paper I, my co-authors and I argue that one needs to rethink the question of representationality in order to establish the link between the model and the phenomenon. We highlight that classical scientific models are largely built on the idea that every model element is representing an interpretable aspect of the phenomenon (Frigg & Nguyen, 2021). We show that it is not individual model elements, but only the entire model that can be interpreted in terms of the phenomenon. To assess the link between the model and the world, the only access point is through strong predictive performance. To introspect the aspect of the phenomenon that is represented by the whole model – this is, to zoom into this aspect – XAI techniques tailored to this task must be used. We note that our approach is applicable to any model where it is not the model elements that represent the phenomenon, but the entire model. On the clarificatory side, Paper I also dampens the unwarranted hopes by Sullivan (2020); Watson (2022); Zednik & Boelsen (2022) to interpret ML models as representing causal relationships; associations is all that supervised ML can offer us.[5]

The second major debate within philosophy of science to which my work has contributed concern the role of XAI in science and how much opacity ultimately limits the use of ML in science (Cichy & Kaiser, 2019; Creel, 2020; Boge, 2021; Zednik & Boelsen, 2022). Boge (2021) describes two different dimensions of ML opacity, opacity with respect to the concepts learned with ML algorithms and opacity with respect to ML model predictions. He points out the problem that, especially in the first case, a gap may arise between scientific discovery and scientific explanation. In Paper I, my co-authors and I strengthen his arguments and describe that people will have difficulty understanding features learned by ML models due to the distributed representation in neural networks. We conclude that applying model-specific XAI techniques to understand these features, while intriguing, is hopeless and will lead re-

[5] This might change if additional assumptions are in place (Peters et al., 2017).

searchers astray. Boge's second opacity concerns the functional opacity of the prediction model; many noted that model-agnostic XAI could help make the functional properties more transparent, but they remained unclear on what opacity they want reduced (model or world) and what explanations they hope to find (causal or associational) (Cichy & Kaiser, 2019; Zednik & Boelsen, 2022; Roscher et al., 2020). Watson (2022) was clearer about the former, but, as we argue in Paper I, is wrong about the latter, thinking that XAI methods provide causal explanations about the world. In Paper I, we clearly separate between explanations that aim to audit the model from those that aim to draw scientific inference, but also describe their interaction. Importantly, in Paper I and Paper II, we argue that if we strive for scientific inference with XAI, it is crucial to integrate only ML predictions on realistic data.

Finally, I see the significance of my work also in the connections it allows to draw between classical statistical modeling and ML modeling, which is particularly relevant to philosophers of statistics. For example, my co-authors and I have shown how ML modeling and classical statistical modeling can be fused to perform statistical inference using XAI techniques. The important contribution of my work is here to provide a representation of the underlying data generating mechanism via statistical decision theory. This link between the ML model and the underlying data generating mechanism, which is strongly emphasized in classical statistical modeling (Romeijn, 2022), will allow to transfer a whole range of other concepts from statistics to ML, such as hypotheses testing. Sampling of data is another big problem in both statistics and ML, which has been touched by Paper I, Paper II, and Paper V. We argue that permuting individual features or intervening on them risks creating unrealistic data (Hooker & Mentch, 2019). Think of a case where the data describes basketball players and we change a individual players size from 2.10 meters to 1.50 without accounting or the resulting changes in his weight or playing skills. With such unrealistic data, the ML model must extrapolate and therefore cannot usually be reliably interpreted.[6] My co-authors and I therefore emphasize in Paper II the importance of conditional sampling of features and also its difficulty.

[6] Even worse, data might not only be unrealistic but completely impossible, imagine a person with 2.10 meter but a weight of 2 kg.

3 Limitations

Now that I have contextualized the contributions of my work across literatures and fields, I would like to use this section to point out the limitations of my work.

General Limitations of the Four "Whats" As I argued before, the purpose of the explanation determines the correct explanandum. Since there are so many possible purposes of XAI explanations, there is still room for confusion about the correct explanandum. In my work, I have only addressed the question of the explanandum for a very few selected purposes such as scientific inference, adversarial attacks, or algorithmic recourse, focusing on a limited number of methods such as CE, PDP, and PFI. There is still much to be done, and some purposes of XAI may still need to be found.

The list of four explananda may be considered incomplete. Indeed, there are quite different explanations when the goal is to learn high-level concepts with XAI or to introspect model elements (Olah et al., 2020, 2017; Bau et al., 2018). Also, for other learning paradigms such as unsupervised or reinforcement learning, we can go beyond the simplistic setting of explaining how $\boldsymbol{X}$ relates to $\hat{\boldsymbol{Y}}$ or $\boldsymbol{Y}$. Nonetheless, for standard supervised learning setups and with a focus on model-agnostic XAI, the four levels are probably exhaustive and can guide many current XAI discussions.

Specific Limitations of Paper I and Paper II Paper I and Paper II, as described above, are similar in spirit, although they put different emphasis on the conceptual and the formal work. Thus, they share some of their limitations. The most important limitation, in my opinion, is that the proposed confidence intervals are based on a very strong assumption, namely, learner unbiasedness. Learner unbiasedness means that the learning algorithm learns the optimal prediction model in expectation about the learning situations. This assumption is enormously far from any guarantees that statistical learning theory can give us so far for complex ML models such as random forests or neural networks (Bishop & Nasrabadi, 2006). Learner unbiasedness requires that we provide our learning algorithm with the right inductive bias by choosing an appropriate optimizer, select solid hyperparameters and choosing the right model class for the problem. Perhaps learner unbiasedness is not such a strong constraint in practice after all. Human experts have control over which models they include when estimating confidence. They will only strive for high-performing models and adjust the inductive bias accordingly.

The second major limitation is the conditional sampling that must be performed in order to apply XAI methods for scientific inference in practice. Conditional sampling is similarly difficult to the original problem we were trying to to solve, which was learning an aspect of the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$. Just like ML models, conditional samplers must be learned from data, which introduces another source of error and uncertainty. This learning process of conditional samplers is ignored in both Paper I and Paper II, but it is necessary to apply these techniques in practice.

A third limitation I see is the following: my work has shown that XAI allows to draw scientific inferences and learn about the phenomenon, but it has also made clear that for some problems there are much simpler and even more accurate techniques to draw the same inference without using XAI. Let me use an analogy to illustrate the problem: Suppose you have build an extremely complex but fairly accurate Lego model of the city Munich, which contains a counterpart of all humans, cars, trees, lakes, and so on. Now, suppose that the question you want to answer with this model is very simple: how long is the queue at my favorite falafel restaurant right now? Find the restaurant in your Lego model, maybe move a few pieces aside to get there and count the Lego figures. However, if the length of the queue is your *only question*, you don't even need to build the Lego model. You could just walk into the restaurant and count the people, and your answer would probably be very accurate. Training a complex ML model and then applying XAI techniques is similar to building a Lego representation of a city just to answer a few specific questions about the world. Using XAI for inference is only useful if you have a complex predictive model anyway and want to use it as an add-on to draw conclusions.

Paper I and Paper II share some, but not all, weaknesses. For instance, one limitation of Paper II is its focus on PDP and PFI without giving a clear rationale for why exactly these two methods should be relevant for scientific inference. This is improved in Paper I, where a question regarding the phenomenon guides the choice of the XAI method. On the other side, Paper I does not provide the experimental and formal depth of Paper II, so it remains unclear how well the results from Paper I are transferable to non-ideal environments.

Specific Limitations of Paper III The main limitation of Paper III is the following: while conceptually it provides clear guidance on how to formally separate CEs and AEs via their relation to the true target label (misclassification), it does not allow for an operationalization of misclassification that would make the optimization problem different for AEs and CEs. This is a general gap in AE research. However, in the paper I describe ways to achieve such an operationalization using causal models.

Specific Limitations of Paper IV Unlike Paper III, Paper IV does provide a formalization for changing the state of the prediction target using causal models. Our formalism from Paper IV can therefore be operationalized and used in practice. The main weakness of Paper IV is the great amount of causal knowledge required for such an operationalization, which is generally not readily available but needs domain experts; in particular, the individualized recourse for which we need SCMs might has more theoretical than practical relevance. It should be noted, however, as we argue in the paper, it is not possible to tackle the recourse problem without such causal knowledge.

Specific Limitations of Paper V As a commentary and survey paper of XAI methods with emphasis on their pitfalls, the limitations of Paper V lie in its limited coverage of the field and its problems. The focus is not only limited to model-agnostic XAI methods, but also places a strong emphasis on global rather than local methods. The focus is not only limited to model-agnostic XAI methods, but also places a strong emphasis on global rather than local methods. Furthermore, the focus is only on some methods and problems within model-agnostic XAI that we believe are relevant to researchers.

4 Outlook

I would like to use this final section of the paper to preview open problems and future work, and to speculate on future developments in XAI research.

The Four "Whats" I believe that the various explananda will receive increasing attention in future XAI research. XAI is still trying to find its core concepts, and what XAI explains must necessarily belong to those core concepts. Many discussions that are currently going on about the right sampling techniques, the Rashomon effect of explanations (i.e. many explanations can explain the same explanandum), or the usefulness of particular XAI techniques can be seen as disguised discussion about the right explanandum. XAI techniques that target the same explanandum can be used across applications that share this explanandum. Perhaps we will see further diversification of explanations in XAI, but the four proposed levels will certainly be an essential part of it.

Future Work Paper I distinguishes between scientific models in which each model element represents an aspect of the phenomenon (elementwise representationality) and models where only the whole model represents (holistic representationality). The relationship between elementwise and holistic representationality needs to be explored in further depth. In particular, it would be interesting to describe a classical mechanistic model, e.g. the Hodgkin-Huxley model from computational neuroscience (Hodgkin & Huxley, 1952), with a holistic representational model and recover its different aspects using tools of model analysis.

Paper III shows that current research obtains AE by accident, solving the standard optimization problem for counterfactual explanations. However, this approach does not work, for instance, for tabular data structures in general. Also as ML models get better, it becomes more difficult to attack them. Currently, there is no formalized notion of misclassification, a gap that needs to be addressed. The formalization of recourse in Paper IV would be a good starting point for this endeavor. A formalization of misclassification would be relevant not only from the attacker's perspective, but especially from the defender's perspective. In particular, it would be interesting from the perspective of algorithmic fairness. If users receive CEs that indicate algorithmic errors, they have grounds to contest unfavorable decisions. User contestability would be an ethical and social milestone in the application of ML systems.

Paper IV gives a reformulation of the recourse problem. The superiority of our formulation compared to the one by Karimi et al. (2020) could be more clearly emphasized if both approaches were implemented in a real-world social context where people respond to the recommendations they receive. Since recourse will eventually be put into practice, it is only a matter of time before the consequences can be observed.

Altogether, in the future I would like to continue working on purposes of ML beyond predictive performance, similar to my work on scientific inference and algorithmic recourse. In particular, I would like to work on algorithmic fairness, i.e., the problem that ML model predictions can be based on sensitive attributes. In addition, I would like to further explore the methodological implications of using ML in science and the scientific models that scientists can create.

Future of XAI One might think that XAI is only a temporary study. Once ML models become arbitrarily accurate in their predictions, we no longer need explanations to bolster our trust in the system. We may not understand how these models arrive at their conclusions, but we don't need to because they are as reliable as a calculator adding two numbers together. However, this idea focuses on one purpose of explanations, which is to gain trust. For other purposes, such as algorithmic recourse or scientific inference, it remains unclear why increasing model accuracy would resolve the need for explanations. In the natural or social sciences, for example, the search for explanations and for models with explanatory power is of crucial interest beyond mere predictive accuracy (Longino, 2018; Shmueli et al., 2010).

Even if we were to assume that trust is the only concern, perfect accuracy would not eliminate it. Accuracy is always measured relative to a particular loss function compared to a particular ground truth. The cases where we really demand trust are the cases where the correct loss-function and also the ground-truth are controversial, explanations and transparent decision making are therefore necessary even with perfect accuracy.

All in all, I am optimistic that XAI is there to stay. I hope that XAI will contribute to using ML for the benefit of humanity, and I wish that my work will provide a push (however small) in that direction.

Bibliography

Achinstein, P. (1983). *The nature of explanation*. Oxford University Press on Demand.

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (pp. 275–285).

Antaki, C. & Leudar, I. (1992). Explaining in conversation: Towards an argument model. *European Journal of Social Psychology*, 22(2), 181–194.

Bailer-Jones, D. M. (2003a). Models, theories and phenomena. *Proceedings of Logic Methodology and Philosophy of Science*.

Bailer-Jones, D. M. (2003b). When scientific models represent. *International studies in the philosophy of science*, 17(1), 59–74.

Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. (2017). Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6541–6549).

Bau, D., Zhu, J.-Y., Strobelt, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2018). Gan dissection: Visualizing and understanding generative adversarial networks. *arXiv preprint arXiv:1811.10597*.

Bishop, C. M. & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Boehme, A. K., Esenwa, C., & Elkind, M. S. (2017). Stroke risk factors, genetics, and prevention. *Circulation research*, 120(3), 472–495.

Boge, F. J. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*, (pp. 1–33).

Bokulich, A. (2017). Models and explanation. In *Springer handbook of model-based science* (pp. 103–118). Springer.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3), 199–231.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, (pp. 419–437).

Chen, H., Janizek, J. D., Lundberg, S., & Lee, S.-I. (2020). True to the model or true to the data? *arXiv preprint arXiv:2006.16234*.

Chin-Parker, S. & Cantelon, J. (2017). Contrastive constraints guide explanation-based category learning. *Cognitive science*, 41(6), 1645–1655.

Cichy, R. M. & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in cognitive sciences*, 23(4), 305–317.

Covert, I., Lundberg, S. M., & Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33, 17212–17223.

Creel, K. A. (2020). Transparency in complex computational systems. *Philosophy of Science*, 87(4), 568–589.

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature* (pp. 448–469).: Springer.

De Regt, H. W. & Dieks, D. (2005). A contextual approach to scientific understanding. *Synthese*, 144(1), 137–170.

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Erasmus, A., Brunet, T. D., & Fisher, E. (2021). What is interpretability? *Philosophy & Technology*, 34(4), 833–862.

Fatima, M., Pasha, M., et al. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01), 1.

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177), 1–81.

Friedman, J. H. et al. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1), 1–67.

Frigg, R. (2002). Models and representation: Why structures are not enough. *Measurement*.

Frigg, R. & Nguyen, J. (2021). Scientific Representation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of science*, 71(5), 742–752.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Halpern, J. Y. & Pearl, J. (2020). Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*.

Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3), 416–434.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Hempel, C. G. et al. (1965). *Aspects of scientific explanation*, volume 1. Free Press New York.

Hempel, C. G. & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of science*, 15(2), 135–175.

Hesslow, G. (1988). The problem of causal selection. *Contemporary science and natural explanation: Commonsense conceptions of causality*, (pp. 11–32).

Hiabu, M., Mammen, E., & Meyer, J. T. (2020). Random planted forest: a directly interpretable tree ensemble. *arXiv preprint arXiv:2012.14563*.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107(1), 65.

Hilton, D. J. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, 2(4), 273–308.

Hitchcock, C. R. (1995). Discussion: Salmon on explanatory relevance. *Philosophy of Science*, 62(2), 304–320.

Hodgkin, A. L. & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4), 500–544.

Hooker, G. & Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. *arXiv e-prints*, (pp. arXiv–1905).

Janzing, D., Minorics, L., & Blöbaum, P. (2020). Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics* (pp. 2907–2916).: PMLR.

Jebeile, J. & Kennedy, A. G. (2015). Explaining with models: The role of idealizations. *International Studies in the Philosophy of Science*, 29(4), 383–392.

Josephson, J. R. & Josephson, S. G. (1996). *Abductive inference: Computation, philosophy, technology*. Cambridge University Press.

Kahneman, D. & Tversky, A. (1981). *The simulation heuristic*. Technical report, Stanford Univ CA Dept of Psychology.

Karimi, A.-H., Schölkopf, B., & Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 353–362).

Karimi, A.-H., Von Kügelgen, J., Schölkopf, B., & Valera, I. (2020). Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *Advances in Neural Information Processing Systems*, 33, 265–277.

Kashima, Y., McKintyre, A., & Clifford, P. (1998). The category of the mind: Folk psychology of belief, desire, and intention. *Asian Journal of Social Psychology*, 1(3), 289–313.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning* (pp. 2668–2677).: PMLR.

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.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., & Liang, P. (2020). Concept bottleneck models. In *International Conference on Machine Learning* (pp. 5338–5348).: PMLR.

Krishnan, M. (2020). Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology*, 33(3), 487–502.

Kuhn, T. S. (1970). *The structure of scientific revolutions*, volume 111. Chicago University of Chicago Press.

Lallukka, T., Millear, A., Pain, A., Cortinovis, M., & Giussani, G. (2017). Gbd 2015 mortality and causes of death collaborators. global, regional, and national life expectancy, all-cause mortality, and cause-specifi c mortality for 249 causes of death, 1980-2015: a systematic analysis for the global burden of disease study 2015 (vol 388, pg 1459, 2016). *Lancet*, 389(10064), E1–E1.

Liao, Q. V. & Varshney, K. R. (2021). Human-centered explainable ai (xai): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790*.

Lipton, P. (1990). Contrastive explanation. *Royal Institute of Philosophy Supplements*, 27, 247–266.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3), 31–57.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive psychology*, 61(4), 303–332.

Longino, H. E. (2018). *The fate of knowledge*. Princeton University Press.

Mackay, M. T., Wiznitzer, M., Benedict, S. L., Lee, K. J., Deveber, G. A., Ganesan, V., & Group, I. P. S. S. (2011). Arterial ischemic stroke risk factors: the international pediatric stroke study. *Annals of neurology*, 69(1), 130–140.

Mahendran, A. & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3), 233–255.

Malle, B. F. (2011). Time to give up the dogmas of attribution: An alternative theory of behavior explanation. In *Advances in experimental social psychology*, volume 44 (pp. 297–352). Elsevier.

Martin-Barragan, B., Lillo, R., & Romo, J. (2014). Interpretable support vector machines for functional data. *European Journal of Operational Research*, 232(1), 146–155.

McClure, J. (2002). Goal-based explanations of actions and outcomes. *European review of social psychology*, 12(1), 201–235.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267, 1–38.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 279–288).

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 11(3-4), 1–45.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 417–431).: Springer.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, 5(3), e00024–001.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, 2(11), e7.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3), 441–459.

Pawelczyk, M., Agarwal, C., Joshi, S., Upadhyay, S., & Lakkaraju, H. (2022). Exploring counterfactual explanations through the lens of adversarial examples: A theoretical and empirical analysis. In *International Conference on Artificial Intelligence and Statistics* (pp. 4574–4594).: PMLR.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press.

Qi, Z., Khorram, S., & Li, F. (2019). Visualizing deep networks by optimizing with integrated gradients. In *CVPR Workshops*, volume 2.

Reutlinger, A. & Saatsi, J. (2018). *Explanation beyond causation: Philosophical perspectives on non-causal explanations*. Oxford University Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144).: ACM.

Romeijn, J.-W. (2022). Philosophy of Statistics. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition.

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8, 42200–42216.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16, 1–85.

Russell, S. J. (2010). *Artificial intelligence a modern approach*. Pearson Education, Inc.

Salmon, W. C. (1971). *Statistical explanation and statistical relevance*, volume 69. University of Pittsburgh Pre.

Salmon, W. C. (1979). Why ask,'why?'? an inquiry concerning scientific explanation. In *Hans Reichenbach: logical empiricist* (pp. 403–425). Springer.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Scholbeck, C. A., Molnar, C., Heumann, C., Bischl, B., & Casalicchio, G. (2019). Sampling, intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations. *arXiv preprint arXiv:1904.03959*.

Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710.

Seuß, D. (2021). Bridging the gap between explainable ai and uncertainty quantification to enhance trustability. *arXiv preprint arXiv:2105.11828*.

Shmueli, G. et al. (2010). To explain or to predict? *Statistical science*, 25(3), 289–310.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Slugoski, B. R., Lalljee, M., Lamb, R., & Ginsburg, G. P. (1993). Attribution in conversational context: Effect of mutual knowledge on explanation-giving. *European Journal of Social Psychology*, 23(3), 219–238.

Smith, B. C. (2019). *The promise of artificial intelligence: reckoning and judgment*. MIT Press.

Strevens, M. (2011). *Depth: An account of scientific explanation*. Harvard University Press.

Štrumbelj, E. & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3), 647–665.

Sullivan, E. (2020). Understanding from machine learning models. *The British Journal for the Philosophy of Science*.

Tetlock, P. E. & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of personality and social psychology*, 57(3), 388.

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 10–19).

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

von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust ai. *Philosophy & Technology*, 34(4), 1607–1622.

Vredenburgh, K. (2022). The right to explanation. *Journal of Political Philosophy*, 30(2), 209–229.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31, 841.

Watson, D. S. (2022). Conceptual challenges for interpretable machine learning. *Synthese*, 200(1), 1–33.

Woodward, J. (1989). The causal mechanical model of explanation. *Minnesota studies in the philosophy of science*, 13.

Woodward, J. (2005). *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*. Metaphysics Research Lab, Stanford University, Summer 2021 edition.

Zednik, C. (2021). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology*, 34(2), 265–288.

Zednik, C. & Boelsen, H. (2022). Scientific exploration and explainable artificial intelligence. *Minds and Machines*, (pp. 1–21).

List of Abbreviations

Abbreviations Used in the Thesis

AE	Adversarial Example
AI	Artificial Intelligence
ALE	Accumulated Local Effect
ANN	Artificial Neural Network
BMI	Body-Mass-Index
CART	Classification and Regression Trees
CE	Counterfactual Explanation
CNN	Convolutional Neural Network
CR	Causal Recourse
DGP	Data-Generating Process
DL	Deep Learning
DN	Deductive-Nomological account of explanation by Hempel
DNN	Deep Neural Network
ER	Elementwise Representationality
FANOVA	Functional ANalysis Of VAriance
GAN	Generative Adversarial Network
GAM	Generalized Additive Model
GOFAI	Good Old Fashioned AI
HR	Holistic Representationality
ICE	Individual Conditional Expectation
ICI	Individual Conditional Importance
ICR	Improvement-focused Causal Recourse
IML	Interpretable Machine Learning
LASSO	Least Absolute Shrinkage and Selection Operator
LSTM	Long Short Term Memory
LIME	Local Interpretable Model-agnostic Explanations
MC	Monte Carlo
ML	Machine Learning
MSE	Mean Squared Error
NLP	Natural Language Processing
PCA	Principal Component Analysis
PCC	Pearson Correlation Coefficient
PDP	Partial Dependence Plot
PFI	Permutation Feature Importance

RF	Random Forest
RL	Reinforcement Learning
SAGE	Shapley Additive Global importancE
SCM	Structural Causal Model
SHAP	SHapley Additive exPlanations
SR	Statistical-Relevance account of explanation by Salmon
SVM	Support Vector Machine
XAI	eXplainable Artificial Intelligence

Timo Freiesleben

Curriculum Vitae

Research Interests

Areas of Specialization	Explainable AI, Causality, Philosophy of AI
Areas of Competence	Ethics of AI, Philosophy of Science, Machine Learning, Philosophy of Statistics, Decision Theory, Probability Theory, Logic

Education

10/2019–Now	**Ph.D. in Systemic Neurosciences**, *Graduate School of Systemic Neurosciences München (LMU)*, Project: What Does Explainable AI explain? Supervised by Prof. Dr. Stephan Hartmann, Dr. Álvaro Tejero Cantero, Prof. Dr. Paul Taylor, Prof. Dr. Agnieszka Wykowska
10/2018–09/2019	**M.Sc. in Computer Science**, *Ludwig-Maximilians-Universität München (LMU)*, Taken courses on Deep Learning & AI Without Graduation
10/2016–09/2018	**M.A. in Logic and Philosophy of Science**, *Munich Center for Mathematical Philosophy (MCMP), Ludwig-Maximilians-Universität München (LMU)*, *Very Good*, Thesis on **Incorporating Intuitions into Decision Making Rationally** Supervised by Dr. Rush Stewart and Prof. Dr. Dr. Hannes Leitgeb
10/2012–09/2016	**B.Sc. in Mathematics**, *Eberhard Karls Universität Tübingen*, *Very Good*, Thesis on **Ramification and Arithmetic Schemes** Supervised by Prof. Dr. Jürgen Hausen
08/2015–07/2016	**Erasmus Exchange Year**, *University of Oslo*, With a focus on Mathematical Logic and Computability Theory
09/2010–07/2012	**Abitur**, *in the Wirtschaftsoberschule at the KS-Künzelsau*, *Very Good*
09/2008–07/2010	**Advanced Technical College Entrance Qualification in Business Informatics**, *GvSS Heilbronn*

Work Experience

Academic

11/2020–Now **Main Instructor**, *LMU Munich Center for Mathematical Philosophy & Department of Statistics*, München

Tasks: Design of course content (lectures, exercises, etc.), teaching, supervision of student projects and contact partner for student matters.

Explainable Artificial Intelligence, *MCMP & Statistics Department*, Jointly with Gunnar König, Winter Term 21/22

Causality and Machine Learning, *Statistics Department*, Jointly with Gunnar König and Susanne Dandl, Sommer Term 21

Philosophy of Artificial Intelligence, *MCMP*, Jointly with Prof. Stephan Hartmann, Winter Term 20/21

Ethics of Artificial Intelligence, *Statistics Department*, Jointly with Florian Pfisterer, Christoph Molnar, Gunnar König, and Susanne Dandl, Winter Term 20/21

10/2016–11/2020 **Teaching Assistant**, *LMU Munich Department of Mathematics & Munich Center for Mathematical Philosophy*, München

Tasks: Designing and correcting assignments/exams, giving tutorials, programming, contact partner for student matters.

Formal Methods II: Models and Simulations, *MCMP*, Led by Dr. Rush Stewart, Summer Term 20

Central Topics in Philosophy of Science, *LMU*, Led by Dr. Jürgen Landes, Winter Term 19/20

Linear Algebra 1, *Mathematics Department*, Led by Dr. Peter Philip, Winter Term 18/19

Linear Algebra 2, *Mathematics Department*, Led by Prof. Dr. Fabien Morel, Summer Term 18

Linear Algebra 1, *Mathematics Department*, Led by Prof. Dr. Fabien Morel, Winter Term 17/18

Topology and multivariable differential calculus, *Mathematics Department*, Led by Prof. Dr. Franz Merkl, Summer Term 17

Analysis 1, *Mathematics Department*, Led by Prof. Dr. Franz Merkl, Winter Term 16/17

Non-Academic

03/2019–09/2019 **Software Developer (working student)**, *Zentrum Digitalisierung.Bayern*, Garching,
Project: Working on the national research project MEMAP which contributes to the German energy transition strategy. MEMAP (Multi-Energy Management and Aggregation-Platform) optimally matches the local electricity- and heat demand/production for districts
Tasks: My work focused mainly on the software development of the platform in the programming language Java. In particular, I had the following tasks:
- programming the OPC-UA interfaces for handling live-data
- developing a Jetty-websocket and a website for online access to the platform (HTML, Javascript,etc.)
- configuration of server data for providing optimization results

Scholarships & Prizes

10/2019–09/2022 **Graduate School of Systemic Neuroscience Neurophilosophy Stipend**, *Ph.D. research stipend*

25/07/2019 **Mobility Innovation Competition @ Campus**, *3rd prize in Startup competition*, Team: DeepGuardian
Deep-learning-software equipped camera board for violence detection that respects data privacy.

07/2018 **Oskar-Karl-Forster-Scholarship**, *book stipend*

06/2012 **School-Prize**, *best Abitur*

Conferences, Workshops, Talks, etc.

30/06/2022-01/07/2022 **Hannover-MCMP-Wuppertal Network Workshop: Philosophy of Science**, *University of Wuppertal*, Presentation on "Scientific Inference With Interpretable Machine Learning"

21/06/2022-24/06/2022 **ACM Conference on Fairness, Accountability, and Transparency (FAccT)**

13/06/2022 **Panelist at Science Summit of the Joint Research Centre of the European Commission**, *Topic: Science through the AI lens*

09/06/2022-10/06/2022 **LMU-Cambridge Strategic Partnership Workshop, Topic: "AI in Science: Foundations and Applications"**, *Presentation on "Scientific Inference With Interpretable Machine Learning"*

09/11/2021-12/11/2021 **Workshop: Philosophy of Science Meets Machine Learning**, *University of Tübingen*, Presentation on "To Explain and to Predict – Explanatory Machine Learning Models in Science"

24/07/2021 **ICML workshop, Algorithmic Recourse**, *Online Event*, Poster on A Causal Perspective on Meaningful and Robust Algorithmic Recourse

19/05/2021 **MCMP-colloquium talks, Embrace the Complexity: The Paradigm Shift in Science From Statistics to Machine Learning**, *München, Germany (Online Event)*, Jointly with Christoph Molnar

12/04/2021-14/04/2021	**NIAS-workshop, Explainable Medical AI: Ethics, Epistemology, and Formal Methods**, *Leiden, the Netherlands (Online Event)*
17/07/2020	**ICML workshop, XXAI: Extending Explainable AI Beyond Deep Models and Classifiers**, *Vienna, Austria (Online Event)*, Poster on Pitfalls to Avoid when Interpreting Machine Learning Models
29/06/2020-03/07/2020	**Summerschool: Regularization Methods for Machine Learning**, *Genova, Italy (Online Event)*, Led by Prof. Lorenzo Rosasco
17/02/2020 - 18/02/2020	**Workshop on Machine Learning: Prediction Without Explanation?**, *Karlsruhe (KIT)*, Talk on Counterfactual Explanations & Adversarial Examples
14/01/2020	**Guest Lecture in CTPS course**, *MCMP*, Topic: The Wisdom of Crowds
27/07/2018 - 28/07/2018	**Workshop on Decision Theory & the Future of Artificial Intelligence**, *München (Jointly organized by the MCMP, the CFI, and the CSER)*
22/06/2017 - 26/05/2017	**Masterclass with Graham Priest on Paraconsistent Logic**, *München (LMU)*

Academic Service and Organization

Reviewing	**Synthese, ICML, ACM FAccT, Minds and Machines**
Workshop Co-Organizer	**LMU-Cambridge Strategic Partnership, Topic: "AI in Science: Foundations and Applications", 9-10 June 2022, Munich**
Reading Group Organizer	**MCMP, Topic: "Philosophy of Machine Learning", jointly with Tom Sterkenburg, summer term 2022, Munich**

Skills

Languages	German (native speaker), English (fluent), Spanish (very good command), Norwegian (good command).
Computer Skills	MATLAB/Octave (++), Python (+), Java (++), NetLogo (+++), JavaScript (++), HTML (++), PHP (+), WebPPL (+), LaTeX(+++), SQL (++).

List of Publications

Peer-reviewed articles

Freiesleben, T. (2022). The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds & Machines*, 32, 77-109, doi: https://doi.org/10.1007/s11023-021-09580-9

König, G., **Freiesleben, T.** and Grosse-Wendrup, M. (forthcoming 2023). Improvement-focused Causal Recourse (ICR). *Proceedings of the AAAI Conference on Artificial Intelligence 2023*.

Molnar, C., König, G., Herbinger, J., **Freiesleben, T.**, Dandl, S., Scholbeck, C., Casalicchio, C., Grosse-Wentrup, M. and Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. *In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science, vol 13200. Springer*, 39–68. doi: https://doi.org/10.1007/978-3-031-04083-2_4.

Unpublished articles (under review)

Freiesleben, T., König, G., Molnar, C. and Tejero-Cantero, A. (unpublished). Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena. *Under review at 'The British Journal for Philosophy of Science'*.

Freiesleben, T., Molnar, C., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., Wright, M. and Bischl, B. (unpublished). Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. *Under review at 'Machine Learning', status: major revision, this is the revised version sent to the journal.*

Acknowledgements

"Knowledge can never replace friendship."

— Patrick Star

Four years ago, I was very hesitant whether it was a good idea to pursue a PhD; writing my master's thesis was a stressful and lonely endeavor, and I feared a PhD would mean the same thing for several more years. So after my master's, I decided to take a year off to figure out if a PhD should be part of my life plan, and if so, to look for an interesting PhD program and a topic I would like to work on. I selected three PhD programs, and decided to leave academia if I didn't make it into one of the programs.

Today I can say that I really enjoyed my PhD and feel very lucky to have had the opportunity to join the GSN and pursue a topic that I am really passionate about. Sure there have been ups (teaching or workshops) and downs (COVID and paper rejections) but all in all I had a very good time, learned a lot, and had the freedom to follow my interests. The fact that my PhD was so much fun is mainly thanks to a number of wonderful people who supported me and whom I would like to thank at this point.

First and foremost, I am extremely grateful to my main supervisor Stephan Hartmann. Stephan has been an inspiring mentor who had always time when I needed help, supported me already in my application at the GSN and gave me immensely useful academic guidance. He always had me in mind when it came to cool projects such as teaching courses, organizing workshops, grant writing, scientific collaboration, and any other aspect of academic life. The MCMP in general has provided me with an ideal and probably unique environment for conducting research across the faculty lines of philosophy of science, cognitive science, computer science, statistics, and social sciences. I feel very fortunate and proud to be part of the MCMP family.

Mil gracias go to my second supervisor, Alvaro Tejero-Cantero, for his immense support. In particular, I would like to thank him for taking the time

to have extensive and detailed discussions about the content of our collaboration, for the occasional entertaining discussion side trips on related philosophical problems, and for the helpful resources he pointed me to. I would also like to thank my third and fourth supervisors, Paul Taylor and Agnieszka Wykowska, for their support, academic guidance, and especially for opening up different perspectives on my topic. In addition, I am very grateful to the GSN for providing me with a welcoming interdisciplinary community, interesting courses outside my area of expertise, a clear framework for my PhD, and financial support that gave me the freedom to focus on my research.

Tons of thanks go to the interpretable ML group led by Giuseppe Casalicchio from the statistics department at the chair of Professor Bernd Bischl. I felt super welcome in the group and learned so much from all the group members. I would especially like to thank my friend Gunnar König, with whom I collaborated the most, for all the long and super insightful discussions, shared teaching experiments and Mensa dates. Great thanks also go to Christoph Molnar for the fun collaborations, the Schrebergarten barbecues, and our provocative joint presentation at the MCMP, Giuseppe Casalicchio for all the nice weekly discussions, the organization and the opportunity to join the group, and Susanne Dandl, Christian Scholbeck, and Julia Herbinger for joint teaching events, insightful discussions, and the fun meetings.

Special thanks go also to my fellow PhD students, colleagues, and friends at the LMU, Gasper, Johannes, Tom, Rush, Sander, Louis, Rolf, Silvia, Naftali, Conrad and Maria for all the conversations and Mensa dates. Great thanks also to the GSN neurophilosophy group and particularly Stephan Sellmaier, who always had a sympathetic ear and good advises.

Warm thanks go also to all my friends outside of the academic realm, without whom my life would be so much poorer: Gabriel, Marina, Dennis, Christopher, Timur, Jörg, Adrian, Markus, Lukasz, Susi, Tobi, Anja, Bene, Alex, Elio, Julia, Thomas, Franzi, Klavdija, Antonia, Patrick, David, Klara, Sophia, Nakul, Eelco, and all the others that I forgot mentioning. Thanks for abiding all my lengthy blabla about my PhD, distracting me with wonderful discussions about all and sundry, and joining me for wonderful events (hiking, games night, Isar walks, soccer, cooking, eating, and all other activities of this endless list).

Zuletzt möchte ich ein riesiges Dankeschön an meine gesamte Familie sagen. Ihr wart immer da und habt mich durch mein ganzes bisheriges Leben unterstützt und mir so die Gelegenheit gegeben weiter zu Lernen und zu Wachsen. Ganz besonders geht dieser Dank an meine Eltern Regine und Bernd, und an meine drei Brüder, Sven, Andy und Jens.

Eidesstattliche Versicherung / Affidavit

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation "What Does Explainable AI Explain?" selbstständig angefertigt habe, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

I hereby confirm that the dissertation "What Does Explainable AI Explain?" is the result of my own work and that I have only used sources or materials listed and specified in the dissertation.

München, den 06.12.2022
Munich, date 06.12.2022

.....
Timo Freiesleben

Author Contributions

Freiesleben, T., König, G., Molnar, C. and Tejero-Cantero, A. (unpublished). Scientific Inference With Interpretable Machine Learning: Analyzing Models to Learn About Real-World Phenomena. *Under review at 'The British Journal for Philosophy of Science'.*

T.F. wrote large parts of the paper and developed the initial idea. A.TC., G.K., and C.M. added valuable new ideas, proofread and helped revise the paper. A.TC. helped in the design of Figures 1,3,4, and 7. G.K. wrote large parts of the section on causal learning and A.TC. contributed a paragraph on mechanistic models.

Freiesleben, T., Molnar, C., König, G., Herbinger, J., Reisinger, T., Casalicchio, G., Wright, M. and Bischl, B. (unpublished). Relating the Partial Dependence Plot and Permutation Feature Importance to the Data Generating Process. *Under review at 'Machine Learning', status: major revision, this is the revised version sent to the journal.*

T.F., C.M, and G.K. contributed equally to this work. C.M. developed the initial idea and wrote the initial draft. **T.F.** generalized the initial definitions and theorems to arbitrary sampling procedures and contributed the corresponding paragraphs concerning the sampling and extrapolation problem in Section 2 and Section 5. Moreover, **T.F.** and G.K. made conceptual contributions (particularly on Section 2), contributed some of the proofs (particularly Theorem 3), restructured the manuscript, and contributed the motivating example. Also, **T.F.** carefully revised and proofread all proofs and formal definitions. G.K. contributed the application section. C.M, M.W., J.H., and T.R. implemented and run the simulation study. **All authors** added valuable new discussion points and helped revise the text.

Freiesleben, T. (2022). The Intriguing Relation Between Counterfactual Explanations and Adversarial Examples. *Minds & Machines*, 32, 77-109, doi: https://doi.org/10.1007/s11023-021-09580-9

T.F. is the only author.

König, G., **Freiesleben, T.** and Grosse-Wendrup, M. (forthcoming 2023). Improvement-focused Causal Recourse (ICR). *Proceedings of the AAAI Conference on Artificial Intelligence 2023*.

G.K. had the initial idea, G.K. and **T.F.** developed the story and the philosophical foundation together. G.K. wrote large parts of the paper, developed the proofs and wrote the code. **T.F.** came up with the running example, lead writing Section 4, checked the proofs and contributed to Sections 1, 2, 9 and 10. **All authors** helped to revise and proofread the paper.

Molnar, C., König, G., Herbinger, J., **Freiesleben, T.**, Dandl, S., Scholbeck, C., Casalicchio, C., Grosse-Wentrup, M. and Bischl, B. (2022). General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models. *In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, KR., Samek, W. (eds) xxAI - Beyond Explainable AI. xxAI 2020. Lecture Notes in Computer Science, vol 13200. Springer*, 39–68. doi: https://doi.org/10.1007/978-3-031-04083-2_4.

C.M. initiated and coordinated the project. **T.F.** wrote Section 5.2 and partially wrote Section 4. and 11. **T.F.** and S.D. contributed paragraphs within each pitfall on local IML methods. **All authors** proofread and revised the paper. The other pitfalls were written by the co-authors.

.....	.....
Timo Freiesleben	Prof. Dr. Stephan Hartmann
.....	.....
Christoph Molnar	Gunnar König

Munich, 6 December 2022