Explainable Machine Learning and its Limitations

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
M.Sc. Sebastian Bordt
aus Gehrden

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation	27.9.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Ulrike von Luxburg
2. Berichterstatter:	Prof. Dr. Matthias Hein

Abstract

In the last decade, machine learning evolved from a sub-field of computer science into one of the most impactful scientific disciplines of our time. While this has brought impressive scientific advances, there are now increasing concerns about the applications of artificial intelligence systems in societal contexts. Many concerns are rooted in the fact that machine learning models can be incredibly opaque. To overcome this problem, the nascent field of explainable machine learning attempts to provide human-understandable explanations for the behavior of complex models. After an initial period of method development and excitement, researchers in this field have now recognized the many difficulties inherent in faithfully explaining complex models. In this thesis, we review the developments within the first decade of explainable machine learning. We outline the main motivations for explainable machine learning, as well as some of the debates within the field. We also make three specific contributions that attempt to clarify what is and is not possible when explaining complex models. The first part of the thesis studies the learning dynamics of the human-machine decision making problem. We show how this learning problem is different from other forms of collaborative decision making, and derive conditions under which it can be efficiently solved. We also clarify the role of algorithmic explanations in this setup. In the second part of the thesis, we study the suitability of local post-hoc explanation algorithms in societal contexts. Focusing on the draft EU Artificial Intelligence Act, we argue that these methods are unable to fulfill the transparency objectives that are inherent in the law. Our results also suggest that regulating artificial intelligence systems implicitly via their explanations is unlikely to succeed with currently available methods. In the third part of the thesis, we provide a detailed mathematical analysis of Shapley Values, a prominent model explanation technique, and show how it is connected with Generalized Additive Models, a popular class of interpretable models. The last part of the thesis serves as an interesting case study of a connection between a post-hoc method and a class of interpretable models.

Zusammenfassung

In der letzten Dekade hat sich das Maschinelles Lernen von einem Teilbereich der Informatik zu einer der bedeutendsten wissenschaftlichen Disziplinen unserer Zeit entwickelt. Obwohl dies beeindruckende wissenschaftliche Fortschritte mit sich gebracht hat, gibt es nun zunehmende Bedenken hinsichtlich des Einsatzes von Künstlicher Intelligenz in gesellschaftlichen Kontexten. Ein Hauptproblem besteht dabei in der Intransparenz der komplexen gelernten Funktionen. Um dieses Problem zu lösen versucht das Feld des Erklärbaren Maschinellen Lernens allgemeinverständliche Erklärungen für das Verhalten komplexer Modelle zu finden. Nach einer anfänglichen Phase der Entwicklung vieler verschiedener Erkläralgorithmen haben Wissenschaftler nun die vielen Schwierigkeiten erkannt, die mit dem Erklären komplexer Modelle verbunden sind. In dieser Arbeit betrachten wir die Entwicklungen im ersten Jahrzehnt des Erklärbaren Maschinellen Lernens. Wir skizzieren die Hauptmotive für Erklärbares Maschinelles Lernen, und einige der wichtigsten Debatten in diesem Bereich. Weiter leisten wir drei Beiträge zur Erkärbarkeit komplexer Modelle. Im ersten Teil der Arbeit untersuchen wir die Lerndynamik zwischen Mensch und Maschine. Wir zeigen, wie sich dieses Lernproblem von anderen Formen der gemeinsamen Entscheidungsfindung unterscheidet und leiten Bedingungen ab, unter denen effizientes Lernen möglich ist. Außerdem klären wir die Rolle algorithmischer Erklärungen im Lernprozess. Im zweiten Teil der Arbeit untersuchen wir die Eignung einer prominenten Klasse von Methoden - sogenannter local post-hoc Erkläralgorithmen - in gesellschaftlichen Kontexten. In Bezug auf den draft EU Artificial Intelligence Act argumentieren wir, dass die Methoden nicht in der Lage sind, die in dem Gesetz verankerten Transparenzkriterien zu erfüllen. Der dritte Teil der Arbeit enthält eine detaillierte mathematische Analyse von Shapley-Werten, einer prominenten Technik der Modellerklärung. Wir zeigen, wie Shapley-Werte mit Generalized Additive Models, einer prominenten Klasse interpretierbarer Modelle, zusammenhängen. Der letzte Teil der Arbeit dient als interessante Fallstudie für die Zusammenhänge zwischen Erklärmethoden und interpretierbaren Modellen.

Acknowledgements

I would like to thank Ulrike von Luxburg for her support and academic guidance. Without her mentorship, this thesis would not have been possible.

I would also like to thank all the different people with whom I have been collaborating over the course of my PhD. I have learned a lot from all of you: Leena C. Vankadara, Debarghya Ghoshdastidar, Michèle Finck, Eric Raidl, Uddeshya Upadhyay, Zeynep Akata, Suraj Srinivas, Himabindu Lakkaraju, Michal Moshkovitz and Gyorgy Turan.

The members of the tml-group have made my time in Tübingen enjoyable, even during Covid: Michael Lohaus, Solveig Klepper, Luca Rendsburg, Moritz Haas, Sascha Meyen, David Künstle, Damien Garreau and Michaël Perrot. I hope we are going to enjoy many more summer evenings on top of Österberg. Also thank you Patrizia Balloch for helping with the many administrative issues that arose over the course of my PhD.

I had a great time at the Summer Cluster on Interpretable Machine Learning at the Simon's Institute for the Theory of Computing in Berkeley. I would especially like to thank Rich Caruana, Shai Ben-David, Gyuri Turan, Zachary Lipton, Simina Brânzei, Bin Yu, Tosca Lechner, Michal Moshkovitz and Sina Fazelpour for many interesting discussions that helped to sharpen my view of explainable machine learning. It was great to meet many of you in Chicago again last month.

Throughout my time in Tübigen, Michael and Lamy have been a constant support, most notably while fighting global pandemics. I would also like to thank Martin, Katie, Dominik, Konstantin, and Auguste for many enjoyable evenings in Tübingen.

While science is rooted in reality, it is the fantastic realms that keep up our imagination. Thank you Justus, Mareike, Kai, Timo and Aylin. Thank you also Ruben, Simon, Philipp and Lorenz.

Finally, I would like to thank all my family and friends who have supported me on this journey, and Suraj Srinivas for a very helpful discussion.

Contents

Chapter 1

Introduction

We begin the thesis by tracing the growing need for explainable machine learning (Section 1.1). We then discuss why the classic paradigm of evaluating machine learning models solely by their test accuracy is increasingly being seen as insufficient (Section 1.2). Starting with a famous paper by Matthew Zeiler and Rob Fergus, we then outline the development of explainable machine learning as a research area (Section 1.3). Based on this, we review a number of different motivations for explainable machine learning (Section 1.4) and debate the promises and limitations of the field (Section 1.5). Finally, we lay out the specific contributions of this thesis (Section 1.6).

1.1 Explainable Machines - Why now?

At the time of the writing of this thesis, public debates about the restriction and regulation of artificial intelligence are at an all-time high. The most immediate reason for this is breakthroughs in generative AI, exemplified by ChatGPT, a language model optimized for dialogue (OpenAI, 2023; Ouyang et al., 2022), and Stable Diffusion, a model that generates realistically-looking images (Rombach et al., 2022). While most observers agree that the technology itself is impressive - ChatGPT has changed our understanding of what is possible in natural language processing (Bubeck et al., 2023) - there are increasing concerns about its effects on society and democracy (Acemoglu and Johnson, 2023). Among others, concerns about the regulation of artificial intelligence have led to a number of hearings before the US Senate (Kang, 2023).

Taking a step back, generative AI is not the only form of machine learning to cause a stir in recent years. In 2016, for example, investigative journalists at ProPublica wrote a much-discussed article about the COMPASS program, a piece of software that assists judges in deciding cases of criminal bail in the United States (Angwin et al., 2016). The journalists claimed that the program was biased against blacks, a claim that has since been much discussed (Rudin et al., 2020).

A main concern about the usage of artificial intelligence systems in societal contexts is that these systems are incredibly opaque. This is problematic insofar as these systems are usually being deployed by institutions that are already relatively pow-

erful, leading to concerns that artificial intelligence might adversely affect the fundamental rights of individuals and the balance of power in society (Acemoglu and Johnson, 2023). Given that many machine learning systems have already been found to be discriminatory and amplify historical biases, these are not empty concerns but major hurdles towards the adoption of artificial intelligence (Barocas et al., 2019).

It is important to distinguish two different reasons for the opaqueness of currently existing systems. One reason is that companies often attempt to keep any details about their developed models a secret. For example, it is not known what exactly the inputs to the COMPASS program are. In their analysis, the journalists at ProPublica had to reconstruct the behavior of the underlying model from publicly available data about criminal records (Larson et al., 2016). The case of ChatGPT is similar. While the general principles of the underlying technology are well-known (Vaswani et al., 2017; Ziegler et al., 2019), any details about the training data, schedule, and various stages of fine-tuning with human feedback remain unknown to the public or general scientific community.

The most important reason for the opaqueness of today's machine learning models, however, is simply the current state of the technology. Modern machine learning models consist of millions of parameters. While they are mathematically well-defined functions, their inner workings are not directly transparent to any human. For this reason, these models are often referred to as black boxes.

Why do intransparent black box models dominate modern machine learning research? One reason for this can be found in the history of the field of machine learning. To abandon debates about model internals and data-generating processes, machine learners invented the test set paradigm. According to this research paradigm, a model is evaluated solely by its accuracy on held-out test data, ignoring any model internals and the way in which the model was obtained (Breiman, 2001b). As a consequence, machine learners developed increasingly large and complex models, without ever caring about model interpretability.

While this historical development is certainly important, many would argue that there are other, more fundamental reasons as to why current models are intransparent. In fact, Leo Breiman gave an argument for accurate but not interpretable models more than 20 years ago, in his famous article "Statistical modeling: The two cultures" (Breiman, 2001b). Breiman argues that

> *"[...] nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable."*

If we accept this argument, then it is obvious to expect that the behavior of an accurate model for natural phenomena must also be partially unknowable. In other words, it might be impossible to interpret the behavior of many accurate models in human-understandable terms. With Breiman, many would argue that the opaqueness of currently existing models is not a failure of the technology, but a necessary property of accurate models that arises directly from the properties of the problem that we attempt

to solve.[1] According to this argument, the black box properties of natural phenomena are also the reason why we need the methods of machine learning in the first place. If there were simple and interpretable solution strategies to problems such as image classification and complex gameplay, scientists would have solved these problems by explicit programming.

In this current moment of highly accurate black box models, the nascent field of explainable machine learning attempts to make the behavior of complex systems transparent by providing human-understandable "explanations". At a high level, this comes with two main motivations. One motivation is the societal need to gain transparency and control over complex machine learning systems, as outlined in the introductory paragraphs above. The other motivation is entirely scientific, and rooted in a relatively straightforward critique of Breiman: To what degree, exactly, are complex natural phenomena "partially unknowable"? Since we have found valid interpretations of complex phenomena like quantum mechanics, why can't we find interpretable yet accurate solutions to other complex phenomena?

As we are going to see in this thesis, the field of explainable machine learning has had notable successes, including, but not limited to, the building of relatively accurate interpretable models. As the field has matured over the last couple of years, however, it has also become clear that most of the explanation algorithms that have been proposed in the literature are much less potent than what has been suggested by their inventors. As such, at a fundamental level, most of the interesting research questions in explainable machine learning are still unanswered.

1.2 The Rise and Fall of Accuracy for Model Evaluation

In this section, we briefly review the shortcomings of evaluating models by their predictive accuracy on held-out test data. Since the idea of accuracy as the sole criterion of model performance is opposed to the idea of model interpretability (Breiman, 2001b), this sets the stage for our discussion of explainability research in the next section.

Historically, focusing on the predictive accuracy on held-out test data has been a key innovation of machine learning research that set the field apart from other disciplines. Under this research paradigm, sometimes called the common task framework (Donoho, 2017), the performance of a statistical model is judged solely by its predictive accuracy, ignoring any model internals and the way in which the model was obtained. Arguably, this approach allowed for a lot of innovation in model design and contributed to the success of machine learning as a scientific discipline.

Around 2012, machine learning models began to "solve" challenging predictive tasks in terms of their accuracy on held-out test data. This is exemplified by AlexNet (Krizhevsky et al., 2012) and increasing performance on the ImageNet Large Scale Visual Recognition Challenge (Deng et al., 2009). In 2014, the best-performing model on

[1] A modern version of this argument is given by Dziugaite et al. (2020).

this benchmark reached an accuracy roughly comparable with that of human annotators (Russakovsky et al., 2015).

Relatively quickly, machine learning researchers began to observe that models that had high accuracy on held-out test data could still dramatically fail to solve the real-world problems that they were meant to address. A particularly prominent failure case is documented by Buolamwini and Gebru (2018), who showed that commercial facial recognition software was significantly worse at detecting darker-skinned females than white-skinned males. In a certain sense, the reason for this was an avoidable failure of model building: The data sets that were available at the time almost exclusively contained images of white-skinned people. As a consequence, in this particular case, the problem could be addressed by collecting more representative and equitable data sets of human faces. At a fundamental level, however, the underlying problem of data set composition remains hard to address: Even very big data sets scarcely represent the real-world problems that we would actually like to solve. At the moment, this is again more than evident in discussions about the composition of the training data of large language models.

Apart from the difficulty of collecting a truly representative data set, another important limitation of the test set paradigm is that it relies, by construction, on the i.i.d. assumption. This means that we assume that the test examples follow *exactly* the same distribution as the training examples. The real world, however, is frequently characterized by distribution shifts. This umbrella term describes all the different ways in which the data that we collect about the same problem changes over time, such as when, for example, the invention of a new drug changes the relationship between blood pressure and health status.

In fact, there are even more failure cases of the test accuracy, such as adversarial examples. What is important for us, however, is only that the failures of test accuracy to reliably translate into desirable real-world behavior have led machine learning researchers to develop novel criteria by which model behavior can be judged. And while explainable machine learning is by far not the only candidate - many researchers would invoke notions of model robustness, for example - more interpretable models remain an important contender to achieve more desirable real-world behavior.

1.3 Explainable Machine Learning: An Overview

In this section, we give a brief overview of recent developments within the field of explainable machine learning. We mostly abstain from enumerating all the different methods and focus on the main intellectual developments within the field. As is the case with many topics in machine learning, aspects of the questions that are studied in explainable machine learning are also being considered in earlier literature on statistics. Concrete examples are attempts to measure the importance of different variables (Breiman, 2001a) and build interpretable models (Lin et al., 2020). In order to be concise, we restrict ourselves to the machine learning literature.

The modern literature on explainable machine learning started around ten years ago with two publications that attempted to make the inner workings of neural networks transparent. These publications are "Visualizing and Understanding Convolutional Networks" by Matthew Zeiler and Rob Fergus (Zeiler and Fergus, 2014) and "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps" by Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman (Simonyan et al., 2013). In the first paper, the authors analyzed a convolutional neural network similar to the AlexNet model that had been released the year before (Krizhevsky et al., 2012). Quoting from the abstract, their objective was to introduce

> *"a novel visualization technique that gives insight into [...] the operation of the classifier. [..] This enables us to find model architectures that outperform Krizhevsky et a. on the ImageNet classification benchmark."*

Today, understanding and improving models are still among the main motivations of explainable machine learning (compare Section 1.4.1 below). In the second paper, the authors proposed to compute the gradient of the class score of a convolutional neural network with respect to the input image (Simonyan et al., 2013). This technique is still the basis of many gradient-based interpretability methods. Interestingly, both papers do not yet use the words "explanation" or "explain". Instead, they employ the term "understanding".

Apart from these two papers, early contributions to the field of explainable machine learning have been made by Rich Caruana, Cynthia Rudin, and Been Kim (Christian, 2020, Chapter 3). Caruana and Rudin are known for their work on interpretable model building (Caruana et al., 2015; Rudin et al., 2022), a topic that we discuss in Section 1.5.3. Been Kim was among the first researchers who explicitly advocated for another main objective of explainable machine learning: Human-machine collaboration. In her 2015 PhD thesis, titled "Interactive and interpretable machine learning models for human machine collaboration" (Kim, 2015), Kim wrote

> *"I envision a system that enables successful collaborations between humans and machine learning models by harnessing the relative strength to accomplish what neither can do alone."*

The perspective of Kim is still very relevant today, since the debates about the degree to which machines should act autonomously, or be decision-support tools for humans, are still far from settled (Narasimhan et al., 2022).

After a period of increasing interest in the topic (Bach et al., 2015), around 2016/2017 the field of explainable machine learning began to take shape. Among others, this is evidenced by two very influential papers on post-hoc feature attribution methods. The first paper, by Marco Ribeiro, titled "'Why Should I Trust You?': Explaining the Predictions of Any Classifier", proposed to explain any function by learning an interpretable model locally around the prediction (Ribeiro et al., 2016). The second paper, by Scott Lundberg, titled "A unified approach to interpreting model predictions" proposed the SHAP method, a game-theoretic approach to feature attribution that is still

heavily used today (Lundberg and Lee, 2017). Both papers stand out in two important ways. For one, they explicitly use the terminology to "explain" functions. While this is not the first use of this terminology, these two papers exemplify that it became commonly used around the time. What is more, both papers stand out for their exceedingly broad claims. Specifically, the papers claimed that the respective methods could explain "any function", and directly fulfill challenging interpretative goals like trust.

The year 2017 also saw the publication of "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" (Selvaraju et al., 2017) and "Axiomatic attribution for deep networks" (Sundararajan et al., 2017). These works are similar in spirit to Simonyan et al. (2013), using gradients and backpropagation to explain the predictions of deep neural networks. Unlike LIME and SHAP, which do not make use of model internals and only rely on querying access, these two papers make use of the specific architecture of deep neural networks.

The growing interest in explainable machine learning around 2016/2017 is also evidenced by a number of panels and workshops at the NeurIPS conference, the premier conference in machine learning. In 2016, there was a NeurIPS panel on "Explainable AI". In 2017, there were workshops on "Interpreting, Explaining and Visualizing Deep Learning - Now what?" and "Transparent and interpretable Machine Learning in Safety Critical Environments". Finally, in 2020, there was a tutorial by Himabindu Lakkaraju, Julius Adebayo, and Sameer Singh on "Explaining ML Predictions: State-of-the-art, Challenges, and Opportunities".

After 2017, the papers on LIME, SHAP, Grad-CAM, and Integrated Gradients gave rise to an entire literature of papers that attempted to improve the respective methods. Prominent examples from this literature are "Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks" (Chattopadhay et al., 2018), and "Feature relevance quantification in explainable AI: A causal problem" (Janzing et al., 2020).

The increasing popularity of explainable machine learning around this time also gave rise to attempts to conceptualize the field (Doshi-Velez and Kim, 2017), as well as an increasing amount of scepticism as to what the different proposed methods were really doing (Lipton, 2018). It also gave rise to books and surveys that reviewed the different proposed methods (Molnar, 2020; Rudin et al., 2022; Samek et al., 2021).

The extremely broad claims in papers on LIME and SHAP, as well as many other early contributions in the field of explainable machine learning, illustrate an early optimism to "solve" the problem of model understanding. As the field gained traction and popularity, however, researchers increasingly started to question the inner workings of the proposed methods, as well as some of the overly broad claims in early papers. In 2018, Julius Adebayo and co-authors published the paper "Sanity checks for saliency maps" (Adebayo et al., 2018). This widely recognized paper empirically tested a number of different saliency methods for how sensitive they were to the underlying model. The paper found that many saliency maps were not sensitive to the weights of the last layer of a neural network, meaning that they were insensitive to the ultimate predictions. This famous paper is the first contribution in a line of work that

started to empirically probe the properties of the different proposed explanation algorithms. Other prominent papers from this literature include the Remove-And-Retrain (ROAR) benchmark (Hooker et al., 2019), as well as the follow-up work by Julius Adebayo "Debugging tests for model explanations" (Adebayo et al., 2020). As researchers began to critically analyze the different proposed methods, it also became apparent that almost all of them were subject to adversarial attacks (Dombrowski et al., 2019; Slack et al., 2020). In addition to this empirical work, researchers also began to study the properties of different explanation algorithms theoretically. A notable early contribution to this literature is the analysis of LIME by Damien Garreau (Garreau and von Luxburg, 2020).

In addition to a critical analysis of explanation algorithms, researchers have also attempted to move beyond the paradigm of feature attributions, which dominated much of the first decade of interpretability research (Section 1.5.2). Early works that proposed different notions of explanations are "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR" (Wachter et al., 2017), "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)" (Kim et al., 2018), and "Anchors: High-Precision Model-Agnostic Explanations " (Ribeiro et al., 2018). These works introduced counterfactual explanations, concepts, and certificates, all of which are now widely accepted notions of model explanations.

1.4 Objectives of Explainable Machine Learning

In addition to a large variety of different methods and algorithms, explainable machine learning is also characterized by a large variety of different motivations. What is the purpose of explaining a model? Sometimes the motivations are very explicit, as in the title of the LIME paper (Ribeiro et al., 2016). At other times the motives behind explanations are more implicit, given via benchmarks and example applications of particular methods (Lundberg and Lee, 2017). The large variety of methods and motivations makes it surprisingly hard to assess the field of explainable machine learning. In fact, this is not a new phenomenon. In 2016, Zachary Lipton observed in "The Mythos of Model Interpretability" (Lipton, 2018) that

> *"[...] the task of interpretation appears underspecified. Papers provide diverse and sometimes non-overlapping motivations for interpretability, and offer myriad notions of what attributes render models interpretable."*

Similarly, researchers Finale Doshi-Velez and Been Kim observed in 2017 that (Doshi-Velez and Kim, 2017)

> *"[..] there is little consensus on what interpretability in machine learning is and how to evaluate it [..]"*

To provide a basis for our discussion in Section 1.5, this section gives a short overview of the most common motivations for explainable machine learning.

1.4.1 Model Debugging, Understanding and Improvement

Model understanding and improvement are the main objectives in some of the earliest works on explainable machine learning (Simonyan et al., 2013; Zeiler and Fergus, 2014). From the perspective of most applied machine learning researchers, they remain among its most important objectives today. An example of this is given by Julius Adebayo's 2022 PhD thesis, titled "Towards Effective Tools for Debugging Machine Learning Models" (Adebayo, 2022). In comparison with other objectives of explainable machine learning, model debugging and improvement are relatively modest goals. In practice, a model developer might use a variety of tools to debug a model and choose the most appropriate ones on a case-by-case basis. In this scenario, it can already be a success case if at least some of the methods are useful at least some of the time. Wrong or unhelpful explanations don't hurt much. In the worst case, they might lead to a failed attempt to improve a model. A concrete sucess case of explainable machine learning for model debugging is the ability of feature attribution methods to highlight localized spurious features in images (Adebayo et al., 2020).

Model understanding in its general form is both adjacent and alternative to explainable machine learning. An example of this is given by the recent work of Been Kim on the acquisition of chess knowledge in AlphaZero (McGrath et al., 2022). The authors write that

> *"[..] the system [..] appears to learn concepts analogous to those used by human chess players. [..] probes applied to AlphaZero's internal state enable us to quantify when and where such concepts are represented in the network."*

The main difference between explainable machine learning, as it is commonly perceived, and model understanding as exemplified by the above quote is that understanding must not necessarily come in the form of "explanations". Another concrete example, timely during the writing of this thesis, is the attempts to understand the behavior and capabilities of large language models like GPT-4 (Bubeck et al., 2023).

1.4.2 Human-AI Collaboration

The idea to develop computer programs that assist human decision makers - we can think of the somewhat stereotypical example of a doctor who is assisted by a computer program - goes back at least to the development of expert systems in the 1970s (Russell, 2010). As we have seen in Section 1.3, improving human-AI collaboration is also one of the initial goals of explainable machine learning. One of the main reasons for this objective is that human domain experts are often dissatisfied with opaque black box models. Instead, they would like to gain insights into how a computer program arrived at a decision and incorporate these insights into their own reasoning.

In terms of the proposed methods, somewhat confusingly, human-AI collaboration has often seen the same tools as model understanding and debugging (for example, saliency maps for images). The important difference, however, is that the human who receives the explanations is not a machine learning engineer who wants to debug a

model, but a domain expert who is going to make a potentially consequential decision. As we are going to discuss in more detail in Section 1.5.2, the goal of improving human-AI collaboration is much more demanding than model debugging and improvement.

Similarly to model understanding, human-AI collaboration in its general form exceeds the field of explainable machine learning. A concrete example is a recent line of work on "learning to defer" which studies the learning problem when a computer program can decide to make a decision on its own or defer to a human expert (Narasimhan et al., 2022).

1.4.3 Transparency, Regulation and Alignment

As machine learning has increasingly impacted the world, policymakers have begun to call for the regulation and oversight of the usage of machine learning systems (Kang, 2023). As we have already outlined in Section 1.1, a main concern is that the opaqueness of machine learning models might adversely affect the balance of power in society (Acemoglu and Johnson, 2023). Another concern, shared by some but not by all machine learning researchers, is that models might become increasingly intelligent while not being aligned with human preferences.

Perhaps unsurprisingly, explainable machine learning has been seen as a potential candidate to provide transparency and oversight of complicated machine learning models. At a very high level, the intuitive idea is that intelligent machines might justify and explain their decisions in a way that is similar to human experts. More concretely, it is sometimes believed that explanations might constrain the adverse behavior of a system, in the sense that a system that is discriminatory would necessarily have to reveal this in its explanations. Unfortunately, as we are going to argue extensively in the second part of the thesis (Section 1.6.2), such beliefs are largely misplaced, at least given currently existing explanation algorithms.

1.5 Debates and Limitations

Ten years after Zeiler and Fergus published their paper on visualizing and understanding neural networks, the field of explainable machine learning has seen early excitement, considerable growth, and increased skepticism. In this section, we give a selective overview of current debates within the field of explainable machine learning.

1.5.1 Explaining Any Function

An important point where the debate within the field of explainable machine learning has progressed from 2017/2018 is that the problem of explaining arbitrary black box functions is increasingly being seen as ill-posed. In other words, explaining requires assumptions.

We give three main arguments why this is the case. The first two are empirical observations, and the third is of theoretical nature. The first empirical observation

is that for almost all of the methods in explainable machine learning, there are some scenarios where they "work well", and others where they don't. Papers that introduce new methods usually contain applications to showcase the good performance of these methods. Subsequent research then identifies failure modes where the proposed method is not able to hold up to its initial promises. The generality of the cases where the method works is then subject to intense debate. As a concrete example, consider the ability of feature attribution methods to highlight spurious correlations. The LIME paper contains a famous application of spurious correlation detection with dogs, wolves, and snow (Ribeiro et al., 2016). Later research by Julius Adebayo has tried to identify the conditions under which feature attribution methods can identify spurious correlation and found that there are many scenarios where they cannot (Adebayo, 2022; Adebayo et al., 2020). Still, many researchers believe, probably justifiably so, that feature attributions are a useful tool to detect at least certain forms of spurious correlations.

The second empirical observation is that explanation algorithms are generally subject to adversarial attacks. By an adversarial attack, we mean a modification to the function f that we want to explain that keeps the predictions of the function on the data distribution constant but arbitrarily modifies the explanations. For LIME and SHAP, this has been demonstrated in (Slack et al., 2020). For gradient-based explanation methods, a general argument is contained in (Dombrowski et al., 2019).

The third argument for why explaining any function is not possible is of theoretical nature. In the literature on the theory of machine learning, it is a well-known fact that there exists no universal learning algorithm (Wolpert, 1996). In other words, there exist so many possible functional relationships that every learning algorithm must invariably choose a preference for some functions over others. Given this fundamental impossibility result for learning functions, it seems highly unlikely that there exists a universal explainer for explaining functions. Interestingly, however, the literature on explainable machine learning has so far not seen any impossibility results that are as convincing as those for learning.

To summarize, while the field of explainable machine learning has seen many papers that propose to explain arbitrary functions, researchers now believe that this is generally not possible. Instead, we must identify the particular conditions and assumptions under which explainability is possible. In most interesting cases, the specific form and practical relevance that these assumptions will take are yet to be discovered.

1.5.2 The Utility of Feature Attributions

In the last decade, the most prominent type of explanation has been the feature attribution. Feature attributions comprise scalar attributions for tabular data, saliency maps for images, and highlighted words in sentences. Many different explanation algorithms ultimately provide feature attributions, including LIME, SHAP, Grad-CAM, and Integrated Gradients (Lundberg and Lee, 2017; Ribeiro et al., 2016; Selvaraju et al., 2017; Sundararajan et al., 2017).

Perhaps due to their popularity, there has been a lot of debate about the utility of feature attributions. This debate can be confusing because all of the different objectives of explainable machine learning have at one point or another been associated with feature attributions (Section 1.4). A literature that begins with Zeiler and Fergus (2014) uses feature attributions and visualizations in order to debug, understand and improve models. This includes spurious correlation detection, already discussed above (Adebayo et al., 2020). However, Scott Lundberg, the inventor of SHAP, has suggested that his method might also be useful for doctors who want to understand the predictions of black box models (Lundberg et al., 2020). In addition, there are many who have suggested that saliency methods might be useful for doctors in medical imaging tasks, leading to studies like "Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy" (Sayres et al., 2019). Also with regard to transparency and regulation, many have suggested that, for example, the fairness of classifiers might be assessed using feature attribution methods.

In the domain of model debugging and improvement, feature attributions have arguably been a success case. A prime example of this is their ability to detect localized forms of spurious correlation (Adebayo, 2022). However, as we have discussed in Section 1.4.1, model debugging and improvement are also the most modest goal of explainable machine learning. A more challenging domain where feature attribution methods seem to be useful, at least in certain cases, is scientific discovery. For example, Janizek et al. (2023) report that feature attributions can accelerate data-driven cancer pharmacology.

With regard to the objective of human-AI collaboration, feature attribution methods, unfortunately, do not seem to bring any real benefit. For example, there does not seem to exist a single high-quality study that unambiguously demonstrates the benefits of saliency maps on doctors' decision-making. In one of the few existing high-quality studies on the topic of human machine decision-making in the medical domain, "Human–computer collaboration for skin cancer recognition" (Tschandl et al., 2020), the authors did not even try any saliency methods but instead resorted to prototype-based explanations. In the end, the main result of this study was that the relatively simple mechanism of delivering multi-class probabilities allowed doctors to make the best diagnoses. Similarly, Sayres et al. (2019), who explicitly tested the efficacy of saliency maps to assist doctors in grading images for diabetic retinopathy, conclude that

> *"For most cases, Grades + Masks was as only effective as Grades Only [...]"*

implying that feature attributions did not bring any real benefit on the considered task.

With regard to transparency and regulation, it is important to note that there have already been many discoveries of biases in algorithms. This includes the gender shades study discussed in Chapter 1.2, as well as many other examples collected in Barocas et al. (2019, Chapter 1). However, none of these discoveries seem to have come through feature attributions methods, or explainable machine learning more generally. Instead, the literature on algorithmic bias develops clever context-specific tests.

In summary, feature attributions have turned out to be useful for model debugging, understanding and improvement. While they have also been heralded for more complex tasks, there does not seem to exist any empirical evidence in order to support these claims.

1.5.3 The Success of Interpretable Models

One of the biggest success cases of explainable machine learning are inherently interpretable models. In contrast to the approach of first building a model, then interpreting it, this approach aims to build models that are interpretable from the start. The two most important classes of interpretable models are small decision trees and Generalized Additive Models (GAMs) with few interactions (Caruana et al., 2015; Rudin et al., 2022). The most important result of recent years is that these models are often able to reach competitive accuracy on tabular data. Interpretable models have also shown promising capabilities at detecting defects in data sets (Lengerich et al., 2022). This makes them promising candidates both for model debugging and improvement, as well as for human-AI collaboration in domains such as health care. Some scholars have even argued that post-hoc explanation algorithms should not be used for tabular data at all (Rudin, 2019).

Unfortunately, interpretable models also face an important limitation. Despite a number of recent attempts at this problem (Chen et al., 2019), they do not work well for image and text data. Thus, while interpretable models are an incredibly useful tool, they are unlikely to replace black box models in many important domains.

1.6 Thesis Contributions

This thesis is based on the following publications.

Bordt, S. and von Luxburg, U. (2022) A Bandit Model for Human-Machine Decision Making with Private Information and Opacity. *In International Conference on Artificial Intelligence and Statistics (AISTATS)*. https://proceedings.mlr.press/v151/bordt22a.html

Bordt, S., Finck, M., Raidl, E., von Luxburg, U., (2022) Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts. *In FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency*. https://dl.acm.org/doi/abs/10.1145/3531146.3533153

Bordt, S. and von Luxburg, U. (2023) From Shapley Values to Generalized Additive Models and back. *In International Conference on Artificial Intelligence and Statistics (AISTATS)*. https://proceedings.mlr.press/v206/bordt23a.html

I also published the following workshop paper.

Bordt, S., Upadhyay, U., Akata, Z., von Luxburg, U. (2023) The Manifold Hypothesis for Gradient-Based Explanations. *In Explainable AI for Computer Vision Workshop, Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2206.07387

I also co-authored the following paper on kernel clustering which was accepted for an oral presentation at AISTATS.

Vankadara, L.C., Bordt, S., von Luxburg, U., Ghoshdastidar, D. (2022) Recovery Guarantees for Kernel-based Clustering under Non-parametric Mixture Models. *In International Conference on Artificial Intelligence and Statistics (AISTATS)*. https://proceedings.mlr.press/v130/vankadara21a.html

The following paper is currently available as a pre-print on arxiv.

Bordt, S. and von Luxburg, U. (2023) ChatGPT Participates in a Computer Science Exam *arXiv preprint*. https://arxiv.org/abs/2303.09461

We now outline the contributions of the different parts of the thesis and how they relate to the research questions within the field of explainable machine learning.

1.6.1 A Bandit Model for Human-Machine Decision Making with Private Information and Opacity

In the first part of the thesis, we consider the learning dynamics of the human-machine decision making problem. Improving human-AI collaboration is one of the main objectives of explainable machine learning (Section 1.4.2). Nevertheless, the learning dynamics of this problem - How can a human and a machine jointly learn to solve difficult problems? - has received little explicit treatment. We introduce a formal framework to study the learning dynamics of the human-machine decision making problem. Our main objective is to clarify the way in which human-machine decision making is different from other forms of collaborative decision making problems. We study the role of two salient aspects that set human-machine decision making apart:

(1) the presence of private information that is accessible only to one of the two decision makers, and (2) opacity, that is an imperfect understanding between the two decision makers.

Formally, our framework is a two-player variant of the contextual bandit model (Lattimore and Szepesvári, 2020). Private information is modeled as part of the context vectors of the respective players. Opacity is modeled via the policy spaces of the respective players. This means that there is opacity if a player does not know another player's policy space. The learning objective in our model is a straightforward extension of the contextual bandit problem: The two players try to find decision rules that minimize the expected regret.

In our model, we are then able to show that both private information and opacity can present significant barriers to learning. Specifically, in the *absence* of private information and opacity, learning in our model is easy. Under the presence of private information *or* opacity, however, efficient learning becomes impossible, at least in the worst case. Intuitively, the two players are stuck with trial and error on all possible combinations of policies. An interesting insight from our formal modeling approach is that private information and opacity can have very similar consequences.

Intuitively, a main conclusion from the first part of the thesis is that human-machine interaction depends on good priors of what comprises successful interaction. In other words, since there is little potential to learn sophisticated strategies of human-machine interaction, we are essentially stuck with trial-and-error on a few strategies that we deem plausible in the first place.

Importantly, the first part of the thesis also informs about the role of algorithmic explanations in the human-machine decision making problem. This is because the hardness results in the paper also apply to the problem of learning explanations. By this, we mean the problem setup where a computer has access to a parameterized space of explanation algorithms and tries to learn the explanations that work best for the human decision maker. Intuitively, the results in the first part of the thesis imply that there is no efficient way in which a computer can automatically identify the best explanations. This again means that we need to rely on our prior knowledge of what might comprise useful explanations for human decision makers.

1.6.2 Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts

In the second part of the thesis, we study the suitability of a particular class of methods - local post-hoc explanation algorithms - in societal contexts. As outlined in Section 1.4.3, it is often believed that explainable machine learning might be useful as a regulatory tool, in order to enhance the transparency of automated decision making systems. We ask whether local post-hoc explanation algorithms (for example, feature attributions) can fulfill the transparency requirements inherent in the draft Artificial Intelligence Act. The Artificial Intelligence Act is a major piece of EU legislation that attempts to specifically regulate AI. As such, it presents an important case study for the potential of explainable machine learning in societal contexts. Using an interdisci-

plinary approach including philosophy, law and computer science, we argue that local post-hoc explanation algorithms are unable to achieve the transparency objectives that are inherent in the law.

Our argument rests on the observation that the law usually attempts to regulate what we call adversarial decision making contexts. In an adversarial decision making context, the incentives of the different actors - the entity that deploys the machine learning system, the provider of the explanations, and the individual that is affected by the machine prediction, are not aligned. As a consequence, the provider of the explanations has an incentive to manipulate the explanations to her end. From a technical perspective, it is extremely hard if not altogether impossible to safeguard against such manipulations. The reason for this is the high degree of ambiguity of post-hoc explanations in realistic application scenarios. Ground-truth explanations of complicated black-box functions do not exist, and from a technical perspective, the problem of providing post-hoc explanations is undetermined. In other words, the provider of the explanations has to make many choices that allow her to manipulate the explanations towards her end.

An important question about machine learning in societal contexts is to what degree the auditing of such systems is possible. In most cases, the entity that deploys the machine learning system - for example, a bank or a university - would not like the system to come under too much scrutiny, in order to avoid the detection of any potential form of malpractice. However, regulators could potentially specify that systems must be accessible for certain forms of testing.

Overall, the second part of the thesis provides a cautionary tale regarding the potential of explainable machine learning for transparency and regulation in societal contexts. At any rate, post-hoc explanations are unlikely to expose undesirable model behaviors such as biases, given that the provider of the explanations has an incentive to hide them. An interesting alternative to the usage of post-hoc explanations in critical applications might be the usage of interpretable models (Section 1.5.3). While these models can also hide indirect effects that occur due to the correlation between the different variables, they are overall much more transparent and potentially much simpler to audit.

1.6.3 From Shapley Values to Generalized Additive Models and back

The third part of the thesis develops connections between post-hoc explanation algorithms and interpretable models (Section 1.5.3). Specifically, we outline theoretical connections between Shapley Values - a prominent feature attribution method - and Generalized Additive Models (GAMs) - a popular class of interpretable models. This is interesting insofar as post-hoc methods and interpretable models are often framed as different or even opposing approaches towards interpretability (Rudin, 2019). By highlighting the connections between the two - if only for one particular pair of methods - we suggest that there are general ideas of interpretability that underlie both interpretable models and post-hoc explanation techniques. In addition, the third part of the thesis also presents a detailed analysis of Shapley Values as a model explanation

technique, identifying GAMs without variable interactions as the class of functions that Shapely Values are able to explain without loss of information.

In order to outline the connections between Shapley Values and GAMs in some generality, we introduce n-Shapley Values. n-Shapley Values are a family of local post-hoc explanation algorithms that explain individual predictions with interaction terms up to order n (that is, n-Shapley Values extend feature attributions with terms for variable interactions up to order n). We then show that n-Shapley Values are able to faithfully represent GAMs with variable interactions up to order n. This means that if the function that we want to explain can be represented as a GAMs of order n, then n-Shapley Values will implicitly provide such a representation, in the sense that the different terms involved in the explanations correspond to the evaluations of the respective component functions of the GAM. This result provides a direct link between the complexity of the model class and the complexity of the explanations that are required in order to faithfully explain it. A consequence of this general result is that the original Shapley Values are able to faithfully represent Generalized Additive Models of order $n = 1$, that is functions without variable interactions.

Overall, the third part of the thesis provides an interesting example of what a detailed theoretical analysis of a post-hoc explanation algorithm can look like. In particular, we demonstrate exactly how the post-hoc method relates to the properties of the function that we want to explain. What is more, the third part of the thesis also reveals the limitations of an entirely mathematical analysis of model explanations. In particular, if one believes that the presented analysis provides a relatively complete characterization of the mathematical properties of Shapley Values, which seems plausible, then it becomes clear that some of the more ambitious goals of explainable machine learning are likely not directly amenable to mathematical analysis. For example, the presented mathematical analysis of Shapley Values does not directly allow us to answer whether these explanations would be useful as a decision aid for doctors.

Chapter 2

Publications

2.1 A Bandit Model for Human-Machine Decision Making with Private Information and Opacity

A Bandit Model for Human-Machine Decision Making with Private Information and Opacity

Sebastian Bordt
University of Tübingen
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Ulrike von Luxburg
University of Tübingen
Max Planck Institute for Intelligent Systems
Tübingen, Germany

Abstract

Applications of machine learning inform human decision makers in a broad range of tasks. The resulting problem is usually formulated in terms of a single decision maker. We argue that it should rather be described as a two-player learning problem where one player is the machine and the other the human. While both players try to optimize the final decision, the setup is often characterized by (1) the presence of private information and (2) opacity, that is imperfect understanding between the decision makers. We prove that both properties can complicate decision making considerably. A lower bound quantifies the worst-case hardness of optimally advising a decision maker who is opaque or has access to private information. An upper bound shows that a simple coordination strategy is nearly minimax optimal. More efficient learning is possible under certain assumptions on the problem, for example that both players learn to take actions independently. Such assumptions are implicit in existing literature, for example in medical applications of machine learning, but have not been described or justified theoretically.

1 Introduction

The number of applications where machine learning informs human decision makers is steadily growing (Board of Governors, 2007; Angwin et al., 2016; Tonekaboni et al., 2018). In this work, we argue for a specific perspective on machine learning systems that inform human decision makers: We want to understand them as attempts to solve joint human-machine decision making problems where both sides have to learn to act optimally. This perspective will help to understand limitations and potential pitfalls of such systems. We believe that this is an important step towards robust and reliable systems (Rahwan et al., 2019).

Our motivation is the growing number of applications where machine learning advises human decision makers. For example:

(1) The COMPAS program that assists judges during criminal trials (Angwin et al., 2016). The program provides a risk assessment score for defendants in criminal law. Judges then use this score, among others, to decide whether a defendant should await trial at home or in jail, and to determine the length of prison sentences (Kleinberg et al., 2018; Forrest, 2021).

(2) Cardiac arrest and other forms of adverse event prediction. In medicine and beyond, it can be of great value to know when adverse events such as cardiac arrest are likely to occur (Tonekaboni et al., 2018; Shamout et al., 2020; Baker et al., 2020). This can often be predicted based on a limited amount of information. Computer programs alert doctors when a patient's condition is likely to become critical. Doctors respond with a treatment adapted to the patient's condition, which may include ignoring the alert.

(3) Diabetic retinopathy detection (Raghu et al., 2019). Deep learning has shown great capabilities to detect diabetic retinopathy in pictures of the eye. This has led to computer programs that inform doctors by assigning scores to to images. Doctors incorporate these scores into their decision making (Beede et al., 2020).

In all three examples, machine learning provides ad-

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

A Bandit Model for Human-Machine Decision Making with Private Information and Opacity

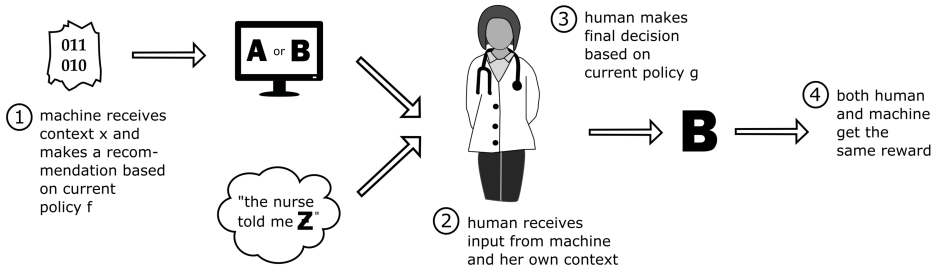


Figure 1: Illustrated application of our model: The computer is advising a human doctor. The computer recommends to perform action A or B. After consulting her additional private information, the human finally decides to perform action B.

vice, but final decisions are left to the human. Moreover, human decision makers base their decisions on additional *private information* that is unavailable to the machine. In the COMPAS example, the judge obtains additional information from the trail and the interaction with the defendant, attorney and prosecutor (Lakkaraju et al., 2017). In the medical example, private information might consist of non-digitized parts of the patient's medical history, or diseases that run in the family of the patient (Goldenberg and Engelhardt, 2019). Even in the diabetic retinopathy example, the final treatment decision is typically based on more information than just the picture of the eye.

In addition to the presence of private information, it has long been argued that human-machine cooperation is hampered by a certain degree of *opacity* (Leonelli, 2020). Indeed, despite a lot of work on explainable machine learning, computers cannot explain their decisions to humans the way other humans can, and computers cannot really understand free-form human explanations.

How can we design computer programs that optimally advise human decision makers in tasks such as (1)-(3)? What does "optimally" even mean in these contexts? To provide precise answers to these questions, we propose a *contextual bandit model with two players* that aims to capture the most important properties of the above decision problems.

The two players in our model, who we refer to as "the human" and "the machine", interact according to the following protocol (illustrated in Figure 1). In every round, the first player (the machine) receives private contextual information and makes a recommendation to the second player (the human). This recommendation can be a suggested action, but it can also be a confidence region, a colorfully highlighted image or any other summary of the received context. In the COMPAS example, the recommendation is the risk assessment score. Given the recommendation and her own private contextual information, the human finally decides on an action. Conditional on context and the chosen action, a reward signal is obtained. Action and reward are observed by both players, and they share the same goal: to maximize the obtained rewards.

We endow each of the players with a finite set of decision rules or policies, which is the simplest possible learning setting. The goal is to minimize the minimax regret with respect to the two decision rules that work best together. We first analyze the case where the human does not attempt to learn (Section 4). However, we believe that the fact that human decision makers have to learn how to "interpret" machine recommendations is a crucial aspect of human-machine decision making. In cardiac arrest prediction, for example, doctors have reported to learn over time how to interpret machine alerts and integrate them into existing clinical practice. We therefore also consider the problem where human and machine *both* have to learn (Section 5).

The main objective of this paper is to gain a theoretical understanding of an emerging number of human-machine decision making problems, such as (1)-(3). By considering the interaction between two abstract decision makers in the presence of private information and opacity, we aim to provide a general analysis of the potential and limitations of human-machine decision making. While our main intention is to set a theoretical baseline for more applied work in human-machine interaction, we also hope that our proposed model and newly introduced problems will spark theoreticians interest into various aspects of the human-machine learning problem. Our main contributions are the following.

- We prove that private information and opacity significantly impact the hardness of two-player decision making. Private information and opacity each lead to a worst-case *lower bound* of order $\sqrt{TN_1}$ (Theorem 3). Here N_1 is the number of policies of the first player. Without private information and opacity, the two players can obtain an efficient expected regret of $\sqrt{2TK\ln(N_1N_2)}$ (Proposition 1). Here N_2 is the number of policies of the second player.
- We show that a simple coordination device – telling the machine which policy to use – allows the human to learn efficiently. Specifically, the P2-EXP4 algorithm allows to *upper bound* the expected minimax regret by $\sqrt{2TKN_1\ln(N_1N_2)}$ (Theorem 4), also in the presence of private information and opacity.
- We derive a criterion – policy space independence – that allows to learn with an expected regret of $\sqrt{8T\max\{K,|\mathcal{R}|\}\ln(\max\{N_1,N_2\})}$ (Theorem 6). Here $|\mathcal{R}|$ is the number of possible machine recommendations. If policy space independence holds and $|\mathcal{R}|$ is small, the two players can learn efficiently.
- In Sections 6 and 7, we show that various approaches in the literature can be better understood within the context of our model. In particular, policy space independence is implicit in much of the existing literature. The peculiar case of treatment recommendations is left as Conjecture 7.

2 Our model: The computer reports to the human, who then decides

Formally, our model is a contextual bandit model with two players, depicted in Figure 2. In round $t = 1, \dots, T$, Player 1 (the machine) first observes context $x_t \in \mathcal{X}$. Player 1 then chooses a recommendation $r_t \in \mathcal{R}$, potentially at random. Here, $\mathcal{R}$ is the space of all possible recommendations that the first player can make. Next, Player 2 (the human) observes context $z_t \in \mathcal{Z}$ and the chosen recommendation r_t. Player 2 then, potentially at random, chooses an action $a_t \in A$. This action is revealed to both players, and they receive a reward signal $y_t \in [0,1]$. Here $\mathcal{X}$ and $\mathcal{Z}$ are arbitrary spaces of private contexts (one for each player), and $A = \{1, \dots, K\}$ is a finite set of K actions.

2.1 Formal setup

Both players are endowed with a finite set of policies. Their common goal is to take optimal actions. Let

In round $t = 1, ..., T$

1. Context $x_t \in \mathcal{X}$ is revealed to Player 1
2. Player 1 decides on a recommendation $r_t \in \mathcal{R}$
3. Context $z_t \in \mathcal{Z}$ and recommendation r_t are revealed to Player 2
4. Player 2 decides on an action $a_t \in A$
5. Reward $y_t \in [0,1]$ and action a_t are revealed to both players

Figure 2: Interaction in our contextual bandit model.

$\Pi_1 \subseteq \mathcal{R}^{\mathcal{X}}$ be a finite set of policies for the first player, and $\Pi_2 \subseteq A^{\mathcal{R}\times\mathcal{Z}}$ a finite set of policies for the second player. Given two policies $f \in \Pi_1$ and $g \in \Pi_2$, we obtain the resulting joint policy $\pi(x,z) = g(f(x),z)$. This joint policy is a complete decision rule for the problem, translating context into actions. Let $\Pi = \Pi_2 \times \Pi_1$ be the space of all combinations of policies that the two players can possibly realize. For a tuple $\pi = (g,f) \in \Pi$, we slightly abuse notation and write $\pi(x,z) = g(f(x),z)$ to refer to the corresponding joint policy.[1] Moreover, we denote the number of policies $N_1 = |\Pi_1|$ and $N_2 = |\Pi_2|$. We have $N = |\Pi| = N_1N_2$.

An algorithm for the two players is a pair $A = (A_1, A_2)$. Here $A_1 = (A_{1,t})_{t=1}^T$ and $A_2 = (A_{2,t})_{t=1}^T$ are two collections of measurable functions that specify the decision rules of both players at all points in time. The domains of these functions specify which variables are observable to which player at what time. Thus, $A_{1,t}$ is a function of $x_1, \dots, x_t$, whereas $A_{2,t}$ is a function of $r_1, \dots, r_t$ and $z_1, \dots, z_t$. The details of this can be found in Supplement A.1.

Let $\mathcal{D}$ be a probability distribution over $\mathcal{X}\times\mathcal{Z}\times[0,1]^A$. We consider an i.i.d. contextual bandit model where tuples (x_t, z_t, Y_t) are i.i.d. draws from $\mathcal{D}$. Let $Y(\pi) = \mathbb{E}_{(x,z,Y)\sim\mathcal{D}}[Y(\pi(x,z))]$ be the expected reward of a joint policy π. Let $\pi^\star \in \arg\max_{\pi\in\Pi} Y(\pi)$ be a policy combination that maximizes the expected reward. The expected regret after T rounds is given by $\mathrm{Reg}_T = \mathbb{E}\left[T\,Y(\pi^\star) - \sum_{t=1}^T Y_t(a_t)\right]$, where the expectation is over $\mathcal{D}$ and the randomly selected actions and recommendations. The central quantity of analysis is the minimax regret, given by $R_T = \inf_A \sup_{\mathcal{D}} \sup_{|\Pi_1|=N_1} \sup_{|\Pi_2|=N_2} \mathrm{Reg}_T$.

[1] Depending on Π_1 and Π_2, different tuples (g,f) can give rise to the same policy $\pi : \mathcal{X}\times\mathcal{Z} \to A$.

2.2 First thoughts and discussion of modelling assumptions

Private context. This is our approach to model private information. In real-world decision making problems such as (1)-(3), humans often have access to information that is not available to any algorithm. A reason for this might be that some information, such a detailed health record, is not yet available in electronic form. However, we also believe that in many of the tasks where machine learning is increasingly being deployed at, formulating all relevant aspects as inputs to an algorithm is impossible. This is because machine learning is increasingly being deployed in social contexts where researchers have long accepted the fact that its impossible to exhaustively collect all relevant variables (Angrist and Pischke, 2008). Our model also allows for private contextual information of the machine. In the medical domain, an algorithm might have access to a patient's genome data, which could never be entirely surveyed by a human. Unobserved variables might also occur in unexpected situations, such as when both decision makers coordinate a decision based on the same image. Here algorithms have been shown to rely on high-frequency patterns that are imperceptible to humans (Ilyas et al., 2019; Makino et al., 2020).

Private policy spaces. We model opacity by keeping knowledge about the policy spaces to the respective players. Intuitively, this means that the players cannot deliberate about what happened: The machine does not know which actions the human would have chosen had it chosen a different recommendation. Similarly, the human does not know which recommendations the machine considered but decided against. While policy spaces are private, we place no restrictions on the algorithms that both players might run.

The space of recommendations. The space of recommendations $\mathcal{R}$ is the interface by which the first player can transmit information to the second player (Goodrich and Schultz, 2007). For the first player, it plays the role of an action space (providing a recommendation is the action that the first player takes). For the second player, it resembles additional contextual information. In the analysis, will turn out to be useful to restrict the size of the space of recommendations (Section 6). A large space of recommendations allows the machine to provide the human with rich contextual information. This includes the scenario where the machine attempts to "explain" predictions in some rich space. A concrete example of this would be when the machine provides a saliency map (Simonyan et al., 2014; Selvaraju et al., 2017). In contrast, a small space of recommendations allows the machine to suggest concrete actions, or to raise an alert. A priori, it seems unclear which of these two approaches will be more useful. On one hand, we might want the machine to provide the human with as much information as possible. On the other hand, it might be more efficient if the machine directly suggests which actions to perform. In Sections 3-5, we remain agnostic about the nature of the space of recommendations. The special case of treatment recommendations is discussed in Section 7.

What makes the model difficult? For both players, the difficulty arises from the fact that contextual information and policy space of the other player are unknown. This gives rise to a *coordination problem*. Each player would like to find the optimal policy that works best in combination with the strategy chosen by the other player. This is difficult because knowledge about the other player's decision problem is limited.

Online learning. Our model is an online learning model. This allows us to study the process by which the two decision makers coordinate and arrive at decisions. In practice, an algorithm would always be trained on a historical dataset before it starts to interact with a human decision maker. However, if we continuously gather data in order to retrain and improve our algorithm, we are implicitly engaging in an online learning procedure. We are directly considering an online learning model since this allows us to study the principal limitations and possibilities of various approaches. For more details on online and repeated supervised learning we refer the reader to Supplement E.

Worst-case analysis. Intuitively, a strategy of the two players might work well for some decision problems and fail for others. Considering the minimax regret means that we would like to find guarantees that can be achieved under *all possible circumstances*. That said, it is interesting to ask how much better the two players can do if we assume that the decision problem is 'benign' – a question that we turn to in Section 6.

3 Two baselines for the expected regret

How well can we expect the two players to coordinate, and what are the consequences of private information and opacity for two-player decision making? To provide answers to these questions, we are first going to consider our model *without* private information and opacity. No private information means that $\mathcal{X} = \mathcal{Z}$ and $x_t = z_t$ for all t. No opacity means that the algorithm of the first player is also a function of the policy space of the second player and vice-versa.

Proposition 1. ***(Regret without private information and opacity)*** *Without private information*

and opacity, the two players can obtain an expected regret of

$$\sqrt{2TK\ln(N_1N_2)}.$$

All proofs are deferred to the Supplement. The regret bound in Proposition 1 is as good as we can expect at all.[2] It is the same regret that would be achieved by a hypothetical single decision maker who had access to all contextual information and both policy spaces, using EXP4 (Auer et al., 2002; Lattimore and Szepesvari, 2019). Proposition 1 demonstrates that our hardness result (Theorem 3) is a consequence of private information and opacity, and not due to the way in which the two players interact in our model.

A second baseline is given by the coordination strategy where players naively try all policy combinations. This strategy also works *with* private information and opacity. Using the MOSS algorithm by Audibert and Bubeck (2010):

Proposition 2. ***(Naively trying all policy combinations)*** *By treating the policies in* Π *as different arms of a stochastic bandit, we obtain*

$$R_T \leq O\left(\sqrt{TN_1N_2}\right).$$

Under private information and opacity, can the two players do better than what is suggested by Proposition 2? The question becomes whether it is possible to *move* N_1 *and* N_2 *inside the logarithm*, at the expense of a factor of K. Why is this important? A regret bound of order $\sqrt{N}$ means that the policy space is not dealt with efficiently. It corresponds to systematic trial and error on every single policy. Quite to the contrary, a regret bound of order $\sqrt{\ln N}$ means that the decision maker can compare many policies simultaneously. It prepares the way to deal with infinite policy spaces and learn rich function classes (Beygelzimer et al., 2010).

4 A lower bound for optimal algorithmic advice

Before we turn to the full problem where human and machine both have to learn, we focus on the problem of the machine. That is we assume that the human does not have to learn how to interpret machine recommendations. This is a significant simplification, but the result will be instructive. We are going to show that private information and opacity *each* lead to a lower of order $\sqrt{TN_1}$.

[2]For adversarial contextual bandits, the bound $\sqrt{TK\ln(N)}$ has been shown to be tight up to a factor of $\ln K$ by Seldin and Lugosi (2016). Note that we are concerned with statistical optimality and set computational concerns aside.

Formally, we assume that the second player follows a fixed decision rule that deterministically translates recommendations r_t and contextual information z_t into actions. The first player has N_1 different policies and wants to learn the best one. How difficult is this learning problem? Note that we do not place any restrictions on the space of recommendations $\mathcal{R}$. However, the ultimate number of actions K is small. Can the first player make use of this fact and solve the problem efficiently? In the presence of private information or opacity, this is not the case.

Theorem 3. ***(Lower bound in the number of policies of the first player)*** *Assume that Player 2 only plays actions that are suggested by policies in* Π_2. *Let* $N_2 = 1$ *and* $K = 2$. *There exists a universal constant* $c > 0$ *such that*

$$R_T \geq c\sqrt{TN_1}.$$

The lower bound in Theorem 3 is as strong as it can possibly be. It shows that the first player has to solve a bandit problem that depends *not* on the number of actions K, but on the number of policies N_1.

Supplement A.4 contains two proofs of Theorem 3. The first proof constructs problem instances with private information but without opacity, and the second proof constructs problem instances with opacity but without private information.[3] In conclusion, both private information and opacity can significantly impact on the hardness of two-player decision making.

Remark 1. The reader might be worried by the fact that we fixed Player 2. Indeed, even if the policy space of the second player consists of a single decision rule, it might be optimal to deviate in order to facilitate coordination. To alleviate such concerns, Supplement C presents a problem class that allows the second player to choose actions arbitrarily. We discuss the general issue of encoding information about the policy spaces in actions and recommendations in Supplement F.

5 Efficient learning for a human who controls the machine

We now consider the full problem of human-machine learning where the human learns how to interpret machine recommendations. Intuitively, the human tries to figure out how to act on machine advice. At the same time, the machine tries to determine how to advise the human. How should the two players coordinate? Is it possible that both explore simultaneously?

[3]From a theoretical perspective it might not be surprising that private information and opacity have the same consequences. Ultimately, what matters are the expert predictions that result from the interaction of context and policy (Cesa-Bianchi and Lugosi, 2006).

Algorithm P2-EXP4
Parameters: $\eta > 0$, $\gamma > 0$
Initialization: $Q_1 \in [0,1]^{N_1 \times N_2}$ with $Q_{1,ij} = \frac{1}{N_1 N_2}$
For each $t = 1, ..., T$

1. Player 2 tells Player 1 to play policy i_t according to $q_{ti} = \sum_{j=1}^{N_2} Q_{t,ij}$
2. Player 1 recommends $r_t = f_{i_t}(x_t)$
3. Player 2 chooses action a_t according to (1)
4. Players receive reward y_t and Player 2 estimates $\hat{y}_{tk} = 1 - \frac{1_{\{a_t=k\}}}{q_{t,i_t} p_{tk} + \gamma}(1 - y_t)$
5. Player 2 propagates rewards to policies $\hat{Y}_{t,ij} = 1_{\{i_t \neq i\}} + 1_{\{i_t = i\}} \hat{y}_{t,g_j(r_t,z_t)}$
6. Player 2 updates Q_t using exponential weighting

$$Q_{t+1,ij} = \frac{\exp(\eta \hat{Y}_{t,ij}) Q_{t,ij}}{\sum_{l,m} \exp(\eta(\hat{Y}_{t,lm}) Q_{t,lm}}$$

Figure 3: The P2-EXP4 algorithm allows the second player to explore efficiently.

From Theorem 3 in the previous section, we already know that the learning problem of the machine is hard, even if the human sticks to a single policy. We are now going to show that a simple coordination device allows the human to explore efficiently: *We allow the human to tell the machine which policy to use.* Intuitively, we can perceive the N_1 different policies of the machine as different computer programs. The human tries to learn which of these computer programs to use. In doing so, the human can explore with exponential weighting, but only on its own policy space, and not on the policy space of the machine. This idea is formalized in the P2-EXP4 (Player 2-EXP4) algorithm, depicted in Figure 3. Theorem 4 shows that P2-EXP4 nearly allows to match the lower bound in Theorem 3.

Theorem 4. ***(Logarithmic regret in the number of policies of the second player)*** *The P2-EXP4 algorithm with* $\eta = \sqrt{2\log(N_1 N_2)/(TKN_1)}$ *and* $\gamma = 0$ *satisfies*

$$R_T \leq \sqrt{2TKN_1 \ln(N_1 N_2)}.$$

The proof of Theorem 4 is in Supplement A.5. We now describe the algorithm. Player 2 maintains a probability distribution Q_t over the space of all policies Π. In every round t, Player 2 first chooses a policy f_{i_t} for Player 1 by drawing i_t from the marginal distribution of Q_t over Π_1. After obtaining a recommendation r_t and context z_t, Player 2 draws a_t according to the induced probability distribution over actions

$$\mathbb{P}(a_t = k) = p_{tk} \quad \text{with} \quad p_{tk} = \frac{\sum_{j=1}^{N_2} Q_{t,i_t,j} 1_{\{g_j(r_t,z_t)=k\}}}{\sum_{j=1}^{N_2} Q_{t,i_t,j}}. \tag{1}$$

With the reward signal y_t, Player 2 computes importance-weighted reward estimates for all policies and then uses an exponential weighting scheme to update Q_t.

Remark 2. The proof of Theorem 4 relies on the fact that the updates performed by P2-EXP4 are equivalent to the updates performed by EXP4 on a related bandit problem with KN_1 actions. As a consequence, all results for EXP4 carry over to P2-EXP4. In particular, for $\gamma > 0$, P2-EXP4 is a variant of EXP4-IX (Neu, 2015). This implies that P2-EXP4 enjoys high-probability regret guarantees. Furthermore, Theorem 4 also holds when contexts and payoffs are determined by an adversary.

6 Efficient learning for the machine, subject to further assumptions

We now discuss additional assumptions on the structure of the problem that allow for more efficient learning. The first idea is to restrict the size of the space of recommendations $\mathcal{R}$. If the machine directly recommends actions, for example, we have $\mathcal{R} = A$. The second idea is to resolve the coordination problem. This can be done via an assumption on the function spaces of both players that we term *policy space independence*. While policy space independence is an abstract criterion, we outline a number of practical examples where it is satisfied.

This section also relates our work to a number of recently proposed techniques for human-machine interaction (Madras et al., 2018; Raghu et al., 2019; Wilder et al., 2020). We will show that policy space independence is implicit in much of the existing literature on human-machine decision making.

6.1 Policy space independence

We now give an abstract condition that *resolves the coordination problem* and allows both players to learn independently. It is an assumption on the policy spaces. The rationale is that assumptions on the policy spaces can implicitly define how human and machine interact.

Definition 5 (Policy space independence). *We say that the two policy spaces* Π_1 *and* Π_2 *are independent with respect to* $\mathcal{D}$ *if, for all* $f_1, f_2 \in \Pi_1$ *and all*

$g_1, g_2 \in \Pi_2$,

$$\begin{aligned} & Y\big(g_1(f_1(x), z)\big) - Y\big(g_1(f_2(x), z)\big) \\ = \; & Y\big(g_2(f_1(x), z)\big)) - Y\big(g_2(f_2(x), z)\big). \end{aligned}$$

Intuitively, whether policy f_1 performs better than policy f_2 does not depend on the policy chosen by the second player. Similarly, whether policy g_1 performs better than policy g_2 does not depend on the policy chosen by the first player. Hence, the learning problems of both players are decoupled. The following theorem shows that policy space independence allows to efficiently learn both policy spaces.

Theorem 6 (Logarithmic regret under policy space independence). *Under policy space independence, if both players explore independently using EXP4,*

$$R_T \leq \sqrt{8T \max\{K, |\mathcal{R}|\} \ln(\max\{N_1, N_2\})}.$$

The proof of Theorem 6 is in Supplement A.6. In contrast to Theorem 4, both N_1 and N_2 appear inside the logarithm. This is at the expense of a factor $|\mathcal{R}|$.

6.2 Allocating decisions between human and machine

If $|\mathcal{R}|$ is small and policy space independence holds, the two players can obtain an efficient expected regret (Theorem 6). But what does this amount to in practice? First note that we can constrain the policy space of the human by specifying rules for how to interact with the machine. For example: "If the machine depicts 'action a', then perform action a". This leads to the following example: Policy space independence holds when there exists a fixed rule that allocates every decision to *either* the human *or* the machine. In medical applications, this would mean that there exists some procedure that determines whether a given case should be decided by the doctor or the machine. For diabetic retinopathy detection, such a procedure was recently proposed by Raghu et al. (2019), who also demonstrate that the approach can lead to substantial benefits in practice. In our model, the rule can be any predicate $P(z)$, that is the human decides who decides. It can also be any predicate $P(x)$, that is the machine decides who decides. Importantly, in order to satisfy policy space independence, the rule cannot be learned while the decision makers learn themselves. We formally show in Supplement B how fixed rules that allocate decisions result in policy space independence.

6.3 Learning to defer

Another example of policy space independence is given by learning to defer (Madras et al., 2018; Mozannar and Sontag, 2020). Learning to defer is characterized by two assumptions. First, the human is a fixed decision maker who does not learn. Second, the space of recommendations is given by $\mathcal{R} = \mathcal{A} \cup \{\texttt{D}\}$, where `D` denotes that the decision is deferred to the human. As can be seen from Definition 5, fixing any of the two decision makers always results in policy space independence. According to Theorem 6, the regret of learning to defer is thus bounded by $\sqrt{8T(K+1)\ln(N_1)}$.[4]

6.4 Other approaches

With some notable exceptions (Hilgard et al., 2019), the literature on human-machine decision making often relies on assumptions similar to fixed rules that allocate decisions and learning to defer (De et al., 2020a,b). It is usually assumed that the human is a fixed decision maker whose performance on the given task can be queried or deferred to (Wilder et al., 2020; Pradier et al., 2021). Specifically, the human does not have to learn how to interact with the machine. Moreover, machine recommendations usually equal actions, with some room for special recommendations in order to involve the human. Viewed through the lens of our model, all of these approaches satisfy policy space independence. In light of Theorem 6, they all allow for efficient learning.

7 How difficult are treatment recommendations?

In the last section, we have seen that the learning problem of the machine can be simplified by (1) choosing $\mathcal{R} = \mathcal{A}$ and (2) fixing the human decision maker. In Section 4, we have seen that the condition $\mathcal{R} = \mathcal{A}$ is crucial (after all, the lower bound was derived for a fixed human decision maker). But is it equally necessary to choose $|\Pi_2| = 1$? This is interesting because $\mathcal{R} = \mathcal{A}$ is satisfied, among others, in screening scenarios. These are the binary classification problems studied in the literature on fairness and machine learning (Kleinberg et al., 2019; Barocas et al., 2019). Here a decision problem might be whether to give a loan or to admit a student to a university. It is often argued that such machine suggestion should still be reviewed by humans (De-Arteaga et al., 2020).

In our model, binary predictions that are reviewed by humans correspond to $\mathcal{R} = A = \{0, 1\}$ and $|\Pi_2| > 1$ (assuming that the human learns when to override machine predictions). If either $N_1 = 1$ or $N_2 = 1$, EXP4 allows to bound the expected regret by $\sqrt{4T\ln(N_2)}$ and $\sqrt{4T\ln(N_1)}$, respectively. Therefore, consider the corner case $N_1 = N_2$. If we assume that the sec-

[4] For $N_2 = 1$, the constant could be improved to 2.

ond player can tell the first player which policy to use, P2-EXP4 allows to bound the expected regret by $\sqrt{8TN_1 \ln(N_1)}$. We conjecture that this is tight up to a constant factor, i.e. that treatment recommendations are difficult.

Conjecture 7. ***(Lower bound in the number of policies if $\mathcal{R}$ and A are small)*** *Let $\mathcal{R} = A = \{0, 1\}$ and $N_1 = N_2$. We conjecture that there exists a universal constant $c > 0$ such that*

$$R_T \geq c\sqrt{TN_1 \ln N_1}.$$

Supplement D details a problem instance that we believe to be worst-case.[5]

8 Related Literature

Researchers have long asked how humans can interact with computers and robots (Sheridan and Verplank, 1978; Goodrich and Schultz, 2007; Parasuraman et al., 2000). In machine learning, researchers increasingly study how humans and automated decision making systems can interact (Tonekaboni et al., 2019; Carroll et al., 2019; Lucic et al., 2020; De-Arteaga et al., 2020). A number of recent works have argued that joint human-machine decision making can outperform a single human or a single machine (Lakhani and Sundaram, 2017; Raghu et al., 2019; Patel et al., 2019). Human-computer interaction and the social sciences study the different ways in which machine recommendations can influence and alter human decisions (Dietvorst et al., 2015; Green and Chen, 2019).

Multi-player multi-armed bandits (Kalathil et al., 2014; Boursier and Perchet, 2019; Martínez-Rubio et al., 2019), economic game theory (Mas-Colell et al., 1995; Von Neumann and Morgenstern, 2007) and combinations thereof (Sankararaman et al., 2021) also study the interaction between multiple players. However, models in economic game theory are *competitive*, and the *cooperative* models in multi-player multi-armed bandits, often inspired by applications in wireless networks (Avner and Mannor, 2016), are *symmetric*. In contrast, interaction our model is cooperative and *asymmetric* – only the second player decides on a payoff-relevant action. Insofar as implicit communication between the two players is concerned, our work probably relates most closely to Bubeck et al. (2020), who study implicit communication in a symmetric collision problem (compare also Supplement F).

[5] The reader might wonder whether interaction terms between K and $\mathcal{R}$ appear in any bound. Beyond the special regime $\mathcal{R} = A = \{0, 1\}$ and $N_1 = N_2$, this might well be the case.

9 Discussion

The consequences of private information and opacity. We have shown that private information and opacity can have a significant effect on human-machine decision making. In the worst-case, the machine cannot advance beyond simple trial and error on a small number of policies (Theorem 3). Does this imply that we can never obtain good results in general human-machine decision making problems where we cannot make plausible assumptions on the presence of private information and opacity? Not necessarily. It does, however, imply that we need good priors for what comprises successful human-machine cooperation on a given task. Note that in practice, researchers often obtain a small number of candidate machine policies from historical data, then evaluate which one works best with human decision makers (Sayres et al., 2019; Tschandl et al., 2020). This approach is closely related to running the P2-EXP4 algorithm: The policy space of the machine consists of the candidate decision rules that were obtained form historical data. In the absence of further assumptions about the problem, we show this approach to be essentially minimax optimal. In some applications, it might be relatively easy to come up with good machine policies. There are, however, also problems where it is hard say how the machine should best inform the human. Consider the example where the machine informs the human about an image: While there have been many empirically successful attempts at such problems, there is still a big debate about post-hoc explainability methods, what properties they should have, and whether they should be used at all (Adebayo et al., 2018; Rudin, 2019).

Different modalities of human-machine decision making. We have seen in Section 6 that our model possesses sufficient generality to analyze a wide array of interaction protocols between humans and machines. Of course, there are many different settings of human-machine decision making, and our model can only serve as first step towards a formal analysis. From a theoretical perspective, it remains an interesting open question whether there are weaker assumptions than policy space independence that allow for efficient learning (Theorem 6). One might also ask whether distributional assumptions that restrict the influence of unobserved variables on the outcome can result in improved bounds. From a practitioner's point of view, the most important question is which assumptions are plausibly satisfied in applications.

Prediction problems. In many decision support systems, machine learning is merely used to solve a specific prediction or classification problem, whose outcome is then transferred to the human. Examples are

scores to predict criminal recidivism, cardiac arrest and severity of diabetic retinopathy. While such an approach is a straightforward way of human-machine interaction, nothing guarantees that the approach will be successful. For example, there have been numerous concerns about the consequences of COMPAS scores on the decision making of judges (Forrest, 2021). In our view, the belief that the humans should be informed with the scores of a particular prediction problem is a very strong prior on the policy space of the machine: the policy space consists of a single policy. In order to credibly identify successful forms of human-machine cooperation, we should however consider a variety of plausible machine policies, and also account for the fact that human decision makers have to learn how to interact with them. This is exactly the setting that we consider in this paper.

Human learning model. In our model, there are no constraints on the algorithm that the two players might run. We also remain entirely agnostic about the policy spaces of both players. This serve the purpose of generality and keeps our work closely aligned with the extant literature on contextual bandits. However, these two assumptions are also major simplifications, especially insofar as the human decision maker is concerned. Indeed, it is a well-known fact that humans are not perfectly rational decision makers and have problems to deal with probabilities (Gigerenzer and Kurzenhaeuser, 2005; Kahneman and Frederick, 2005). A human would not be able to correctly perform the updates prescribed by P2-EXP4, MOSS, or any other bandit algorithm for that matter. The results provided in this paper apply to two perfectly rational decision makers who have access to arbitrary computational and cognitive resources. Two decision makers who only have access to limited computational and cognitive resources might hope to achieve as much, but will in general not be able to do any better. Of course it is an interesting question to ask how specific behavioral assumptions on the the human decision maker, such as bounded rationality (Selten, 1990) or biases when dealing with machine recommendations (Green and Chen, 2019) influence optimal interaction. In the context of our model, such assumptions might take the form of assumptions on the policy space of the human, or the way in which the second decision maker selects policies in every round. This might be an interesting avenue for future research. Note, however, that in the context of our model, "the human" does not necessarily correspond to a single (biological) human. In most applications that we are interested in (compare (1)-(3) in Section 1), there are many different judges or doctors that interact with a given machine learning system. While these human decision makers certainly learn individually how to interpret machine recommendations, they also engage in a collective learning procedure (Rakoff, 2021). While questions around the correct modelling of human-machine interaction are certainly very interesting, our objective in this paper is not to propose a universal model of human-machine interaction. Instead, our objective is to propose a model that is as simple as possible while still being able to capture the relations that we are interested in.

Exploration in high-stakes decision making problems. In many human-machine decision making problems, direct exploration is highly problematic (for example in medical applications). In these applications, it is often impossible to explore according to an online algorithm during deployment. Instead, exploration is only possible during certain development stages (e.g. when we evaluate in a controlled study how doctors respond to different kinds of machine recommendations). In bandit models in particular, there are a number of different approaches – such as batching and offline learning – that can be taken in order to model constraints on exploration (Amani et al., 2019; Liu et al., 2020). In any case, full online learning, that is the modelling approach taken in this paper, can only serve as a simple theoretical model for the process in which algorithmic decision aids are developed, tested and refined in practice (compare also Supplement E).

Ethical impact. This work discusses statistical efficiency, which is in itself not a sufficient criterion to justify automation. This is especially true in medicine, an area that is believed to experience the widespread deployment of machine learning systems in the future (Froomkin et al., 2019; Grote and Berens, 2020). Automated decision making may also arise in undesired contexts. However, it remains important to understand it in the scenarios where it is desirable. As our work concerns theoretical foundations, theorems and proofs, we do not believe that it will have immediate negative consequences.

Acknowledgements

We would like to thank Sébastien Bubeck, Nicolò Cesa-Bianchi and Thomas Grote for helpful discussions. We would also like to thank Ronja Müller and Ruben Thoms for helping to create Figure 1. This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645), the BMBF Tübingen AI Center (FKZ: 01IS18039A), and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

S. Amani, M. Alizadeh, and C. Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 2019.

J. D. Angrist and J.-S. Pischke. *Mostly harmless econometrics*. Princeton University Press, 2008.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. ProPublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 2010.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 2002.

O. Avner and S. Mannor. Multi-user lax communications: a multi-armed bandit approach. In *IEEE International Conference on Computer Communications*, 2016.

S. Baker, W. Xiang, and I. Atkinson. Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach. *Scientific Reports*, 2020.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *CHI Conference on Human Factors in Computing Systems*, 2020.

A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Board of Governors. *Report to the Congress on Credit Scoring and its Effects on the Availability and Affordability of Credit*. Board of Governors of the US Federal Reserve System, 2007.

E. Boursier and V. Perchet. SIC-MMAB: synchronisation involves communication in multiplayer multi-armed bandits. In *Advances in Neural Information Processing Systems*, 2019.

S. Bubeck, Y. Li, Y. Peres, and M. Sellke. Non-stochastic multi-player multi-armed bandits: Optimal rate with collision information, sublinear without. In *Conference on Learning Theory*, 2020.

M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the Utility of Learning about Humans for Human-AI Coordination. In *Advances in Neural Information Processing Systems*, 2019.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

A. De, P. Koley, N. Ganguly, and M. Gomez-Rodriguez. Regression under human assistance. In *AAAI*, 2020a.

A. De, N. Okati, A. Zarezade, and M. Gomez-Rodriguez. Classification under human assistance. *arXiv preprint arXiv:2006.11845*, 2020b.

M. De-Arteaga, R. Fogliato, and A. Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *CHI Conference on Human Factors in Computing Systems*, 2020.

B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 2015.

K. B. Forrest. *When Machines Can Be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence*. World Scientific, 2021.

A. M. Froomkin, I. Kerr, and J. Pineau. When AIs outperform doctors: confronting the challenges of a tort-induced over-reliance on machine learning. *Ariz. L. Rev.*, 61:33, 2019.

G. Gigerenzer and S. Kurzenhaeuser. Fast and frugal heuristics in medical decision making. In *Science and medicine in dialogue: Thinking through particulars and universals*. Praeger Westport, CT, 2005.

A. Goldenberg and B. Engelhardt. Machine learning for computational biology and health. Tutorial at Advances in Neural Information Processing Systems Conference, 2019.

M. Goodrich and A. Schultz. Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 2007.

B. Green and Y. Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 2019.

T. Grote and P. Berens. On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, 2020.

S. Hilgard, N. Rosenfeld, M. R. Banaji, J. Cao, and D. C. Parkes. Learning representations by humans, for humans. *arXiv preprint arXiv:1905.12686*, 2019.

A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019.

D. Kahneman and S. Frederick. *A model of heuristic judgment.* Cambridge University Press, 2005.

D. Kalathil, N. Nayyar, and R. Jain. Decentralized learning for multiplayer multiarmed bandits. *IEEE Transactions on Information Theory*, 2014.

J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 2018.

J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the Age of Algorithms. *Journal of Legal Analysis*, 10:113–174, 04 2019.

P. Lakhani and B. Sundaram. Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*, 2017.

H. Lakkaraju, J. Kleinberg, J. Leskovec, J. Ludwig, and S. Mullainathan. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

T. Lattimore and C. Szepesvari. *Bandit Algorithms.* Cambridge University Press, 2019.

S. Leonelli. Scientific research and big data. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Metaphysics Research Lab, Stanford University, Summer 2020 edition, 2020.

Y. Liu, A. Swaminathan, A. Agarwal, and E. Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

A. Lucic, H. Haned, and M. de Rijke. Why does my model fail? Contrastive local explanations for retail forecasting. In *ACM Conference on Fairness, Accountability, and Transparency*, 2020.

D. Madras, T. Pitassi, and R. S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 2018.

T. Makino, S. Jastrzebski, W. Oleszkiewicz, C. Chacko, R. Ehrenpreis, N. Samreen, C. Chhor, E. Kim, J. Lee, K. Pysarenko, et al. Differences between human and machine perception in medical diagnosis. *arXiv preprint arXiv:2011.14036*, 2020.

D. Martínez-Rubio, V. Kanade, and P. Rebeschini. Decentralized cooperative stochastic bandits. In *Advances in Neural Information Processing Systems*, 2019.

A. Mas-Colell, M. D. Whinston, J. R. Green, et al. *Microeconomic theory.* Oxford University Press, New York, 1995.

H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. *arXiv preprint arXiv:2006.01862*, 2020.

G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 2000.

B. N. Patel, L. Rosenberg, G. Willcox, D. Baltaxe, M. Lyons, J. Irvin, P. Rajpurkar, T. Amrhein, R. Gupta, S. Halabi, C. Langlotz, E. Lo, J. Mammarappallil, A. J. Mariano, G. Riley, J. Seekins, L. Shen, E. Zucker, and M. P. Lungren. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine*, 2019.

M. F. Pradier, J. Zazo, S. Parbhoo, R. H. Perlis, M. Zazzi, and F. Doshi-Velez. Preferential mixture-of-experts: Interpretable models that rely on human expertise as much as possible. *arXiv preprint arXiv:2101.05360*, 2021.

M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, and S. Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.

I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, et al. Machine behaviour. *Nature*, 2019.

J. S. Rakoff. Sentenced by algorithm. *The New York Review of Books*, 2021.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

A. Sankararaman, S. Basu, and K. A. Sankararaman. Dominate or delete: Decentralized competing bandits in serial dictatorship. In *International Conference on Artificial Intelligence and Statistics*, 2021.

R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 2019.

Y. Seldin and G. Lugosi. A lower bound for multi-armed bandits with expert advice. In *13th European Workshop on Reinforcement Learning (EWRL)*, 2016.

R. Selten. Bounded rationality. *Journal of Institutional and Theoretical Economics*, 1990.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *International Conference on Computer Vision*, 2017.

F. E. Shamout, T. Zhu, P. Sharma, P. J. Watkinson, and D. A. Clifton. Deep interpretable early warning system for the detection of clinical deterioration. *IEEE Journal of Biomedical and Health Informatics*, 2020.

T. B. Sheridan and W. L. Verplank. Human and computer controlof undersea teleoperators. Technical report, MIT Man-Machine Systems Laboratory, Cambridge, MA, 1978.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *ICLR Workshop*, 2014.

S. Tonekaboni, M. Mazwi, P. Laussen, D. Eytan, R. Greer, S. D. Goodfellow, A. Goodwin, M. Brudno, and A. Goldenberg. Prediction of Cardiac Arrest from Physiological Signals in the Pediatric ICU. In *Machine Learning for Healthcare Conference*, 2018.

S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: Contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134*, 2019.

P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, et al. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 2020.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 2007.

B. Wilder, E. Horvitz, and E. Kamar. Learning to complement humans. *arXiv preprint arXiv:2005.00582*, 2020.

Supplementary Material: A Bandit Model for Human-Machine Decision Making with Private Information and Opacity

A Proofs of theorems in the main paper

A.1 Additional definitions

Let $H_{1,t} \in (\mathcal{X}\times\mathcal{R}\times A\times[0,1])^t$ and $H_{2,t} \in (\mathcal{R}\times\mathcal{Z}\times A\times[0,1])^t$ be the histories of Player 1 and Player 2, respectively. Let $\mathcal{D}(X)$ denote the space of probability distributions over a space X, and $\mathcal{F}(X)$ the set of all finite subsets of X. An algorithm A is a pair $A = (A_1, A_2)$ of two collections of measurable functions $A_1 = (A_{1,t})_{t=1}^T$ and $A_2 = (A_{2,t})_{t=1}^T$. For $t = 1$, we have $A_{1,1} : \mathcal{F}(\mathcal{R}^{\mathcal{X}})\times\mathcal{X} \to \mathcal{D}(\mathcal{R})$ and $A_{2,1} : \mathcal{F}(\mathcal{A}^{\mathcal{R}\times\mathcal{Z}})\times\mathcal{R}\times\mathcal{Z} \to \mathcal{D}(\mathcal{A})$. For $t = 2,\dots,T$, we have $A_{1,t} : \mathcal{F}(\mathcal{R}^{\mathcal{X}}) \times \mathcal{X} \times H_{1,t-1} \to \mathcal{D}(\mathcal{R})$ and $A_{2,t} : \mathcal{F}(\mathcal{A}^{\mathcal{R}\times\mathcal{Z}}) \times \mathcal{R} \times \mathcal{Z} \times H_{2,t-1} \to \mathcal{D}(\mathcal{A})$. In Section 5, we allow Player 2 to tell Player 1 which policy to use. This means that there is an additional collection of measurable functions $(A_{3,t})_{t=1}^T$ with $A_{3,1} : \mathcal{F}(\mathcal{A}^{\mathcal{R}\times\mathcal{Z}}) \to \mathcal{D}(\{1,\dots,N_1\})$ and $A_{3,t} : \mathcal{F}(\mathcal{A}^{\mathcal{R}\times\mathcal{Z}}) \times H_{2,t-1} \to \mathcal{D}(\{1,\dots,N_1\})$ for $t = 2,\dots,T$. These functions specify the (possibly randomized) policies that Player 2 tells Player 1 to use. A_1 consists of the fixed functions that implement the said policy choices for the first player. Additionally, the history of Player 2 and domain of functions in A_2 contain the policy that Player 1 was told to use.

A.2 Proof of Proposition 1

Proof. Without private information and opacity, the two players can perform actions that are equivalent to EXP4 run on the joint policy space Π (the EXP4 Algorithm is reproduced in Supplement Figure 4). Note that without opacity, both players have access to Π. Since $x_t = z_t$, they are also able to evaluate $\pi(x_t, z_t)$ for all $\pi \in \Pi$. Hence, a trivial solution would be that the second player ignores the recommendations made by the first player and simply performs EXP4. The result then follows from the standard analysis of EXP4 (Lattimore and Szepesvari, 2019, Theorem 18.1). A solution more in line with the interaction in our model would be that the first player recommends, in each round, r_t according to $\mathbb{P}(r_t = r) = q_{tr}$ where

$$q_{tr} = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} Q_{t,ij} 1_{\{f_i(x_t)=r\}}.$$

Here $Q_t \in \mathbb{R}^{N_1\times N_2}$ is the matrix maintained by EXP4 as described in Supplement Figure 4. The second player would then choose a_t according to $\mathbb{P}(a_t = k) = p_{tk}$ with

$$p_{tk} = \frac{\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} Q_{t,ij} 1_{\{f_i(x_t)=r_t \wedge g_j(r_t,z_t)=k\}}}{q_{t,r_t}},$$

i.e. there is a policy $g_t \in \Pi_2$ s.t. $a_t = g_t(r_t, z_t)$, while the action is again chosen exactly as in EXP4. □

A.3 Proof of Proposition 2

Proof. Both players privately label their policies from $0,\dots,N_1-1$ and $0,\dots,N_2-1$. Before the game starts, both players agree on a deterministic strategy for solving an N-armed stochastic bandit problem. In round t, where arm $0 \le i \le N-1$ is to be pulled in the N-armed stochastic bandit problem, for $i = a\cdot N_2 + b$ with $0 \le b < N_2$, Player 1 plays policy a and Player 2 plays policy b. Since a deterministic strategy determines the next arm to be pulled solely on the basis of past pulled arms and obtained rewards, both players know which of the N arms is to be pulled in each round. Agreeing on MOSS (Minimax Optimal Strategy in the Stochastic case), a variant of UCB, allows the two players to bound the minimax regret by $25\sqrt{TN}$ (Audibert and Bubeck (2010), Theorem 24). □

A.4 Proof of Theorem 3

A.4.1 Proof with private information

Proof. The idea is to construct a decision problem where the first player has to solve an N_1-armed stochastic Bernoulli bandit. The result then follows from the lower bound for stochastic Bernoulli bandits (e.g. Exercise 15.4 in Lattimore and Szepesvari (2019)). Note that Player 2 has only a single policy, i.e. $\Pi_2 = \{g\}$. Thus, the assumption that Player 2 only plays actions that are suggested by policies in Π_2 effectively fixes A_2, the algorithm of the second player.

Let $(X_{1t}, \dots, X_{N_1,t}) \in \{0,1\}^{N_1}$ be the payoffs associated with an N_1-armed stochastic Bernoulli bandit in round t. By assumption $K = 2$, so $A = \{1, 2\}$. Player 1 does not need to receive any context, so let $\mathcal{X} = \{\emptyset\}$. Choose $\mathcal{R} = \{1, \dots, N_1\}$ and $\Pi_1 = \{f_i | f_i = i, i = 1, \dots, N_1\}$. That is Player 1 has N_1 policies, and policy f_i constantly suggests recommendation i. In effect, recommendations and policies are really the same, namely the arms of a stochastic bandit. Let $\mathcal{Z} = \{0,1\}^{N_1}$ and $\Pi_2 = \{g\}$ with $g(r, z) = 1 + z_r$. For simplicity, let the payoff of Action 1 be 0 in all rounds. Conversely, let the payoff Action 2 be 1 in all rounds. Let the context vector z_t of Player 2 be given by the payoffs associated with the Bernoulli bandit, i.e. $z_t = (X_{1t}, \dots, X_{N_1,t})$.

In round t, when arm $i \in \{1, \dots, N_1\}$ of the Bernoulli bandit has payoff X_{it}, Player 2 assigns recommendation i to action $1 + X_{it}$. This results in a reward of X_{it}. Thus, in round t, where Player 1 chooses recommendation $r_t \in \{1, \dots, N_1\}$, the observes reward is $X_{r_t,t}$. To sum up, in every round, Player 1 incurs the reward of one of the arms of the Bernoulli bandit, and this arm can be freely chosen by choosing the recommendation. Since z_t is not observed by Player 1, the payoffs of all other arms of the Bernoulli bandit remain unknown. Every algorithm for Player 1 gives rise to an algorithm for stochastic Bernoulli bandits and vice-versa, and we obtain the lower bound. □

A.4.2 Proof with opacity

Proof. As above, let $A = \{1, 2\}$ and $\mathcal{R} = \{1, \dots, N_1\}$. Let $(X_{1t}, \dots, X_{N_1,t}) \in \{0,1\}^{N_1}$ be the payoffs associated with an N_1-armed stochastic Bernoulli bandit in round t. Now, in every round, both players receive the same context vector $x \in \{1, \dots, M\}$. The recommendations of policies of Player 1 are as before and independent of the context vector, $\Pi_1 = \{f_i | f_i = i, i = 1, \dots, N_1\}$.

The important part is the policy of Player 2, which is based on a function $\hat{g} : \{1, \dots, M\} \to \{0,1\}^{N_1}$. Instead of obtaining the payoffs of the Bernoulli bandit directly as contextual information, Player 2 now uses the private function $\hat{g}$ to obtain these payoffs from x. Naturally, $\hat{g}$ is not known to Player 1. As above, the policy of Player 2 is given by $g(r, x) = 1 + \hat{g}(x)_r$ and action payoffs are fixed to 0 and 1.

Let the context vector be uniformly distributed over $\{1, \dots, M\}$. We have to make sure that the same context vectors do not appear too often, since otherwise the first player could start to infer the payoffs associated with them. By choosing M large enough, context vectors up to time T are unique with probability arbitrarily close to 1.

We still have to specify how to choose $\hat{g}$ as a function from $\{1, \dots, M\}$ to $\{0,1\}^{N_1}$. For N_1 and M fixed, there are only finitely many of these functions. In order to realize a single desired Bernoulli bandit, draw $\hat{g}$ according to the probability distribution $\hat{\mathcal{D}}$ given by

$$\mathbb{P}_{\hat{\mathcal{D}}}(\hat{g}) = \prod_{i=1}^{M} \mathbb{P}\Big((X_{11}, \dots, X_{N_1,1}) = \hat{g}(i)\Big).$$

In other words, for all $i = 1, \dots, M$, the distribution of $\hat{g}(i)$ over $\{0,1\}^{N_1}$ is exactly that of the Bernoulli bandit.

By the same argument as in the prove with unknown context, if $\hat{g}$ is drawn according to $\hat{\mathcal{D}}$, Player 1 has to solve the Bernoulli bandit given by $(X_{1t}, \dots, X_{N_1,t})$. Now recall that the minimax regret is given by

$$R_T = \inf_{A_1} \sup_{\mathcal{D}} \sup_{|\Pi_1|=N_1} \sup_{|\Pi_2|=1} \mathrm{Reg}_T \,.$$

In particular,

$$\sup_{|\Pi_2|=1} \mathrm{Reg}_T \geq \sup_{\hat{\mathcal{D}}} \mathbb{E}_{\hat{g}\sim\hat{\mathcal{D}}}\Big[\,\mathrm{Reg}_T\,\Big],$$

EXP4
Parameters: $\eta > 0$, $\gamma > 0$
Initialization: Vector $Q_1 \in [0,1]^N$ with $Q_{1i} = \frac{1}{N}$
For each $t = 1, ..., T$

1. Receive context x_t
2. Choose action a_t according to $p_{tk} = \sum_{i=1}^{N} Q_{ti} 1_{\{\pi_i(x_t)=k\}}$.
3. Receive reward y_t and estimate $\hat{y}_{tk} = 1 - \frac{1_{\{a_t=k\}}}{p_{tk}+\gamma}(1-y_t)$
4. Propagates rewards to experts $\hat{Y}_{ti} = \hat{y}_{t,\pi_i(x_t)}$
5. Player 2 updates Q_t using exponential weighting

$$Q_{t+1,i} = \frac{\exp\left(\eta \hat{Y}_{ti}\right) Q_{ti}}{\sum_j \exp\left(\eta \hat{Y}_{tj}\right) Q_{tj}}$$

Figure 4: EXP4. Adapted from Algorithm 11 in Lattimore and Szepesvari (2019).

which shows the lower bound in terms of the minimax expected regret for N_1-armed stochastic Bernoulli bandits. □

A.5 Proof of Theorem 4

Proof. Recall the EXP4 algorithm, reproduced in Supplement Figure 4. The idea of the proof is as follows. In P2-EXP4, Player 2 maintains a probability distribution over the space of all policy combinations Π and performs importance-weighted updates. Player 2 does not know the policy space and context of Player 1. Therefore, in every round, he only obtains information on policy combinations where f_{i_t}, the function that the first player actually played, is present. This restricts Player 2 and does not allow him to perform the sames updates as EXP4. However, assume that all policy combinations where f_{i_t} is not present had suggested different actions than the policy combinations where f_{i_t} is present. In this case, the updates in P2-EXP4 would be equivalent to the updates of EXP4. Therefore, we now construct a bandit problem where two different policies of Player 1 never suggest the same action, and show that Algorithm 1 is equivalent to EXP4 on this related bandit problem.

Consider the adversarial contextual bandit problem with KN_1 actions and policy space

$$\begin{aligned}\tilde{\Pi} = \Big\{ h_{i,j} \,\big|\, & i = 1, \dots, N_1,\ j = 1, \dots, N_2, \\ & h_{i,j} : \mathcal{X} \times \mathcal{Z} \to \{1, \dots, KN_1\}, \\ & h_{i,j}(x,z) = (i-1)K + g_j(f_i(x), z) \Big\}.\end{aligned}$$

This policy space consists of N policies, and there exists a natural bijection I between $\tilde{\Pi}$ and Π given by $h_{i,j} \mapsto g_j(f_i(\cdot), \cdot)$. Let the adversarial payoffs of this new problem be a function of the (adversarial or i.i.d.) payoffs of the original problem, namely

$$\tilde{x}_t = (x_t, z_t)$$

and

$$\tilde{Y}_t(k) = Y_t\Big(1 + ((k-1) \bmod K)\Big),$$

for all $t = 1, \dots, T$ and $k = 1, \dots, KN_1$. Here $Y_t \in [0,1]^A$ contains the payoffs of the original problem, and $\tilde{Y}_t \in [0,1]^{\{1,\dots,KN_1\}}$ the payoffs of the new problem. By construction,

$$\tilde{Y}_t\Big(h_{i,j}(\tilde{x}_t)\Big) = Y_t\Big(g_j(f_i(x_t), z_t)\Big).$$

Therefore,

$$\max_{\tilde{\pi}\in\tilde{\Pi}} \sum_{t=1}^{T} \tilde{Y}_t\Big(\tilde{\pi}(\tilde{x}_t)\Big) = \max_{\pi\in\Pi} \sum_{t=1}^{T} Y_t\Big(\pi(x_t, z_t)\Big). \tag{2}$$

We are now going to show that P2-EXP4 is equivalent to EXP4(-IX) on this adversarial contextual bandit problem. In this proof, we denote all variables related this problem and EXP4 with a $\sim$. For example, $\tilde{a}_t$ is the action chosen by EXP4 in round t, resulting in a payoff of $\tilde{y}_t$. Since both P2-EXP4 and EXP4 are randomized, equivalence means that there exists a coupling of the random variables drawn by both algorithms under which, in all rounds, the probability distribution Q_t maintained by P2-EXP4 is the probability distribution $\tilde{Q}_t$ maintained by EXP4 (with respect to bijection I), $\tilde{a}_t = (i_t - 1)K + a_t$ and $\tilde{y}_t = y_t$.

We proceed by induction over t. The induction hypothesis is that equivalence holds up to round t. This is obviously true in the first round since both Q_t and $\tilde{Q}_t$ are initialized to be uniform. In round t, EXP4 chooses an action $\tilde{a}_t \in \{1, \dots, KN_1\}$. This action $\tilde{a}_t$ can be uniquely written as $\tilde{a}_t = (\hat{i}_t - 1)K + \hat{a}_t$ for some $\hat{i}_t \in \{1, \dots, N_1\}$ and $\hat{a}_t \in \{1, \dots, K\}$. By construction, it is exactly policies $h_{i,1}, \dots, h_{i,N_2}$ that suggest actions

$$(i-1)K + 1, \dots, iK.$$

Hence,

$$\mathbb{P}\left(\hat{i}_t = i\right) = \sum_{j=1}^{N_2} Q_{t,ij} = \mathbb{P}(i_t = i),$$

where the first equality is due to the induction hypothesis and the second due to the definition of q_{ti} in P2-EXP4. Since they have the same distribution, $\hat{i}_t$ and i_t can be perfectly coupled. Additionally, and already subject to this coupling,

$$\begin{aligned}
\mathbb{P}(\hat{a}_t = k \,|\, i_t = i) &= \frac{\mathbb{P}(\tilde{a}_t = (i-1)K + k)}{\mathbb{P}(i_t = i)} \\
&= \frac{\sum_{j=1}^{N_2} Q_{t,ij} 1_{\left\{h_{i,j}(\tilde{x}_t) = (i-1)K + k\right\}}}{\sum_{j=1}^{N_2} Q_{t,ij}} \\
&= \mathbb{P}(a_t = k \,|\, i_t = i)
\end{aligned}$$

where we used the definition of a_t in Equation (1) of the main paper and the fact that

$$h_{i,j}(\tilde{x}_t) = (i-1)K + k \iff g_j(f_i(x_t), z_t) = k.$$

Thus, conditional on i_t, $\hat{a}_t$ and a_t have the same distribution. Therefore, $\hat{a}_t$ and a_t can be perfectly coupled, too, and we arrive at $\tilde{a}_t = (i_t - 1)K + a_t$. From the definition of $\tilde{Y}_t$, it follows that $\tilde{y}_t = y_t$.

It remains to show that the update $Q_t \to Q_{t+1}$ in P2-EXP4 agrees with EXP4. We have to show that $\hat{Y}_t$ in P2-EXP4 agrees with the importance-weighted reward estimates of EXP4. We distinguish three cases. The first case is $i = i_t$ and $g_j(f_i(x_t), z_t) = a_t$. Here it holds that

$$\begin{aligned}
\hat{Y}_{t,ij} &= 0 + 1 - \frac{1}{q_{t,i_t}\, p_{t,a_t} + \gamma}(1 - y_t) \\
&= 1 - \frac{1}{\tilde{p}_{tk} + \gamma}(1 - \tilde{y}_t).
\end{aligned}$$

The second case is $i = i_t$ and $g_j(f_i(x_t), z_t) \neq a_t$. Here it holds that $\hat{Y}_{t,ij} = 0 + 1 = 1$. The third case is $i \neq i_t$. Here it holds that $\hat{Y}_{t,ij} = 1 + 0 = 1$, too. In all three cases, the update agrees exactly with EXP4.

We have shown that $\sum_{t=1}^{T} y_t = \sum_{t=1}^{T} \tilde{y}_t$. Subtracting this from (2), we see that

$$\max_{\pi\in\Pi} \sum_{t=1}^{T} Y_t\Big(\pi(x_t, z_t)\Big) - \sum_{t=1}^{T} y_t = \max_{\tilde{\pi}\in\tilde{\Pi}} \sum_{t=1}^{T} \tilde{Y}_t\Big(\tilde{\pi}(\tilde{x}_t)\Big) - \sum_{t=1}^{T} \tilde{y}_t. \tag{3}$$

From the analysis of EXP4, e.g. from Theorem 18.1 in Lattimore and Szepesvari (2019), we know that

$$\mathbb{E}\left(\max_{\tilde{\pi}\in\tilde{\Pi}}\sum_{t=1}^{T}\tilde{\pi}(\tilde{x}_t)-\sum_{t=1}^{T}\tilde{y}_t\right)\leq\sqrt{2TKN_1\ln(N_1N_2)}$$

for $\gamma=0$ and $\eta=\sqrt{2\log(N_1N_2)/(TKN_1)}$, which implies the desired bound. □

A.6 Proof of Theorem 6

Proof. Assume that $|\mathcal{R}|<\infty$, otherwise the bound is vacuous. Let f_1 and f_2 be two policies of Player 1. In general, the expected regret under f_1 and f_2 depends on the policy choice of Player 2. Specifically, there might be g_1 and g_2 such that $Y\big(g_1(f_1(x),z)\big)>Y\big(g_1(f_2(x),z)\big)$ and $Y\big(g_2(f_1(x),z)\big)<Y\big(g_2(f_2(x),z)\big)$. Let $\pi_\star=(g_\star,f_\star)$ be an optimal policy combination. Under policy space independence, the quantities

$$\text{Reg}(f)=Y\big(g(f_\star(x),z)\big)-Y\big(g(f(x),z)\big)$$

and

$$\text{Reg}(g)=Y\big(g_\star(f(x),z)\big)-Y\big(g(f(x),z)\big)$$

are well-defined. Moreover,

$$Y\big(\pi_\star\big)-Y\big(g(f(x),z)\big)=\text{Reg}(g)+\text{Reg}(f).$$

That both players explore independently using EXP4 means the following. Player 2 uses EXP4 on A with $\eta_1=\sqrt{2\log(N_2)/(TK)}$ and $\gamma_1=0$. Player 1 considers recommendations as actions and uses EXP4 on $\mathcal{R}$ with $\eta_2=\sqrt{2\log(N_1)/(T|\mathcal{R}|)}$ and $\gamma_2=0$. In round t, there exist policies f_{i_t} and g_{j_t} such that $r_t=f_{i_t}(x_t)$ and $a_t=g_{j_t}(f_{i_t}(x_t),z_t)$. Player 1 solves the adversarial contextual bandit problem with context x_t, action space $\mathcal{R}$ and policy space Π_1. Player 2 solves the adversarial contextual bandit problem with context (r_t,z_t), action space A and policy space Π_2. Player 1 provides adversarial context for Player 2, and Player 2 provides adversarial payoff for Player 1. Because of policy space independence, this independent exploration strategy also controls the joint expected regret.

First note that i_t and j_t are functions of the history and can be considered drawn before the tuple (x_t,z_t,Y_t). The expected regret in round t is given by

$$\mathbb{E}_{(x_t,z_t,Y_t)\sim\mathcal{D}}\Big[Y_t(g_\star(f_\star(x_t),z_t)-Y_t(g_{j_t}(f_{i_t}(x_t),z_t))\Big]=Y\big(g_\star(f_\star(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big).$$

Making use of policy space independence, the right hand side can be rewritten as

$$\begin{aligned}&Y\big(g_\star(f_\star(x),z)\big)-Y\big(g_\star(f_{i_t}(x),z)\big)+Y\big(g_\star(f_{i_t}(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big)\\&=Y\big(g_{j_t}(f_\star(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big)+Y\big(g_\star(f_{i_t}(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big).\end{aligned}$$

Summing over t, the expected regret is given by

$$\begin{aligned}\text{Reg}_T=&\sum_{t=1}^{T}\Big[Y\big(g_{j_t}(f_\star(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big)\Big]\\&+\sum_{t=1}^{T}\Big[Y\big(g_\star(f_{i_t}(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big)\Big].\end{aligned}$$

The first sum is the expected regret in the adversarial contextual bandit problem of the first player. The second sum is the expected regret in the adversarial contextual bandit problem of the second player. From the analysis of EXP4, e.g. from Theorem 18.1 in Lattimore and Szepesvari (2019), we obtain

$$\sum_{t=1}^{T}\Big[Y\big(g_{j_t}(f_\star(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big)\Big]\leq\sqrt{2T|\mathcal{R}|\ln N_1}$$

and

$$\sum_{t=1}^{T}\Big[Y\big(g_\star(f_{i_t}(x),z)\big)-Y\big(g_{j_t}(f_{i_t}(x),z)\big)\Big]\leq\sqrt{2TK\ln N_2},$$

which implies the desired bound. □

B Fixed rules that allocate decisions result in policy space independence

In this section we formalize the example given in Section 6.2. We show that fixed rules that allocate decisions to either the human or the machine result in policy space independence. Let $\mathcal{R} = A$ (treatment recommendations), $D : \mathcal{Z} \mapsto \{0, 1\}$ (the human decides who decides), $\tilde{\Pi}_2 : \mathcal{Z} \to A$ (the human's own decision rules) and

$$\Pi_2 = \{g | g = D(z)\tilde{g}(z) + (1 - D(z))r,\ \tilde{g} \in \tilde{\Pi}_2\}.$$

Here $r = f(x)$ where $f \in \Pi_1$ is the decision rule used by the machine. Now, for all $f \in \Pi_1$ and $g \in \Pi_2$, and all distributions $\mathcal{D}$,

$$\begin{aligned} Y(g(f(x), z)) &= \mathbb{E}_{(x,y,z)\sim\mathcal{D}}[Y(g(f(x), z))] \\ &= \mathbb{P}(D(z) = 0)\mathbb{E}[Y(g(f(x), z))|D(z) = 0] \\ &\quad + \mathbb{P}(D(z) = 1)\mathbb{E}[Y(g(f(x), z))|D(z) = 1] \\ &= \mathbb{P}(D(z) = 0)\mathbb{E}[Y(f(x))|D(z) = 0] + \mathbb{P}(D(z) = 1)\mathbb{E}[Y(\tilde{g}(z))|D(z) = 1]. \end{aligned}$$

Thus,

$$\begin{aligned} Y(g_1(f_1(x), z)) - Y(g_1(f_2(x), z)) &= \mathbb{P}(D(z) = 0)\mathbb{E}[Y(f_1(x)) - Y(f_2(x))|D(z) = 0] \\ &= Y(g_2(f_1(x), z)) - Y(g_2(f_2(x), z)) \end{aligned}$$

for all $f_1, f_2 \in \Pi_1$, $g_1, g_2 \in \Pi_2$ and all distributions $\mathcal{D}$. In the key step of the derivation, we did not use the fact that D was a (measurable) function of Z. Indeed, the sample space can be partitioned with respect to any event D.

C Fixed second player in Theorem 3

In Theorem 3, we assumed that Player 2 only plays actions that are suggested by policies in Π_2. We are convinced that this assumption can be dropped if the problem instances in the respective proofs are modified in the following two ways.

First, in every round, the relation between policies and recommendations should be entirely random. Concretely, let the policies of Player 1 depend on a context vector $x \in \{1, \dots, M\}$. In every round, let x_t be uniform on $\{1, \dots, M\}$. Moreover, choose M large enough such that every context vector occurs at most once up to time T. For every $x \in \{1, \dots, M\}$, randomly draw a permutation $\pi_x \in S_{N_1}$. Choose the policy space of the first player such that given context x_t, policy f_i recommends $\pi_{x_t}(i)$. In effect, up to time T, the policies of Player 1 make random recommendations, subject to the constraint that all recommendations be different.

Second, in every round, it should be entirely random which action gives the payoff of 1. Thus, for every $x \in \{1, \dots, M\}$, randomly drawn one action to give a payoff of 1, and set the payoff of the other action to 0.

In the first proof of Theorem 3 (unknown context), permute the context vector z of Player 2 so that every policy still gets the same payoff as it would in the original construction (considering both π_x and the permuted payoffs). In the second proof of Theorem 3 (unknown policy), let the policy of Player 2 encode the appropriately permuted context vector.

Intuitively, if Player 2 knew which policies suggested which recommendations, Player 2 could effectively learn for Player 1. This is since Player 2 does always know the relation between recommendations and actions. In the given problem instance, the relation between policies and recommendations is impossible to know, at least up to time T.

D Problem instance for Conjecture 7

In this section we give a problem instance for Conjecture 7. We conjecture that it is a worst-case instance for which the lower bound stated in Conjecture 7 holds.

In every round, let one action give a payoff of 0 and the other a payoff of 1. Randomly decide in every round which action gives the payoff of 1. Choose the context vector and policy class of Player 1 such that he uniformly receives one of the 2^{N_1} possible expert recommendations in every round. Ahead of time, select a policy of Player 1 and Player 2, respectively (the optimal policies). In every round, a policy for Player 2 gives a map $\mathcal{R} \to A$. With $\mathcal{R} = A = \{0, 1\}$, there are 4 possible maps that we denote by $(0,0)$, $(1,0)$, $(0,1)$ and $(1,1)$. Here $(1,0)$ is the map that maps recommendation 0 to Action 1 and recommendation 1 to action 0. For the optimal policy of Player 2, let the relation between recommendations and actions be such that every policy of Player 1 except the optimal policy receives an expected payoff of 0.5, and the optimal policy receives an expected payoff of $0.5 + \Delta$. This can be achieved as follows. In round t, where the optimal policy recommends r_t, map recommendation r_t to the action with a payoff of 1 with probability $0.5 + \Delta$. Similarly, map recommendation $1 - r_t$ to the action with a payoff of 1 with probability $0.5 - \Delta$. Note that since context vectors of Player 1 are drawn uniformly at random, each policy makes the same recommendation as the optimal policy exactly half of the time. For all other policies of Player 2, draw one of the 4 possible maps from recommendations to actions according to

$$\mathbb{P}\big((0,0)\big) = 0.25 - \Delta^2, \qquad \mathbb{P}\big((1,0)\big) = 0.25 + \Delta^2$$
$$\mathbb{P}\big((0,1)\big) = 0.25 + \Delta^2, \qquad \mathbb{P}\big((1,1)\big) = 0.25 - \Delta^2.$$

This distribution is chosen such that all other policies have the same marginal distribution over the maps from recommendations to actions as the optimal policy.

Let us quickly outline why we think that this is a difficult problem instance. Imagine that in every round, both players choose a policy according to some decision rule. If both players choose their optimal policy, the expected payoff is $0.5 + \Delta$. Should any of the two players not choose their optimal policy, the expected payoff is 0.5 (for all policy choices of the other player, also the optimal policy). Now consider what happens in the first round of the game. Assume that both players choose a policy uniformly at random (uniformly choosing recommendations, maps or actions does not reveal any information at all). Then, the expected payoff of the optimal policies of both players is $0.5 + \frac{\Delta}{N_1}$. Thus, at least in the first round, the magnitude of the signal is $\frac{\Delta}{N_1}$, while the magnitude of the regret is Δ. While the magnitude of the signal increases as the other player starts to identify the optimal policy, this strongly suggests that the regret does not scale logarithmically in N_1.

E Online learning and repeated supervised learning

In this section we give some more detail on why online learning is the correct approach to study human-machine decision making. Indeed, full online learning, as studied in our paper, is the most general and unrestricted way to understand how decisions evolve over time. This is despite the fact that machine learning algorithms are often not deployed in an online fashion. One reason for the latter is that online learning entails exploration which usually requires informed consent of the individuals who are impacted by the decisions.

In practice, machine learning algorithms are usually trained on a historical dataset. In a human-machine decision making context, one would then evaluate how well humans perform with the trained algorithm, or a given number of trained algorithms. This might include some form of training for human decision makers plus a randomized controlled trial. If one finds that a given system performs sufficiently well, it might be deployed . Although this procedure is not an explicit online learning procedure, it is subject to the same limitations as online learning, at least insofar as coordination between the two decision makers is concerned. Viewed through the lens of our model, it could be interpreted as follows. First, the human makes a number of decisions, ignoring the machine (this produces the historical dataset). Second, the machine decides on a number of candidate policies (this is the supervised learning part). Third, the human tries to learn how to interpret the candidate policies of the machine (as in Section 5). A slightly different interpretation would be to consider the result of supervised learning as the initial policy space of the machine. More generally, *full online learning is the theoretical limit* of all sorts of procedures that iterate between machine learning on a given dataset, evaluating how well something works with humans in a real-world setting, collecting a bigger dataset, retraining our model in order to improve performance, evaluating again with humans, and so on. Importantly, online learning covers the scenario where we continuously collect data as a given system is running and then re-train it, say, once a year. In fact, full online learning places as few constraints on learning as possible. For example, re-training a system only at fixed intervals introduces an additional constraint often referred to as batching.

F Opacity and implicit communication

In this section we discuss a theoretical subtlety that arises due to the way in which we set the problem up. This gives more details on the discussion at the end of Section 4 and in Supplement C.

We model opacity by keeping knowledge about the policy spaces to the respective players. As is apparent from the definition of the minimax regret in section 2.1, both players first fix the way in which they want to approach the problem (the algorithm), then get to see the respective policy spaces. Importantly, we decided to place *no* restrictions on the algorithm that the two players might run. This is because the algorithm is part of the solution and not part of the problem. It also keeps our work closely aligned with the extant literature on online learning. This assumption has, however, a subtle consequence. Namely, the algorithms of both players can be arbitrarily well adapted. In a sense, before the game starts, the two players are allowed to get together in order to discuss how the problem might be approached. During the game, players might then try to implicitly encode information about policy spaces and context in actions and recommendations – according to some protocol that they agreed upon in advance.

With regard to our original research question, elaborate implicit communication protocols between the two players are of course unrealistic and even violate the idea of opacity. After all, it is implausible that a computer program and a human decision maker would communicate with such means. In this regard, note that we ruled out implicit communication protocols in Theorem 3 by assuming that the second player follows his one (and only) policy.

From a theoretical perspective, the question of whether implicit communication protocols would make a difference nevertheless remains interesting (Bubeck et al., 2020). As we argue in Supplement C, we believe that this is not the case.

2.2 Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts

Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts

Sebastian Bordt
sebastian.bordt@uni-tuebingen.de
University of Tübingen, Germany

Michèle Finck
michele.finck@uni-tuebingen.de
University of Tübingen, Germany

Eric Raidl
eric.raidl@uni-tuebingen.de
University of Tübingen, Germany

Ulrike von Luxburg
ulrike.luxburg@uni-tuebingen.de
University of Tübingen, Germany

ABSTRACT

Existing and planned legislation stipulates various obligations to provide information about machine learning algorithms and their functioning, often interpreted as obligations to "explain". Many researchers suggest using post-hoc explanation algorithms for this purpose. In this paper, we combine legal, philosophical and technical arguments to show that post-hoc explanation algorithms are unsuitable to achieve the law's objectives. Indeed, most situations where explanations are requested are adversarial, meaning that the explanation provider and receiver have opposing interests and incentives, so that the provider might manipulate the explanation for her own ends. We show that this fundamental conflict cannot be resolved because of the high degree of ambiguity of post-hoc explanations in realistic application scenarios. As a consequence, post-hoc explanation algorithms are unsuitable to achieve the transparency objectives inherent to the legal norms. Instead, there is a need to more explicitly discuss the objectives underlying "explainability" obligations as these can often be better achieved through other mechanisms. There is an urgent need for a more open and honest discussion regarding the potential and limitations of post-hoc explanations in adversarial contexts, in particular in light of the current negotiations of the European Union's draft Artificial Intelligence Act.

KEYWORDS

Explainability, Transparency, Regulation, Artificial Intelligence Act, GDPR, Counterfactual Explanations, SHAP, LIME

ACM Reference Format:
Sebastian Bordt, Michèle Finck, Eric Raidl, and Ulrike von Luxburg. 2022. Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts . In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3531146.3533153

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
FAccT '22, June 21–24, 2022, Seoul, Republic of Korea
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533153

1 INTRODUCTION

Explainability is one of the concepts dominating debates about the ethics and regulation of machine learning algorithms. Intuitively, requests for explainability are reactions to the prevalent unease about machine learning algorithms, including concerns regarding discrimination, biases, manipulation, and data protection. The fact that machine learning systems are often "black boxes" is considered a major hurdle towards their implementation, supervision and control, and explainability is often praised as a remedy against such risks. Existing legislation such as the EU General Data Protection Regulation ('GDPR') has sometimes been interpreted as containing a "right for explanation". The draft Artificial Intelligence Act, a piece of proposed EU legislation, alludes to explainability but does, in its current form, not make clear whether and when exactly explainability is legally required. On the technical side, explainability has evolved into its own field of research [33]. The current machine learning literature knows two different approaches towards explainability. One approach is to build machine learning models that are constrained to be "inherently interpretable" [42]. The other approach is to use any machine learning model, even a "back-box", and then employ any of an increasing number of approaches in order to "explain" the behavior of the black-box after the decision has been made ("post-hoc"). Because there exists no general way to summarize the entire behavior of a black-box model, these explanations are usually local, meaning that they only describe the behavior of the function for a single prediction or decision. The natural advantage of local post-hoc explanation methods, such as feature highlighting methods [30, 41] and counterfactual explanations [60], is that they place no constraints on model complexity and do not require model disclosure [7]. This has led a number of researchers to suggest that these methods might be able to comply with existing legal requirements [7, 60].

In this paper, we put forward an important distinction that has not yet been extensively discussed in the literature on explainable AI: whether the explanation's context is adversarial or cooperative. By "cooperative contexts" we broadly summarize situations where all involved parties have aligned interests. This includes model development and debugging, scientific discovery, and, to a degree, areas such as medical diagnosis. In a

cooperative context, the explanation provider and the explanation receiver share the same interests: to identify the most suitable and insightful explanation algorithm for the given problem. In "adversarial contexts", in contrast, parties have opposing interests. This is the case for example when a bank denies a customer a loan and the customer wants to contest the decision because it was discriminatory. Since the explanation provider anticipates that one might use the provided explanations to challenge the functioning of the system, the explanation provider does not have any incentive to provide "true" insights into the functioning of the system; but rather to render the internal functioning of the machine learning system incontestable. Indeed, it has been pointed out repeatedly that post-hoc explanation algorithms can be manipulated or cheated upon [5, 47, 48]. Many machine learning papers on explanation algorithms implicitly consider collaborative contexts where explanations are used to improve machine learning algorithms and can help developers to understand the biases of complex systems, or where they are used in an explorative spirit towards new scientific discoveries [63]. In contrast, the legal discussion focuses predominantly on adversarial scenarios. Here explainability is portrayed as a mechanism to add more transparency, fairness and accountability to AI, and post-hoc explanations are often seen as a technical tool to achieve these goals.

Combining insights from computer science, philosophy and law, we offer a critical multidisciplinary perspective on the usage of post-hoc explanations to achieve transparency and accountability obligations in adversarial contexts. We highlight the blurry legal landscape around explainability as well as the philosophical and technical limitations of post-hoc explanations. In Section 2 we introduce different scenarios – cooperative and adversarial – under which an external examiner might audit a black-box and its generated explanations. We focus on adversarial scenarios – where the explanation provider has opposing interests to the explanation receiver – and local post-hoc explanations – where the explanation explains a single decision for one particular person. In Section 3 we argue that existing and planned legislation,specifically the GDPR and the EU Artificial Intelligence Act, can either be read as portraying explainability as one possible mechanism to achieve more transparency or as presenting it as a free-standing objective. We also highlight the current lack of legal certainty as to how existing legal norms around explainability ought to be interpreted and implemented. These issues have been the source of confusion and uncertainty. This is why we propose to capture the role of explainability by a discussion of its motivations: Explanations are thought to build trust, and also enable actions, such as debugging, contesting, recourse. In Section 4 we show from a philosophical and technical perspective that the goals associated with explainability are unlikely to be achieved by post-hoc explanations. The reason is that the truth assumptions under which explanations are expected to fulfill their legal goal are lacking in the adversarial context. To the contrary, due to the inherent geometric ambiguity of local post-hoc explanations, the explanation provider has a multitude of options to influence explanations in a subtle, undetectable way and to pick those that suit her goals. In Section 5 we show that testing explanations is also problematic. While at best we can test for internal consistency of the explanation with the decision, in more typical cases the explanations become redundant and we would better rely on testing decisions and predictions directly. In Section 6 we conclude and argue that there needs to be a deeper and more honest debate about what the underlying objectives of explainability obligations are. We also argue that one needs to be honest about the fact that using a black-box entails considerable discretion: Neither post-hoc explanation methods, nor regulation can completely compel the deployer of a black-box to align his interests with the public good. As such, if one is absolutely unwilling to award any discretion to the deployer of the black-box, the only solution is to forbid its deployment and favor inherently interpretable or otherwise constrained machine learning methods. The question under which circumstances the deployment of a black-box might still be admissible depends on our ability to examine and audit the black-box. How exactly this might be done is still an area for future research. We hope that our paper contributes to an open discussion regarding the (lack of) potential of post-hoc explanations in the context of the on-going negotiation of the Artificial Intelligence Act.

2 EXPLANATIONS IN COOPERATIVE AND ADVERSARIAL CONTEXTS

In this work we broadly distinguish between "cooperative" and "adversarial" explanation contexts. In a **cooperative context**, all parties involved in the process of building the system, providing explanations and using the system share the same goal: to create a system as good and supportive as possible. Prototypical examples are model debugging and scientific research. But also a company building a medical decision support system, say for skin cancer detection, will closely collaborate with the doctors who use it [53]. The company's goal would be to provide explanations that are as helpful as possible. The situation is very different in **adversarial contexts**, where parties do not share the same goal, such as in the oft-repeated example of a denial of a loan application. Here, the applicant and bank have opposing interests and incentives. Accordingly, should the bank be mandated to provide the applicant with an explanation, this explanation will be shaped by the bank's incentives and existing power asymmetries. For reasons that we outline below, the distinction between cooperative and adversarial contexts is crucial. In particular, we argue that local post-hoc explanations, which have a variety of use-cases in the cooperative scenario, are pointless or even harmful in adversarial contexts.

2.1 Parties involved in the adversarial explanation process

We consider adversarial explanation contexts where an AI decision system is used to make decisions about individuals. Prominent examples are university admissions, job and loan applications, or bail and sentencing decisions. Under existing and planned legislation, such as the EU Artificial Intelligence Act, the *creator of the system* ought to provide information about how the system comes to its decisions (see Section 3 below for a detailed discussion of the legal background). The creator of the system is the entity that has built the machine learning system and uses it to support decision making.[1] The creator could be a private company (such as a bank) or a public entity (such as a university). The *decision subject* is the person about whom the automated system makes a decision: the person who applies for a loan, or the person who applies to for university admission. After the decision has been communicated, the *explanation recipient* asks for an explanation, which is communicated by the *explanation provider*. The explanation recipient could be the decision subject herself, or an external *examiner* who is supposed to investigate the decisions or explanations on behalf of the decision subject or to defend her interests. The explanation provider is typically the creator of the system.[2]

2.2 Machine learning problem: Supervised learning, tabular data, point-wise post-hoc explanations

In our technical discussion, we assume that the inputs $x \in \mathbb{R}^d$ of a decision algorithm are given in **tabular** form. Each dimension of the input encodes a different property of a person, for example age, income, etc. Typically, the number of dimensions d is large: persons are described by dozens or hundreds of features. A machine learning algorithm is used to learn a **decision function** $f : \mathbb{R}^d \to \mathbb{R}$. The resulting decision $y = f(x)$ for input x could be a binary decision ("receives the loan" or "does not receive the loan") or a numeric risk score on which such a decision is based, as in the often discussed COMPAS algorithm to predict recidivism risk. We focus on *supervised machine learning*, where f is learned based on **training data** consisting of pairs $(x_1, y_1), ..., (x_n, y_n)$ with x_i the training points and y_i the training labels. An explanation algorithm E is an algorithm operating on a decision function with the purpose of explaining it. We focus on **local post-hoc explanation algorithms**: The explanation algorithm E gets queried with a data point x and the corresponding decision y, and produces an explanation $E(x, y)$. Internally, the algorithm has access to the decision function f, and in some cases also to the training data. The explanation $E(x, y)$ is supposed to explain why the decision function f came to decide y for x. The explanation can be in linguistic form. For example, "The low income of Mr. Smith was relevant for the refusal of the loan" or "Mr. Smith would have received the loan had his income been 10.000 Euros higher".

[1] The creator is mainly the developer. But since the developer develops the system for a user, their interests typically align. Hence we do not distinguish developer and user, and use the term "creator" instead.
[2] Similar distinctions were introduced by [52].

2.3 Explanation algorithms that fall into this framework

In this paper we consider local post-hoc explanation algorithms such as LIME, SHAP, and DiCE [30, 34, 41]. The explanations generated by these algorithms do not provide a global or holistic view of the decision function f but merely try to explain individual decisions $y = f(x)$. The often-cited advantage of these algorithms is that they work, at least in principle, for *any* decision function [7, 41]. Different algorithms take different approaches as to what constitutes an explanation: LIME and SHAP provide *feature attributions* that aim to quantify the influence of the different input-features for the particular decision. Feature attributions correspond to the linguistic form "The low income of Mr. Smith was relevant for the refusal of the loan". Another approach is to provide *counterfactual explanations* [60]. These explanations are based on searching for a sufficiently close or the closest alternative point x' to the actual input point x that yields a decision $y' = f(x')$ that differs from the original decision $y = f(x)$. Comparing the two we can arrive at factors that are relevant to the decision [24]. The resulting counterfactual explanations have the linguistic form "Mr. Smith would have received the loan had his income been 10.000 Euros higher".

3 LEGAL FRAMEWORK: EXPLAINABILITY IN EU LAW

This paper argues that post-hoc explanation algorithms are unsuitable in adversarial contexts. Before we elaborate this from a philosophical and technical perspective (Section 4), it is important to understand the related legal framework. We focus on European Union law as the EU has often been a first-mover regarding the regulation of data and its analysis, and over time its legislation will likely inspire other jurisdictions (for a broader view, see [21]). Our analysis focuses on the draft Artificial Intelligence Act (AIA), a piece of proposed legislation that would be the first to specifically target AI. This pioneering approach would be a global blueprint for the regulation of AI. In its current form it creates different legal obligations for different AI applications on the basis of the perceived risks. The AIA would apply to general AI systems (Section 3.1). We also consider the General Data Protection Regulation (GDPR), which applies to the processing of personal data (Section 3.2). It will be seen that whereas EU law contains various obligations to provide information about a machine learning algorithm and its functioning, it remains unclear how these legal norms should be implemented from a technical perspective and whether explainability should be understood as a free-standing legal obligation or whether it should rather be seen as one of various mechanisms to achieve algorithmic transparency (Section 3.3). To better

understand the latter we also review their underlying rationales and objectives from a philosophical and legal perspective (Section 3.4).

3.1 The draft Artificial Intelligence Act (AIA)

The current draft of the AIA defines AI systems as "software (...) that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with". Generally, the AIA regulates AI on the basis of its perceived risk by introducing four different categories of AI. Most relevant to our discussion are the two categories of systems that are high-risk, as opposed to systems that are not high-risk (the remaining two categories are practices that are subject to qualified prohibitions, and a residual category of AI systems that includes law enforcement software, emotion recognition system, biometric categorisation systems and deep fakes) [54]. The stronger the risk, the heavier regulatory obligations apply, also regarding transparency and interpretability.

There are two categories of **high-risk AI systems**. First, AI systems that relate to products that are already subject to supranational harmonisation, namely AI systems intended to be used as a safety component of a product, which are themselves products covered by Union harmonising legislation or which are required to undergo third-party conformity assessments. Second, a list of systems that are currently considered to carry a high-risk such as, for instance, biometric identification systems, systems for the management and operation of critical infrastructure, those used in education and employment, some law enforcement systems as well as others (see further Art 3(1) of the draft AIA). Article 13 governs explainability for high-risk AI systems, which have to be "designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to *interpret* the system's output and use it appropriately". Furthermore, users (the entity deploying the AI) need to have access to instructions for use in an appropriate digital format that contains information about the characteristics, capabilities and limitations of performance, including information about the level of accuracy, robustness and cybersecurity, risks to health, safety or fundamental rights, specifications for the input data, expected lifetime of the AI system and necessary maintenance measures. Finally, human oversight must be ensured. These measures are designed to minimize risks to health, safety or fundamental rights. Human oversight shall either be (i) identified and built into the system by the provider before it is placed on the market or put into service, or (ii) identified by the provider before the system is placed on the market or put into service but only implemented by the user.

In its current version, the AIA would thus require that high-risk AI systems are sufficiently transparent to enable the interpretation of the system's output. Is this an explainability obligation? Recital 47 sheds some light on how to interpret these notions. It specifies that high-risk AI systems should be transparent to a "certain degree" to "address the opacity that may make certain AI systems incomprehensible to or too complex for natural persons". To this end, users "should be able to interpret the system output and use it appropriately" through the provision of "relevant documentation and instructions of use". *This does not read like an obligation to make systems explainable in the sense that the way in which data has been processed must be entirely traceable.* Rather, the AIA would require that an "interpretation" of the output must be facilitated through sufficient transparency. Importantly, this does not necessarily seem to imply that an absolute truth must be identified post-hoc (see Sections 4.1 and 4.2 below) but rather the overall functioning of the system and how it comes to an output. *The draft AIA leaves open the question of what transparency and interpretability imply from a technical perspective.* This certainly includes the elements listed in its Article 16 such as technical documentation, keeping logs or quality management systems. Article 13 leaves open whether there are additional requirements and what, exactly, interpretability requires from a technical perspective. If input data ought to be entirely traceable, "black-box" systems cannot be used in high-risk applications. This highlights that it is important to think about the objectives of transparency and explainability. If these can be achieved through alternative means, excluding black-box systems such as deep neural networks from high-risk scenarios (such as healthcare as devices falling under the Medical Devices Regulation qualify as high-risk) might unduly hinder innovation in important domains.

Article 52 AIA creates some general transparency obligations for **AI systems that are not high-risk**. These are general disclosure obligations such as to (i) inform users that they are interacting with an AI system unless this is obvious from context, (ii) users of an emotion recognition system or biometric categorization system shall inform natural persons exposed thereto, (iii) deep fakes must be disclosed as such. Some exceptions apply where the AI is used in the context of law enforcement. These are thus obligations of transparency that require disclosure that AI is used, as opposed to how it is used.

To summarize, *the draft AIA would thus not, in its current form, create a general explainability obligation for machine learning systems.* Such an obligation clearly is not foreseen in relation to AI systems that are not qualified as high-risk. Arguably, there is also no explainability obligation in relation to high-risk AI systems. Rather, what is required is transparency of the system's functioning and output generation. This transparency must make these elements interpretable but not necessarily amount to the provision of an explanation as it is commonly understood in computer science.

3.2 The General Data Protection Regulation (GDPR)

The GDPR creates some general transparency requirements that form part of the data controller's (the entity that determines the purposes and means of processing) general informational obligations vis-à-vis the data subject (the natural person that personal data relates to). In addition, it also contains a specific regime for "solely automated data processing". In contrast to the draft AIA, which creates vague obligations resting on the user, the GDPR creates specific rights for the individual subjected to such decisions.

Article 13 requires that data controllers provide specific information to data subjects where personal data is collected from them at the time of collection such as whether "automated decision-making" is used, and, if so, provide *"meaningful information about the logic involved*[3]*, as well as the significance and the envisaged consequences of such processing"*. Article 14(1)(h) creates the same obligation in cases where data is not directly collected from the data subject. Pursuant to Recital 62 this information does not have to be provided where it is redundant, or where compliance proves impossible or involves a disproportionate effort. The same wording can also be found in Article 15, which deals with the data subject's right to access data. Whereas Articles 13 and 14 relate to the pre-processing stage, data subjects can exercise their rights under Article 15 at any time, including after processing has taken place. This raises the question of whether – despite the identical wording of these provisions – Article 15 may substantively require something different when referring to the "logic" of the automated decision-making process.

There is no general right to an explanation under the GDPR. Some explainability requirements may, however, arise in respect of machine learning algorithms that produce legal effect or similarly significantly affect a data subject. Article 22 creates a qualified prohibition of "solely automated data processing", including profiling. This implies that such techniques can only be used in some circumstances, namely (i) where necessary to enter into or perform a contract between the data subject and controller, (ii) where it is authorized by law or where the data subject has provided explicit consent. In these circumstances automated processing can take place, but the data subject has the right to human intervention and to express her point of view and to contest the decision. Recital 71 mentions an additional element, namely that the data subject has the right "to obtain an *explanation*" after human review of the decision "and to challenge this decision".[4] Recitals, however, do not have the same legally binding force as the text of the GDPR itself.

[3]The exact interpretation of "logic" in the GDPR is not settled but likely does not refer to understandings of this term in philosophy or computer science.
[4]Children should not be subject to automated decision-making.

Over the past years there has been a vivid academic debate around whether the reference to "an explanation" in Recital 71 amounts to a "right to an explanation" that data subjects can exercise via-à-vis controllers [59] [32] [45] [14]. The Article 29 Working Party's guidance suggests that Article 22, read in conjunction with Recital 71, should be understood to require that controllers (i) tell data subjects that they are engaging in automated decision making, (ii) deliver meaningful information about the logic, and (iii) explain the processing's significance and envisaged consequences. The information provided should include details about the categories of data; why data is seen as pertinent; how profiles are built; why the profile is relevant for the decision-making process and how it is used to reach a decision about the data subject. The last three criteria appear to apply to profiling only [36]. Information with respect to the "logic" means "simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision". What is required is "not necessarily a complex explanation of the algorithms used or disclosure of the full algorithm". Nonetheless, the information transmitted to the data subject should be sufficiently comprehensive to "understand the reasons for the decision". Thus, an explanation of algorithms or disclosure of the full algorithm isn't "necessarily" required and that the controller ought to find "simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision". *Unfortunately, this guidance leaves a lot of room for doubt regarding what exactly is required of controllers.* In any event the GDPR does not create a general right to an explanation but applies only to automated decision-making that legally affect the data subject or have similarly significant effects on them.

3.3 Explainability as a sub-component of transparency

While there is a persistent myth that EU law requires that all decisions based on AI are "explainable" our analysis has painted a more nuanced picture. First, there is no overarching explainability norm that would apply to any usage of AI. To what degree secondary law requires explanations has not been authoritatively settled. Ultimately, the Court of Justice of the European Union will need to settle this question in respect of the GDPR. Concerning the draft AIA, however, legislators should clarify in the final text whether explainability is a free-standing legal obligation in respect of high-risk AI systems or whether it should rather be understood as a sub-component of transparency. As shown above, it is indeed possible to read references to explainability as elements of the broader transparency obligation. Article 13 AIA is explicitly about transparency, but the reference that this transparency must allow users to "interpret the system's output" has been understood as an explainability obligation by some. Further iterations should clarify the link between transparency and explainability to enhance legal certainty. An analysis of the history behind the AIA confirms the lack of precision

of the AIA itself. The EU High Level Expert Group on AI's report on the one hand portrayed explainability as a component of transparency. On the other hand, it repeatedly referred to another concept, "explicability", which was introduced as an ethical principle and as the "procedural dimension" of fairness. In contrast, the AIA White Paper made no reference to explainability other than to mention that symbolic reasoning could help make deep neural networks more explainable. This part of the AIA legislative history underlines the lack of consensus about what *exactly* explainability is. Similarly, the GDPR could also be read as referring to explainability as a sub-component of transparency. Articles 12-15 derive from the core data protection principle of transparency in Article 5(1)(a) and likewise, one reading of Article 22 in conjunction with relevant recitals could also be understood as a more general transparency rather than explainability obligation.

This, of course, raises the question of what transparency means and what it should enable. There is broad consensus that the GDPR requires that decisions reached through automated decision making be justifiable. Indeed, Hildebrandt has highlighted that data protection requires "the justification of such decision-making rather than an explanation in the sense of its heuristics" (p. 113 in [18]). Kamimski and Urban deem that justification should enable "understanding, revealing and making challengeable the normative grounds of a decision" (p. 1980 in [21]). Wachter, Mittelstadt and Russell have argued that explainability is ultimately designed to help the data subject understand, contest and alter decisions and that this could also be achieved by counterfactual explanations [60]. If explainability is merely one means of achieving transparency, there needs to be a more thorough discussion as to what other, alternative, means of achieving transparency there are, particularly in situations where explainability strictu sensu proves impossible. Considering the lack of consensus as to how the legislative texts of the AIA and the GDPR ought to be interpreted and applied in practice, it is helpful to consider their underlying objectives.

3.4 Rationale and objectives of explainability norms in an adversarial setting

The vague formulation of explainability rights, coupled with uncertainty regarding their function makes it legitimate to ask whether explanations serve any meaningful purpose. Indeed, as Edwards and Veale [14] have argued, "the search for a legally enforceable right to an explanation may be at best distracting and at worst nurture a new kind of transparency fallacy". This is essentially a warning that if explainability obligations just become a box-ticking exercise, they might give a misleading appearance of compliance rather than to be of any real value to the decision subject. In addition, explainability rights in the GDPR inevitably also suffer from the general shortcomings of the low enforcement of the GDPR.

In order to better understand the above-examined norms we propose to consider their underlying objectives. Before discussing legislative history let us recapitulate what philosophers have identified as main objectives for algorithmic explanations.[5] One major motivation for explainability of AI systems is the hope that this may foster **trust** in these systems [10, 26, 35, 57]. This has been called the "Explainability-Trust" hypothesis [22].[6] The hypothesis is controversial, and it is not exactly clear how explanations would induce trust. The underlying rationale seems to rest on an analogy with human interactions. Consider decisions made by human experts. When the decision doesn't satisfy us, we are drawn to ask for an additional explanation. Given such an explanation, we may check whether it conforms to our expectations about good decision making. If so, this may be a ground for further trusting the decision maker. This is not a one-shot process, but an ever evolving interaction on a long term time-scale. We tend to trust a person that proved repeatedly to predict correctly, make good decisions, or provide well informed explanations. The trust raising potential of an explanation however requires that we can submit explanations and decisions to tests, possibly by delegating it to other experts. The trust raising potential of a single explanation thus presupposes that the explanation provider stays in the information-exchange on the long run: only then does she have an incentive to provide a correct explanation, since an incorrect one would lead to a loss of trust in the long run but not in a one-shot exchange. If an algorithm rather than a human expert makes a decision, we might have similar expectations. We would like to engage in a similar information exchange with an algorithm as we engage with humans. The demand for an explanation is then a demand for a piece of communicative interaction. The hope that this builds trust stems from the intuition that the interaction with the algorithm is similar to the interaction among humans, as depicted above. This assumption may however fail either because the algorithmic explanations cannot be submitted to sensible tests or because the exchange is one-shot and not long run. In the first case, explanations loose their trust raising potential. In the second, the explanation provider may not have the incentive to tell the truth. A second implicit motivation for explainability stems from the idea that information provided by explanations can be **used to perform actions**, and may in fact be needed for such actions. In the adversarial setting, a data subject might want to use an explanation to *contest* a decision [7, 60], by claiming, or arguing that the decision is not right, not good, or not fair. The data subject might also want to use the

[5] Questions regarding "Explanations" have been discussed since the beginning of philosophy, with a strong revival in the philosophy of science of the last century, treating scientific explanations [1, 8, 17, 39, 43], causal explanations [28, 38, 49, 50, 61], and non-causal explanations [40]. We refer the interested reader to [44, 62] and restrict our discussion to the context of machine learning.

[6] For further references, see §2 therein.

explanation for *recourse*, in order to do better next time [4, 55, 60] (see also [2, 29]).[7] But such explanations are only of value when true or correct. A false explanation will not help in doing better next time, and may even be devised such as to render a decision incontestable.

The two motivations from philosophy — building trust and enabling recipients to act — can also be found as objectives in the legal texts. The EU High Level Expert Group on AI described explainability as one **tool to achieve trust** in AI systems [35].[8] The AIA provides that explainability norms are designed to allow users to fully understand the capacities and limitations of high-risk systems, leading again to trust. Partly related to trust, one can understand explainability as a tool for **risk management**, in line with the AIA's overall risk-based approach. Indeed, for high-risk AI systems, transparency must be ensured by monitoring the system's operation, detect signs of anomalies, dysfunctions and unexpected performance in order to counteract automation bias or to potentially intervene in the system (the idea of a "stop button"). The European Commission White Paper also emphasized the risk-based approach and stressed that due to the potential scale of AI systems [11]: a hidden bias or an incorrect assumption of an AI system, say deciding on tens of thousands of university admission decisions, will have a large systemic effect. This differentiates large-scale AI systems from human decision-making systems. In philosophy explanations are considered as a tool towards future actions. Similarly, the legal discussion also portrays explainability as an **enabling right**. The High-Level Expert Group on AI has drawn attention to the fact that to be able to contest decisions, they must be traceable. Also outside the AIA and the GDPR, explainability serves a related purpose. In consumer protection law, explainability is linked to the **unequal power dynamics** between the business and the consumer. In the public administration, it has been argued that being subjected to an intransparent black-box decision would undermine **human dignity** and is also to be avoided, unlike in the private sector, individuals cannot vote with their feet and go elsewhere.

Overall, the motivations for explanations seem to presuppose that such explanations are true or correct. Only then does a single explanation raise trust, and only then can an explanation be used to perform the intended actions, such as contesting or recourse. We will, however, see in the next section that this truth-presupposition for explanations fails in adversarial scenarios of algorithmic post-hoc explanations.

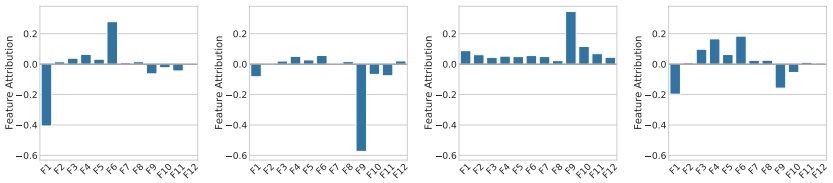
4 THE PROBLEMS WITH POST-HOC EXPLANATIONS IN ADVERSARIAL CONTEXTS

We now discuss the problems with post-hoc explanations in adversarial scenarios. What can we expect from an algorithmic explanation in these contexts? We roughly know what to expect from human explanations. For example, witnesses giving evidence in court are expected to tell the truth. Can we expect something similar of an algorithmic explanation? If the algorithm decided, for example, to reject a loan application, can we expect to discover the true reason why it decided to do so? The answer is that we cannot, for two reasons. First, the algorithm's view of the world is coarse-grained and incomplete, and this significantly restricts the vocabulary available for potential explanations (Section 4.1). Second, even within the limited picture of the world that the algorithm has access to (the "algorithm's own world") uniquely preferred or "ground truth" explanations do not exist (Section 4.2). This directly ties with the computer science perspective of why post-hoc explanations should not be used in adversarial contexts: the task of providing post-hoc explanations is underdetermined. The objective of the adversary explanation provider is to deploy a classifier that has high accuracy and generate post-hoc explanations that cannot be contested by the data subject or an examiner. We argue that due to the high degree of ambiguity inherent to algorithmic explanations, the adversary has sufficient degrees of freedom to devise incontestable explanations – even without explicitly optimizing against a particular explanation method [46, 47]. We identify four key quantities that allow the adversary to influence the resulting explanations: the choice of an explanation algorithm and its particular parameters (Sections 4.3 and 4.4); the exact shape of the high-dimensional decision boundary (Section 4.5); and, when applicable, the choice of the reference dataset (Section 4.6). This section contains a number of figures and simulation results. Additional figures can be found in the supplement. The code to replicate the results in this paper is available at https://github.com/tml-tuebingen/facct-post-hoc.

4.1 The algorithm's view of the world is coarse-grained and incomplete - this limits potential explanations

Learning and explanation algorithms only have access to a coarse-grained description of the real world. Their vocabulary is restricted to certain features, and possible relations between them. The "experience" of such algorithms given by the finite training data is formulated in the restricted vocabulary and provides only a small window to the world. Overall, the algorithm's representation of the real world is coarse-grained and

[7] Other actions belong more properly to the collaborative setting, such as debugging, improving, correcting, learning, understanding and testing.
[8] With the consequence that explainability would also play a role in stimulating the adoption of AI and the competitiveness of the internal market.



(a) SHAP **(b) LIME** **(c) DiCE** **(d) Interventional SHAP**

Figure 1: Different explanation algorithms lead to different explanations. Depicted are the feature attribution explanations of four different explanation algorithms: Exact SHAP for trees [31], LIME [41], DiCE [24], and Interventional SHAP [20]. All four explanation algorithms attempt to explain the prediction for the *same* individual with the *same* decision function (a gradient boosted tree) on the *same* dataset (Adult-Income). The idea of feature attribution explanations is to determine how much each dimension of the input contributed towards the decision. The figures depict these attributions by drawing a bar for each of the 12 input dimensions. The larger the bar, the higher is the influence of the corresponding feature. Some methods distinguish between positive and negative attributions. In the depicted example, the first bar in Panel (a) is relatively large, which indicates that the SHAP algorithm determined that the value of the first feature contributed strongly to the prediction. The DiCE algorithm in Panel (c), in contrast, determined that the value of feature 9 contributed most strongly to the prediction. More figures showing results for other data points can be found in the supplement.

incomplete.[9] The learning algorithm just sees features and training labels. The explanation algorithm, additionally, sees the learning algorithm's association between input and output. This is what we call "the world of the explanation algorithm", and this is all what it can exploit. As a consequence, all the explanation algorithm could talk about are geometric properties in the world of the algorithm: distances of points to the decision surface, proximity between points, their true or predicted labels, the gradient of the decision function at a point, the necessary change of a feature to change the decision, etc. Although a true explanation for a decision might exist in the real world, it might not be represented in the data or other aspects of the algorithm's world, which could thus not provide any such explanation. This is even the case in a cooperative setting. Consider the example of a medical diagnosis of a disease for which a true (say, causal) explanation exists in the real world. If the learning algorithm was trained on feature-based data such as age, blood pressure, etc, the explanation algorithm could suggest that age was the cause. However, in reality the cause for the disease may not be age, but rather a smoking habit that was not represented in the data. So even if a true explanation exists (say, a cause) this may neither be identifiable nor expressible by the explanation algorithm.

4.2 Even within the algorithm's own world, a unique preferred reason does not exist

Even within the limited world that the explanation algorithm has access to, a "true internal reason" why the learned decision function comes to a certain decision does generally not exist. This is particularly the case for complicated black-box functions. Even machine learning experts digging into the learning algorithm or properties of the function could not reveal a unique true reason. All we can do is to provide vague approximations of how the algorithm arrives at its decision, by summarizing which features contributed how much to the decision (the approach of LIME and SHAP), or whether a change in some features would alter the decision (the approach of counterfactual explanations). For example, in the case of a loan rejection, we might want to know whether it was rather our low income or our postal code which determined the decision, and whether we could change something about the decision, if in the future we had a higher income or moved to another area. However, these explanation attempts are all subject to choices. A mathematically unique way to determine how much each feature of a complicated black-box function contributed to the decision does not exist. Consequently, all feature attribution methods rely on particular assumptions and mechanisms in order to construct explanations: LIME, for example, looks at the gradient of the decision function at the point to be explained [15, 41]. SHAP compares the point with other datapoints from a reference population [16, 30]. Yet another approach would be to re-train the classifier on subsets of features or to use counterfactual feature importance, where one looks at the distance to the decision surface in various directions. All these mechanisms and choices seem plausible but, as we will see in the next section, they all deliver different explanations.

[9] Similar issues were discussed in [7, 19].

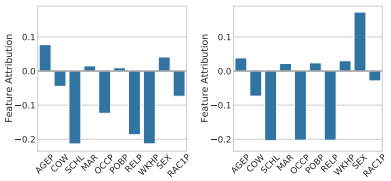
4.3 Different explanation algorithms lead to different explanations

Different explanation algorithms lead to different explanations [25]. This is true even if the algorithms have access to exactly the same information (the geometry of the data, the learned decision function, etc). In an adversarial context, this is problematic because it means that the creator of the system can modify the explanations by choosing a particular explanation algorithm. In practice, different explanation algorithms lead to different explanations even on the most simple machine learning problems. In high dimensions, that is in real-world problems, the difference between the explanations obtained from two different explanation algorithms can be so significant that the explanations are entirely different. This is illustrated in Figure 1. The figure depicts the feature attribution explanations that four different explanation algorithms determined for the *same* individual. From the difference between the four panels in Figure 1 it is quite clear that different explanation algorithms can lead to markedly different explanations, even if they all attempt to explain the same decision for the same individual.[10] Details on the machine learning problem, dataset and explanation algorithms can be found in the supplement.

That different explanation algorithms lead to different explanations is also true for counterfactual explanation methods [34, 60]. Indeed, there is a variety of ways in which the optimization problem can be set up, which in turn leads to different explanations. However, already a single counterfactual explanation method can lead to a large number of counterfactual explanations. In a cooperative context, being able to generate many different counterfactual explanations for the same individual can be beneficial [34]. In an adversarial context this is problematic because there is no principled way to choose among different counterfactual explanations, and the adversary is again awarded considerable discretion to determine explanations. In realistic, high-dimensional applications, the number of potential counterfactual explanations can quickly become very large. Let us illustrate this point on the German Credit Dataset. The German Credit Dataset is a 20-dimensional dataset with features on credit history and personal characteristic. The task is to predict credit risk in binary form. How many different counterfactual explanations exist for a single individual? With a common black-box decision function, more than 100 different counterfactual explanations exist for each individual.

At its core, the fundamental difficulty of explainable machine learning is then the same as in other fields of unsupervised learning: the lack of a ground truth explanation impedes the development of an algorithmic framework to automatically evaluate explanations. Every explanation algorithm needs to make assumptions about which properties of the decision function it seeks to highlight. As a result, it is possible to develop sanity checks for explanation algorithms and exclude unreasonable approaches [3, 9], but not to discern whether any of two post-hoc explanations is "more correct", which would be equivalent to discussing whether any of two different clusterings is "more correct" [58] .

[10]The reader who is acquainted with the internal mechanics of the depicted explanation method might feel that a direct comparison between the different methods is unwarranted, because different methods measure different aspects of the underlying decision function [9]. Note, however, that this is exactly the point that we want to make by explicitly contrasting the different attributions.



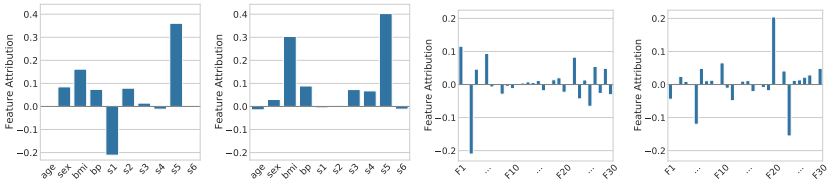
(a) SHAP **(b) LIME**

Figure 2: For any given datapoint, different explanation algorithms might lead to very similar or completely different explanations. In many cases, however, there are both similarities and dissimilarities. The Figure depicts the SHAP and LIME feature attributions for a datapoint in the folktables ACSIncome prediction task [13]: Are these attributions similar or different? More figures showing results for other data points can be found in the supplement.

4.4 The explanation provider can choose between a large number of possible explanation algorithms and parametrizations

Even for a single explanation algorithm, there can be many different parameter choices that all lead to different explanations. LIME explanations, for example, depend on the bandwidth and the number of perturbations [15, 27, 46]. The uniqueness properties of Shapley values non-withstanding, there is a multiplicity of ways in which Shapley values can be operationalized to generate explanations [51]. Counterfactual explanation algorithms depend on the underlying metric chosen to represent closeness (e.g. Euclidean distance vs. $L1$-norm)[11] as well as additional hyperparameters to weight-off between closeness and prediction, and, at least in principle, any number of additional penalty terms [34]. In certain cases, it might be possible to come up with good default parameter choices. For example, recent work has demonstrated how to choose the bandwidth parameter of LIME in a principled way or quantify uncertainty in the resulting explanations [27, 46, 64]. It is also possible to exclude explanation algorithms and parametrizations that are

[11]This originates in the philosophical account: counterfactuals depend on the way one measures proximity between facts and alternative counter-facts [28].



(a) Diabetes, Lin. Regr. **(b) Diabetes, Random Forest** **(c) Cancer, 36% Accuracy** **(d) Cancer, 96% Accuracy**

Figure 3: Explanations depend on the exact shape of the classifiers high-dimensional decision boundary. Panel (a) and (b): On the diabetes dataset, linear regression and a random forest agree for 94% of their predictions. Shown are the SHAP explanations on a data point where the prediction of both methods agree. As we can see, the explanations differ. Panel (c) and (d): the dependence on the decision boundary is subtle. It can even be hard to tell from the explanations whether the classifier had been trained trained at all. On the Wisconsin Breast Cancer dataset, the SHAP explanations of a classifier trained to achieve an accuracy of 96% are hard to distinguish from those of the same classifier trained on random labels. More figures showing results for other data points can be found in the supplement.

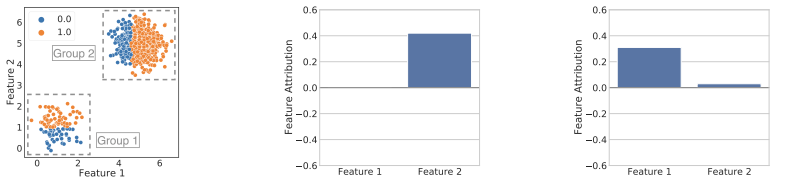
completely unreasonable, for example because they are not sensitive to the decision function [3, 9]. This nevertheless leaves an ever-increasing number of plausible explanation algorithms and corresponding parametrizations. Quite generally, different explanation algorithms vary among many different dimensions, and there is an ever increasing number of suggestions as to how black-box functions might be explained. This can be seen, for example, in the recent work of Covert et al. [12], who summarize 25 existing methods in a unified framework. As already discussed above, there are no fundamental reasons that impede us from using any particular method.[12]

4.5 Explanations depend on the exact shape of the high-dimensional decision boundary

Even if we fix a particular explanation method and its parameters, the generated explanations still depend on the *exact* shape of the learned decision boundary. In high dimensions, there are often many different black-box functions that solve a particular classification problem to a desired accuracy, that is they represent the data sufficiently well. However, these functions often lead to different explanations. To a certain extent, we may say that the exact shape of the learned decision boundary is arbitrary, but since the explanations depend on it, these turn out to be arbitrary as well. One of the reasons for the sensitivity of the explanation to the function's shape is that many explanation methods evaluate the function f at datapoints that are outside the data distribution or at points that are unlike most points from the data distribution. In the adversarial scenario, this is problematic because *the adversary can freely modify the values of the function f outside the data distribution* without changing the classification behavior. Recent work has demonstrated that this property can be used to explicitly manipulate and attack explanation methods [47, 48]. But even without explicit attacks, there are many different choices, in particular hyperparameter and architecture choices, that influence the shape of the decision boundary, and thus the resulting explanations. For an external examiner, this presents a challenging problem: while certain explicit attacks on explanation methods could in principle be detected through code review (see also Section 5.2), it is far less clear how one would argue about choosing one classifier over another, or any particular choice of hyperparameters. This problem is illustrated in Figure 3. Here, we solved the same machine learning problem both with linear regression and a random forest. The two methods have comparable performance on the test set, where 94% of their predictions agree. Nevertheless, the explanations obtained for the two different decision functions can be quite different – even for points that receive the same prediction.

Turning to counterfactual explanations, it is well-known that these depend on the exact shape of the decision boundary. Let us give an example, again using the German Credit Dataset. Consider two different decision functions, a gradient boosted tree and logistic regression. If we generate a number of diverse counterfactual explanations [34] for a typical individual with respect to one decision function, are these also counterfactual explanations with respect to the other decision function (at least as long as both functions arrive at the same decision)? In this simple experiment less than 50% of counterfactual explanations that work for the gradient boosted tree also work for logistic regression. As discussed above, the fact that the explanations depend on the exact shape of the decision boundary is problematic because it allows the creator of the system to influence the resulting explanations. The particular choice of the decision function can

[12]The distinction between two "different" explanation algorithms and different parameter choices for the "same" explanation algorithm is of course a matter of perspective: We might consider the question of distributional versus intervential Shapley values as a question of how to use "the" SHAP method [20], but we might as well perceive it as a discussion as to which of two different methods to use.



(a) Dataset (b) Reference data set: Group 1 only (c) Reference data set: entire dataset

Figure 4: A simple toy example of how the choice of the explanation's reference dataset can influence the resulting explanations. The dataset in Panel (a) consists of two different population groups. The blue and orange color depicts the binary label that the classifier is supposed to predict at each data point (to get an intuition, you might think of the groups as "male" and "female", and the label as "is awarded the credit" or "is not awarded the credit"). Panels (b) and (c) depict the interventional SHAP feature attributions [20] for the *same* data point in Group 1. In Panel (b), the explanation's reference dataset consists of the observations of Group 1 only. In Panel (c), the reference dataset is the entire dataset. The example shows that changing the reference dataset can almost completely change the feature attribution from one feature to another.

even determine whether certain types of counterfactual explanations exist at all. Let us give an example on the Wisconsin Breast Cancer Dataset. To demonstrate the dependence on the decision boundary, we consider again two different decision functions, linear regression and a random forest. For linear regression, there exist a large number of counterfactual explanations that modify only a single variable. For the random forest, it is impossible to find any such counterfactual explanations. This is despite the fact that both classifiers exhibit similarly low test error.

4.6 It is unclear how to choose the reference dataset that many explanations depend on

In recent years, there has been an increased focus on the composition of datasets, for example on the representation of different sociodemographic groups in machine learning datasets [6, 37]. In many real-world problems such as credit lending, the criteria for choosing an appropriate dataset are not clear. In both cooperative and adversarial contexts, the creator of the system has to make numerous choices, many of which can have significant effects on both the shape of the learned decision boundary and the generated explanations. For example, Anders et al. [5] have shown that gradient-based explanations can be manipulated by adding additional variables to the dataset. In this section, we highlight the additional role that the dataset can have on algorithmic explanations, even when *keeping the learned decision boundary constant*. Indeed, while some explanation algorithms such as LIME only rely on the learned decision boundary, other methods such as SHAP and some counterfactual explanation methods make additional use of the data in order to generate explanations. The relevant dataset could be the training data, but it could also be a different dataset. We refer to it as the *reference dataset*.

While the usage of such a dataset to generate explanations can be seen as a remedy to the vagaries of high dimensions, or as a possibility to generate counterfactual explanations that look like they come from the data, this approach is problematic as long as the adversary determines the composition of the dataset. The reason is that whether certain datapoints are included in the dataset or not can determine whether an explanation algorithm provides one or another explanation. Figure 4 illustrates this with a simple example: By deciding between two different reference datasets, one can effectively decide whether one ore another feature was relevant to the decision.

4.7 Bottom line: Post-hoc explanations are highly problematic in an adversarial context

It is extremely important to understand that an explanation algorithm is based on many human choices that are shaped by human objectives and preferences. While many choices are plausible, there is no objective reason to prefer one algorithm over the other, or one explanation over the other. Apart from the explanation algorithm and its particular parameters, explanations are influenced by human choices such as the selection of the classifier and the composition of the dataset. In adversarial contexts it implies that the adversary can choose, among many different plausible explanations, one that suits their incentives. This complicated situation makes it particularly difficult for external observers, including judges and regulatory bodies, to determine whether an explanation is acceptable. Explanation algorithms appear to provide objective explanations, yet as explained above this is not the case (compare Section 4.2).

5 ONCE AN EXAMINER IS ALLOWED TO ASSESS THE PROVIDED POST-HOC EXPLANATIONS, SHE'D BETTER INVESTIGATE THE DECISION FUNCTION DIRECTLY

So far we have discussed explainability obligations in European Union law and their motivation (Section 3), and pointed out theoretical (Sections 4.1-4.2) and practical (Sections 4.3-4.6) shortcomings of post-hoc explanations. In this section, we add yet another component to our argument. In an adversarial setting, it is not only the AI decision system itself but also the corresponding explanation algorithm which might need to be examined by a third party. Even if the examiner only attempts to assess the most basic consistency properties of the provided explanations, that is to check whether the explanations relate to the AI decision system at all, this necessarily requires that the examiner is able to query the AI system. But then, the explanations become entirely redundant: Rather than relying on explanations to enable risk management, provide trust or bias and discrimination detection (compare Section 3.4), the examiner could directly query the AI system for problematic decision behavior. Because the creator of the system and the examiner have competing interests, it is important to distinguish degrees of transparent interaction between the two. Naturally, the examiner would like to have access to as much information as possible, whereas the adversary creator wants to disclose as little information as possible. We distinguish between a minimal and a fully transparent scenario of information disclosure (Sections 5.1-5.2).

5.1 Minimalist scenario where decision function and explanation algorithm can be queried

To determine whether the adversary's explanations actually correspond to the used decision function f instead of being arbitrary justifications not related to the decision process, the examiner needs to be able to query the decision function and the generated explanations.[13] This includes a fair amount of related knowledge, such as which variables are input to the algorithm, but excludes explicit access to the decision function, explanation algorithm, source code and training dataset. A related but slightly more limited version of this scenario arises when individuals jointly collect the decisions and explanations from the creator of the system. In this *minimalist scenario*, the examiner can validate the internal consistency of the provided explanations. Researchers have proposed a number of criteria that the examiner can test for such as faithfulness to the model, robustness to local perturbations, as well as necessity and sufficiency notions for individual feature attributions [3, 24, 56]. The examiner might also want to perform tests as to whether the provided explanations have been manipulated [48]. More importantly however, even just with the ability to query the decision function, the examiner can ignore the explanations and directly investigate the decision function for problematic properties. For example, the examiner could conduct a systematic evaluation of, say, fairness metrics such as equal opportunity and demographic parity, based on an independent reference dataset of her choice (see [6] for these and other notions of fairness and discrimination). Indeed, because the adversary designing the explanation algorithm has no interest in choosing explanations that highlight any discriminatory behavior of the decision algorithm, the examiner is well-advised to simply ignore the explanations and test the decision algorithm directly. Although such tests might be similar to certain explanation algorithms, what is important is that the examiner (as opposed to the creator) designs and implements them. Note that we are *not* saying that the minimalist scenario actually allows the examiner to assess all legally relevant properties of the decision function. What exactly can be assessed with querying access is a question that still requires more research. Our point is that once we have querying access, the explanations are useless.

[13] This means that for any possible datapoint (or individual) x, the examiner is allowed to ask the adversary: "For this hypothetical datapoint x, what would be the decision $y = f(x)$, and what would be the corresponding explanation $E(x, y)$? The adversary would then privately compute both quantities and make them available to the examiner, but not tell the examiner how the computation was performed.

5.2 Fully transparent scenario where algorithms' source code and training data are disclosed

At the opposite end of the minimalist scenario is the *fully transparent scenario* where the examiner is allowed to investigate the decision function, source code and training data. An examiner could then scrutinize whether the explanation algorithms have been implemented according to the state of the art with sensible parameter choices. This directly rules out the possibility for the creator of the system to manipulate explanations. Are post-hoc explanations useful in the transparent scenario, perhaps because the examiner now has the tools to verify whether the adversary has chosen the "correct" explanations? As we have already discussed above, the problem is that there is no notion of "correct" explanation (Sections 4.2 and 4.3). Thus, except for notions of internal consistency [3, 24], there is, in general, nothing the examiner can say about the explanations. Another issue, already observed in Sections 4.5 and 4.6, are hyperparameter choices and decisions regarding the composition of the dataset. For these decisions, it is highly non-trivial to come up with uniquely reasonable defaults: If the adversary has found a particular neural network architecture with hyperparameters that generalize well on the adversary's own dataset, how exactly could the examiner argue that this is inappropriate? Nevertheless, all of these choices can influence the resulting explanations, even if we fix a particular explanation algorithm. Of course, the examiner could scrutinize the source code, re-train the system with different parameters, perform tests on the data, and generate alternative explanations. Some have argued

that this might be sufficient in order to assess a variety of legal requirements [23]. While we think that more research is needed on what can be realistically achieved in the fully transparent scenario, it is quite clear that the examiner can, at least in principle, perform a variety of powerful tests (whether this is achievable in practice, based on the limited resources of an examiner, is yet a different story). At any rate, just as in the minimalist scenario, the examiner is well-advised to examine and test the system on her own, and to ignore the explanations provided by the adversary creator.

6 DISCUSSION

Explainability is often praised as a tool to mitigate some of the risks of black-box AI systems. Our paper demonstrates that in adversarial contexts, post-hoc explanations are of very limited use. From a technical and philosophical point of view these explanations can never reveal the "unique, true reason" why an algorithm came to a certain decision. In complicated black-box models, such a true reason simply does not exist. We moreover demonstrated that post-hoc explanations of standard decision algorithms on simple datasets possess a high degree of ambiguity that cannot be resolved in principle. For these reasons, post-hoc explanations of black-box systems are, to a certain degree, incontestable. In the best case, post-hoc explanation algorithms can point out some of the factors that contributed to a decision — these algorithms are therefore useful for model debugging, scientific discovery and practical applications where all parties share a common goal. In adversarial contexts, in contrast, we demonstrated that local post-hoc explanations are either trivial or harmful. In the worst case, the explanations may induce us into falsely believing that a "justified", or "objective" decision has been made even when this is not the case.

It was also seen that it remains unclear how expectations of explainability in the GDPR or the AIA ought to be interpreted. The GDPR does not give rise to a general explainability obligation, and the draft AI Act currently would only require some degree of explainability in relation to high-risk applications of AI. We call on legislators to formulate related provisions with more specificity in order to create legal certainty in this respect. If the final version of the AIA requires a strong version of explainability for high-risk AI systems, black-boxes simply cannot be used: they cannot be explained directly, and the only indirect means of explaining them — local post-hoc explanations — are unsuitable. In this case, one would have to resort to the use of simple, inherently interpretable machine learning models rather than black-box models (compare [42]) although this may impede innovations. We would expect that these algorithms and their explanations are more robust and less susceptible to manipulation, such that large parts of our criticism would not apply to inherently interpretable models. However, future research needs to clarify whether this is the case, because we are not aware of any research that investigates inherently interpretable machine learning in an adversarial setting. If, on the other hand, explainability in the final version of the AIA is to be understood as one of several means to achieve more transparency in machine learning, other methods than post-hoc explanations might be more suitable to achieve the desired goals of transparency. For example, as far as testing for biases and discrimination is concerned, it is unlikely that the creator of the system will choose to generate explanations that can be used to uncover hidden biases. But there is a much more direct route to assess discrimination than implicitly through explanations. Indeed, external examiners could directly test the system for discriminatory properties [23]. As such, the external examination of black-boxes may be a more suitable means of enabling more accountable AI systems.

The current draft of the AIA already requires documentation regarding the functioning of AI systems. However, one has to be aware of the versatile manipulation possibilities that lie in the development process of AI systems itself, through choice of training data, features, algorithms, parameters, and so on. Even in the fully transparent scenario where the entire development pipeline including the source code is open [23], a considerable leeway for manipulations remains. In order to address these, an external examiner would need access to considerable manpower and resources. Even when training data and source code can in principle be examined, algorithms re-applied or even retrained, actually doing so for a system that has been developed by a large team might be very difficult if not impossible. More research is needed to understand exactly which legal objectives can be satisfied by such extended documentation of AI systems, or whether the documentation would again just serve as a means to provide an appearance of objectivity without any real value.

Overall, we believe that the question of testing and certifying machine learning systems in an adversarial scenario is a research direction that is still heavily underexplored. There is no single way to achieve all the desired transparency and control goals for such AI systems. Even complete transparency, open code, open data might not lead to all the desired goals. For this reason, it is important to investigate in more detail what objective can be achieved by which means, and which goals might not be possible to achieve at all. Only then can we engage in a meaningful debate about responsible use of AI systems in social contexts.

Finally, we recall that our criticism of explainability, in particular local post-hoc explanations, concerns adversarial scenarios. In cooperative scenarios, many interesting discoveries might be made with the help of explainable machine learning.

7 FUNDING DISCLOSURE

This work has been partially supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645), the Baden-Württemberg Foundation (program "Verantwortliche Künstliche Intelligenz"), the BMBF Tübingen AI Center (FKZ: 01IS18039A), the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the Carl Zeiss Foundation. The authors declare no additional sources of funding and no financial interests.

REFERENCES

[1] P. Achinstein. 1983. *The Nature of Explanation.* Oxford University Press, New York.
[2] A. Adadi and M. Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
[3] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. 2018. Sanity checks for saliency maps. In *Neural Information Processing Systems (NeurIPS).*
[4] A.Karimi, G. Barthe, B. Schölkopf, and I. Valera. 2021. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. arXiv:2010.04050
[5] C. Anders, P. Pasliev, A. K. Dombrowski, K. R. Müller, and P. Kessel. 2020. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning (ICML).*
[6] S. Barocas, M. Hardt, and A. Narayanan. 2019. *Fairness and Machine Learning.* fairmlbook.org. http://www.fairmlbook.org.
[7] S. Barocas, A. Selbst, and M. Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *ACM Conference on Fairness, Accountability, and Transparency.*
[8] R. B. Braithwaite. 1953. *Scientific Explanation: A Study of the Function of Theory, Probability and Law in Science.* Cambridge University Press, Cambridge.
[9] O. Camburu, E. Giunchiglia, J. Foerster, T. Lukasiewicz, and P. Blunsom. 2019. Can I trust the explainer? Verifying post-hoc explanatory methods. *arXiv:1910.02065* (2019).
[10] L. Chazette, W. Brunotte, and T. Speith. 2021. Exploring explainability: A definition, a model, and a knowledge catalogue. In *IEEE 29th International Requirements Engineering Conference (RE).*
[11] European Commission. 2020. White Paper on Artificial Intelligence-A European approach to excellence and trust. *Com (2020) 65 Final* (2020).
[12] I. Covert, S. Lundberg, and S.I. Lee. 2021. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research (JMLR)* 22, 209 (2021), 1–90.
[13] F. Ding, M. Hardt, J. Miller, and L. Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. In *Neural Information Processing Systems (NeurIPS).*
[14] L. Edwards and M. Veale. 2017. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke Law and Technology Review* 16 (2017).
[15] D. Garreau and U. von Luxburg. 2020. Explaining the Explainer: A First Theoretical Analysis of LIME. In *Conference on Artificial Intelligence and Statistics (AISTATS).*
[16] S. Ghalebikesabi, L. Ter-Minassian, K. DiazOrdaz, and C. C. Holmes. 2021. On locality of local explanation models. In *Advances in Neural Information Processing Systems (NeurIPS).*
[17] C. Hempel. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science.* Free Press, New York.
[18] M. Hildebrandt. 2019. Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law* 20, 1 (2019), 83–121.
[19] A. Z. Jacobs and H. Wallach. 2021. Measurement and fairness. In *ACM conference on Fairness, Accountability, and Transparency.*
[20] D. Janzing, L. Minorics, and P. Blöbaum. 2020. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics (AISTATS).*
[21] M. Kaminski and J. Urban. 2021. The Right to Contest AI. *Columbia Law Review* (2021).
[22] L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, and S. Sterz. 2021. On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness. In *IEEE 29th International Requirements Engineering Conference Workshops (REW).*
[23] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018), 113–174.
[24] R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma. 2021. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AAAI/ACM Conference on AI, Ethics, and Society.*
[25] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv preprint arXiv:2202.01602* (2022).
[26] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021).
[27] E. Lee, D. Braines, Mi. Stiffler, A. Hudler, and D. Harborne. 2019. Developing the sensitivity of LIME for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications.*
[28] D. Lewis. 1973. *Counterfactuals.* Blackwell.
[29] Q. V. Liao and K. R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv preprint arXiv:2110.10790* (2021).
[30] S. Lundberg and S. Lee. 2017. A unified approach to interpreting model predictions. In *Neural Information Processing Systems (NeurIPS).*
[31] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.
[32] G. Malgieri and G. Comandé. 2017. Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation. *International Data Privacy Law* 7, 4 (11 2017), 243–265.
[33] C. Molnar. 2020. *Interpretable machine learning.* Lulu.com.
[34] R. Mothilal, A. Sharma, and C. Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *ACM Conference on Fairness, Accountability, and Transparency.*
[35] High-Level Expert Group on AI. 2019. Ethics Guidelines for Trustworthy AI.
[36] Working Party. 2016. Guidelines on Automated individual decision-making and Profiling for the purposes of RegulationGuidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679.
[37] A. Paullada, I. Raji, E. Bender, E.and Denton, and A. Hanna. 2021. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2, 11 (2021).
[38] J. Pearl. 2000. *Causality: Models, Reasoning and Inference.* Cambridge University Press, Cambridge.
[39] K. Popper. 1959. *The Logic of Scientific Discovery.* Hutchinson, London.
[40] A. Reutlinger and J. Saatsi. 2018. *Explanation Beyond Causation; Philosophical Perspectives on Non-Causal Explanations.* Oxford University Press, Oxford.
[41] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. Why should i trust you? Explaining the predictions of any classifier. In *22nd ACM SIGKDD international conference on knowledge discovery and data mining.*
[42] C. Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.
[43] W. Salmon. 1971. *Statistical Explanation and Statistical Relevance.* University of Pittsburgh Press, Pittsburgh, PA.
[44] W. Salmon. 1989. Four Decades of Scientific Explanation. In *Scientific Explanation*, Kitcher and Salmon (Eds.). Minnesota Studies in the Philosophy of Science, Vol. 13. University of Minnesota Press, 3–219.
[45] A. Selbst and J. Powles. 2018. Meaningful Information and the Right to Explanation. In *ACM Conference on Fairness, Accountability, and Transparency.*
[46] D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju. 2021. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Neural Information Processing Systems (NeurIPS).*
[47] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society.*
[48] D. Slack, S. Hilgard, H. Lakkaraju, and S. Singh. 2021. Counterfactual Explanations Can Be Manipulated. *arXiv:2106.02666* (2021).
[49] P. Spirtes, C. Glymour, and R. Scheines. 1993. *Causation, Prediction, and Search.* Springer, Berlin.
[50] W. Spohn. 1980. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic* 9 (1980), 73–99.
[51] M. Sundararajan and A. Najmi. 2020. The many Shapley values for model explanation. In *International Conference on Machine Learning*

(ICML).
[52] R. Tomsett, D. Braines, D. Harborne, A. Preece, and S. Chakraborty. 2018. Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. In *ICML Workshop on Human Interpretability in Machine Learning*.
[53] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. Soyer, I. Zalaudek, and H. Kittler. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234.
[54] M. Veale and F. Zuiderveen Borgesius. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International* 22, 4 (2021), 97–112.
[55] S. Venkatasubramanian and M. Alfano. 2020. The Philosophical Basis of Algorithmic Recourse. In *ACM Conference on Fairness, Accountability, and Transparency*.
[56] G. Vilone and L. Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106.
[57] W. J. von Eschenbach. 2021. Transparency and the Black Box Problem: Why We Do Not Trust AI. *Philos. Technol.* 34 (2021), 1607–1622.
[58] U. von Luxburg, R. Williamson, and I. Guyon. 2012. Clustering: Science or Art? *JMLR Workshop and Conference Proceedings (Workshop on Unsupervised Learning and Transfer Learning)* (2012), 65 – 79.
[59] S. Wachter, B. Mittelstadt, and L. Floridi. 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 (06 2017), 76–99.
[60] S. Wachter, B. Mittelstadt, and C. Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
[61] J. Woodward. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.
[62] J. Woodward and L. Ross. 2003. Scientific Explanation. *The Stanford Encyclopedia of Philosophy (Summer Edition 2021)* (2003). https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/
[63] C. Zednik and H. Boelsen. forthcoming. Scientific Exploration and Explainable Artificial Intelligence. *Minds and Machines* (forthcoming).
[64] Y. Zhang, K. Song, Y. Sun, S. Tan, and M. Udell. 2019. Why Should You Trust My Explanation? Understanding Uncertainty in LIME Explanations. *arXiv preprint arXiv:1904.12991* (2019).

A POST-HOC EXPLANATIONS FAIL TO ACHIEVE THEIR PURPOSE IN ADVERSARIAL CONTEXTS: SUPPLEMENTARY MATERIALS

A.1 Code

The python code to replicate all results in this paper is available at https://github.com/tml-tuebingen/facct-post-hoc.

A.2 Datasets

In our experiments, we used the following datasets.

Adult-Income. This dataset contains information about individuals based on the 1994 US Census. It is available from the UCI machine learning repository. We obtained it from the SHAP package https://github.com/slundberg/shap. The dataset contains the 12 features age, workclass, education-num, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, country. In the figures, the features are numbered F1-F12 in this order. The machine learning problem is to predict whether whether an individual's income is over $50,000. We trained a gradient boosted tree which achieved a test accuracy of 87%.

German Credit. The German Credit Dataset is a dataset with 20 different features on individual's credit history and personal characteristic. The machine learning problem is to predict credit risk in binary form. We obtained the dataset from the UCI machine learning repository. We trained a gradient boosted tree which achieved a test accuracy of 76%. We also trained logistic regression which achieved a test accuracy of 74%.

Folktables. Folktables is a Python package that provides access to datasets derived from recent US Censuses https://github.com/zykls/folktables. We used this package to obtain the data from the 2016 Census in California. The machine learning problem is the ACSIncome prediction task, that is to predict whether an individual's income is above $50,000, based on 8 personal characteristics. We trained a gradient boosted tree which achieved a test accuracy of 83%.

Diabetes. The Diabetes dataset is a dataset of diabetes patient records. It is available from the UCI machine learning repository. We obtained it from the scikit-learn machine learning library https://scikit-learn.org. The dataset contains 10 features about each individual at baseline: age, sex, body mass index, average blood pressure, and six blood serum measurements. The machine learning problem is to predict disease progression one year after baseline. We converted the scalar outcome into a binary by thresholding at the median. We trained linear regression which achieved a test accuracy of 71%. We also trained a random forest which achieved a test accuracy of 74%.

Wisconsin Breast Cancer. The Wisconsin Breast Cancer dataset is a tabular dataset with features of breast mass images. The dataset contains 30 features that describe the characteristics of the cell nuclei present in the image. The dataset is available from the UCI machine learning repository. We obtained it from the scikit-learn machine learning library https://scikit-learn.org. The machine learning problem is to predict the binary diagnosis (malignant/benign). We trained linear regression which achieved a test accuracy of 96%. We also trained linear regression on random labels which achieved a test accuracy of 36%.

A.3 Explanation Algorithms

In our experiments, we used the following explanation algorithms.

SHAP The SHAP algorithm was proposed by [30]. We use it via the accompanying python package https://github.com/slundberg/shap. With (gradient boosted) trees, we use the exact computation method proposed in [31]. With all other classifiers, we use the Kernel SHAP method. The approach by Janzing et al. [20] is also implemented in this package. Whenever available, we use parametrizations proposed in the documentaion of the package.

LIME The LIME algorithm was proposed by [41]. We use it via via the accompanying python package https://github.com/marcotcr/lime. Whenever available, we use parametrizations proposed in the documentaion of the package.

DiCE The DiCE algorithm was proposed by [34]. We use it via the accompanying python package https://github.com/interpretml/DiCE. To generate counterfactual explanations, we used the model-agnostic randomized sampling method.

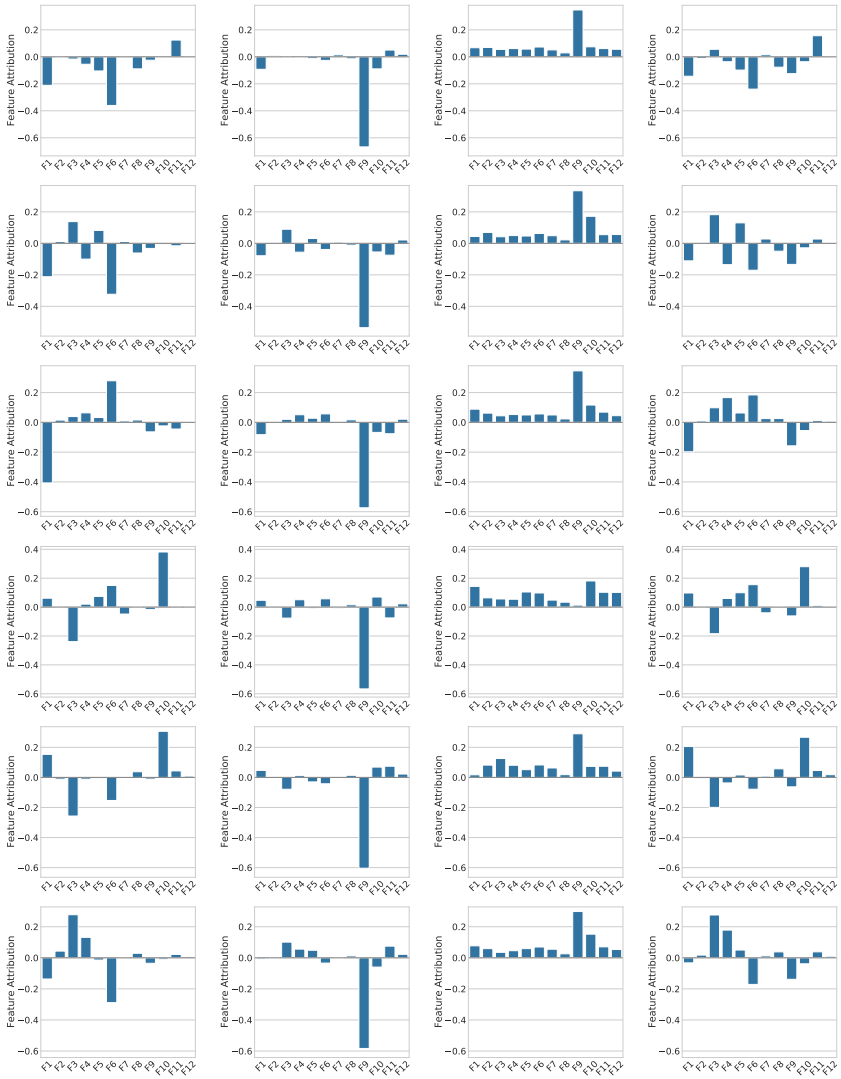
A.4 Figures

To create the figures, we normalized the feature attributions to have $l1$-norm 1.

A.5 Additional Figures

The following pages contain additional figures. These follow the figures in the main paper and depict the first observations from the test set, so they are not hand-selected in any way. The reader might notice that we selected the figures in the main paper from these. Figures for all observations from the test are avaialble with the code that will be made available upon publication.

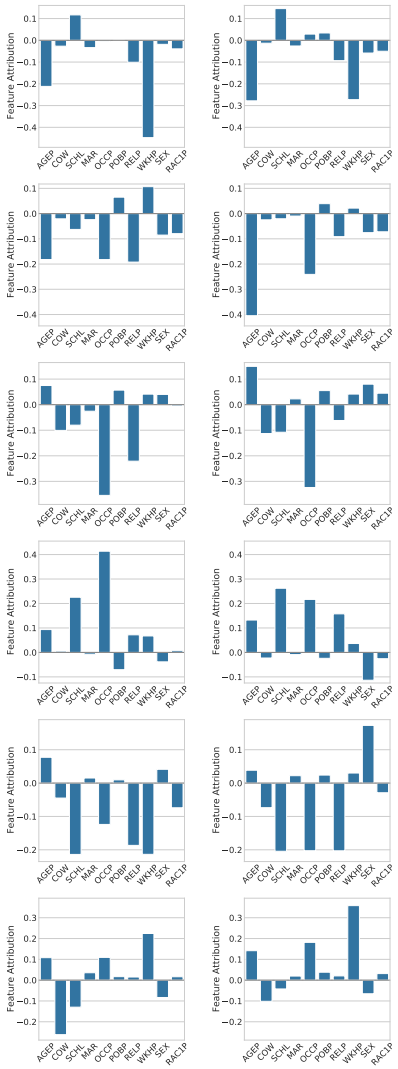
Additional Figures Related to Figure 1 in the Main Paper



(a) SHAP (b) LIME (c) DiCE (d) Interventional SHAP

Figure A.1: Different explanation algorithms lead to different explanations (compare Figure 1 in the main paper). Every row depicts the explanations of the four different explanation algorithms for another individual. The Figure depicts the first 6 observations from the test set.

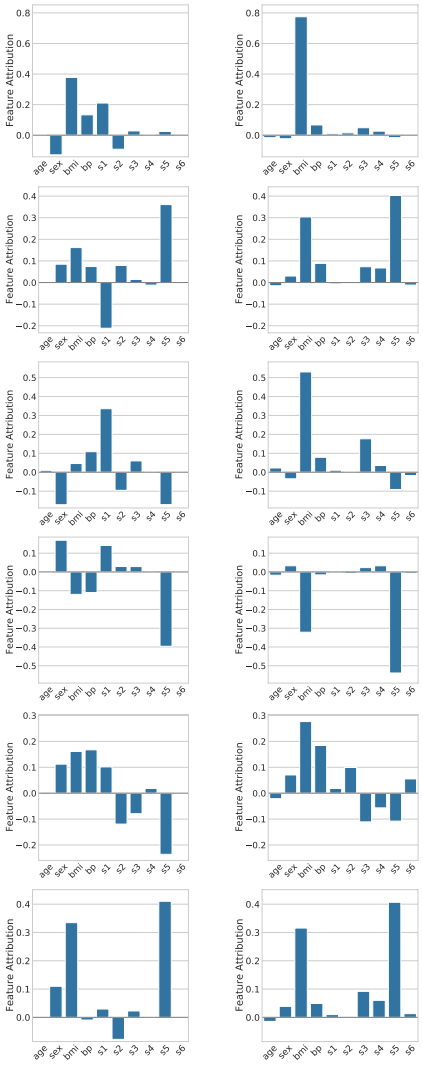
Additional Figures Related to Figure 2 in the Main Paper



(a) SHAP (b) LIME

Figure A.2: For any given datapoint, different explanation algorithms might lead to very similar or completely different explanations. In many cases, however, there are both similarities and dissimilarities (compare Figure 2 in the main paper). Every row depicts the explanations of the two different explanation algorithms for another individual. The Figure depicts the first 6 observations from the test set.

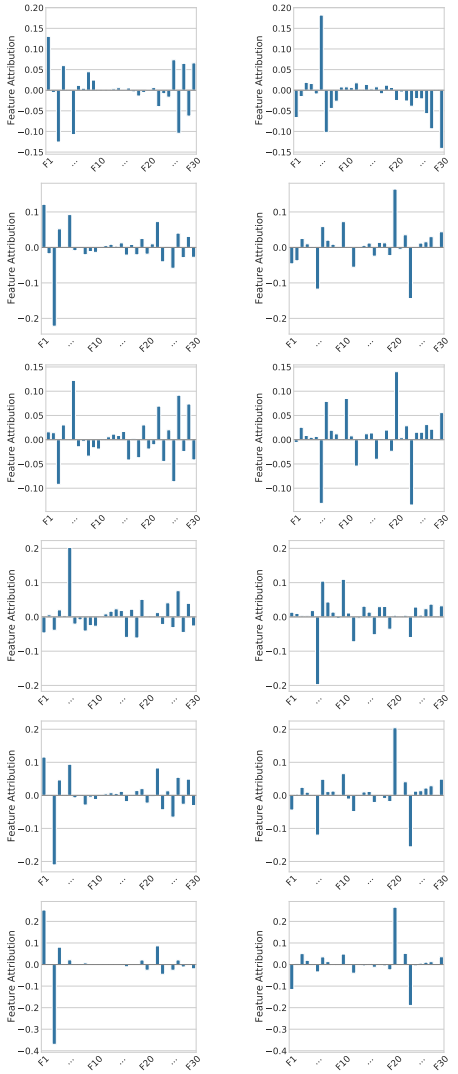
Additional Figures Related to Figure 3 (a), (b) in the Main Paper



(a) Diabetes, Linear Regression (b) Diabetes, Random Forest

Figure A.3: Explanations depend on the exact shape of the decision boundary (compare Figure 3 in the main paper). Every row depicts the explanations of the two different explanation algorithms for another individual. The Figure depicts the first 6 observations from the test set.

Additional Figures Related to Figure 3 (c), (d) in the Main Paper



(a) Breast Cancer, 36% Accuracy (b) Breast Cancer, 96% Accuracy

Figure A.4: Explanations depend on the exact shape of the decision boundary (compare Figure 3 in the main paper). Every row depicts the explanations of the two different explanation algorithms for another individual. The Figure depicts the first 6 observations from the test set.

2.3 From Shapley Values to Generalized Additive Models and back

From Shapley Values to Generalized Additive Models and back

Sebastian Bordt
Department of Computer Science
University of Tübingen

Ulrike von Luxburg
Department of Computer Science and Tübingen AI Center
University of Tübingen

Abstract

In explainable machine learning, local post-hoc explanation algorithms and inherently interpretable models are often seen as competing approaches. This work offers a partial reconciliation between the two by establishing a correspondence between Shapley Values and Generalized Additive Models (GAMs). We introduce n-Shapley Values, a parametric family of local post-hoc explanation algorithms that explain individual predictions with interaction terms up to order n. By varying the parameter n, we obtain a sequence of explanations that covers the entire range from Shapley Values up to a uniquely determined decomposition of the function we want to explain. The relationship between n-Shapley Values and this decomposition offers a functionally-grounded characterization of Shapley Values, which highlights their limitations. We then show that n-Shapley Values, as well as the Shapley Taylor- and Faith-Shap interaction indices, recover GAMs with interaction terms up to order n. This implies that the original Shapely Values recover GAMs without variable interactions. Taken together, our results provide a precise characterization of Shapley Values as they are being used in explainable machine learning. They also offer a principled interpretation of partial dependence plots of Shapley Values in terms of the underlying functional decomposition. A package for the estimation of different interaction indices is available at `https://github.com/tml-tuebingen/nshap`.

1 INTRODUCTION

Local post-hoc explanation algorithms and inherently interpretable models are two of the most prominent approaches in explainable machine learning (Molnar, 2020; Holzinger et al., 2022). Despite a number of arguments about their relative benefits, the differences and similarities between these two approaches remain largely unresolved Rudin (2019). In the current literature, post-hoc explanations and inherently interpretable models are often framed as different concepts, with research papers, book chapters, and tutorials divided along these lines (Lundberg et al., 2020; Molnar, 2020; Lakkaraju et al., 2020). We take a different perspective and highlight the similarities between post-hoc explanations and interpretable models. We do so for the particular case of Shapley Values, a prominent feature attribution method, and GAMs, a popular class of interpretable models.

Post-hoc explanations with Shapley Values. The seminal work by Lundberg and Lee (2017) introduced the SHAP feature attributions. These are based on the literature on Shapley Values in game theory. The authors showed that for linear functions $f(x) = w^T x$ and statistically independent features, the SHAP attributions take the form $\Phi_i = w_i(x_i - \mathbb{E}(x_i))$, thus establishing a link between the post-hoc explanation method and a very simple type of interpretable model. This work has inspired a whole branch of literature on explainable machine learning. Most relevant to us are Shapley Interaction Values (Lundberg et al., 2020), which extend Shapley Values with local interaction effects between pairs of features.

An important building block of our work is the generalization of Shapley Interaction Values towards **n-Shapley Values**, a novel type of Shapley-based post-hoc explanation that is able to incorporate arbitrarily many variable interactions. Similarly to the Shapley Taylor- (Sundararajan et al., 2020) and the Faith-Shap interaction index (Tsai et al., 2022), n-Shapley Values are a parametric family of local post-hoc explanation algorithms that explain individual predictions with interaction terms up to order n. As n increases, the explanations become more complex and expressive and are able to faithfully explain more complex models.

Generalized Additive Models (GAMs hereafter) are a popular class of interpretable models with a restricted form of non-linearity (Hastie and Tibshirani, 1990; Caruana et al., 2015; Agarwal et al., 2021a). Traditionally, GAMs are allowed to exhibit (arbitrary) non-linearity in individual

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

features, but no interaction between features is allowed. GA^{2}Ms (Lou et al., 2012) relax this restriction and allow for interaction between pairs of features. Conceptually, it is straightforward to extend GAMs with interaction effects of any desired order n (this comes, however, at the cost of human interpretability). Important to us, the model class of GAMs suffers from an identification problem. As soon as we introduce variable interactions, the way in which a given function can be written as a GAM is no longer uniquely determined Lengerich et al. (2020).

Shapley-based explanations faithfully explain GAMs. In this work, we show that different kinds of Shapley-based post-hoc explanations (Lundberg and Lee, 2017; Lundberg et al., 2020; Sundararajan et al., 2020; Tsai et al., 2022) are completely faithful to GAMs: if the function to be explained is a GAM, then the explanations recover its individual non-linear component functions. We link the order of the GAM – the maximum degree of variable interaction that is present in a function – with the order of an explanation that we use to explain that function. If the order of the explanation is at least as large as the maximum variable interaction that is (locally) present in the model, then the explanations are guaranteed to recover a faithful representation of the function as a GAM. This result applies to the newly proposed n-Shapley Values, as well as to the Shapley Taylor- and Faith-Shap interaction indices. As a special case, our results imply that the interventional SHAP feature attributions (Lundberg and Lee, 2017; Janzing et al., 2020) are perfectly faithful to GAMs without variable interactions, even if the features are arbitrarily dependent.

What is more, we show that Shapley-based post-hoc explanations of **any function** implicitly solve the problem of representing the function as a GAM (potentially with variable interactions of very high order). This means that our results provide insights into the mechanics of Shapley Values not only if the function to be explained is a lower-order GAM, but any (learned) function, for example a neural network. Concretely, we identify a necessary and sufficient regularity condition – subset compliance – under which a value function gives rise to a well-defined functional decomposition of the function that we attempt to explain. Because this decomposition connects Shapley Values with GAMs, we term it the Shapley-GAM.

Taken together, our results offer a precise **functionally-grounded analysis** of Shapley Values, one of the most widely used approaches in explainable machine learning (Doshi-Velez and Kim, 2017). They also highlight the peculiar properties of these explanations, and the way in which they are different from other feature attribution methods (Covert et al., 2021; Krishna et al., 2022). For example, contrary to popular belief, Shapley Values only depend on the coordinates of the point that we attempt to explain, but not on the local neighbourhood of that point. This in turn implies that the explanations are unrelated to the gradient and do not perform any kind of local function approximation (Han et al., 2022).

We consider n-Shapley Values to be a useful tool for practitioners who want to debug black-box models. Moreover, we introduce a novel method to plot feature attributions of higher order that is consistent with the underlying theory (depicted, for example, in Figure 1). We also introduce a way to estimate the amount of variable interaction that is necessary to represent a given function. Finally, we study the link between accuracy and the average degree of variable interaction present in different standard classifiers (Section 7).

2 RELATED WORK

Shapley Values. The seminal paper by Lundberg and Lee (2017) has led to a line of work that investigates the usage of Shapley Values in explainable machine learning (Chen et al., 2020; Heskes et al., 2020; Slack et al., 2020; Albini et al., 2022). Shapley Values originate in a literature on economic game theory (Shapley, 1953), and our work builds on a particular paper from this literature, namely the seminal work by Grabisch (1997) on additive set functions. The idea to extend Shapley Interaction Values towards n-Shapley Values is closely related to other approaches that also extend the Shapley Value (Grabisch, 1997; Lundberg et al., 2020; Sundararajan et al., 2020; Tsai et al., 2022). The efficient computation of Shapley Values is a topic of ongoing research interest (Lundberg et al., 2020; Jethani et al., 2021). Our results also relate to the debate about the choice of value function (Sundararajan and Najmi, 2020; Janzing et al., 2020). Shapley Values have been explored in various tasks with human decision makers, a topic about which there is much debate (Kumar et al., 2020).

Generalized Additive Models. Generalized additive models originate in statistics (Hastie and Tibshirani, 1990) and have recently become popular in combination with trees (Lou et al., 2012, 2013) and neural networks (Agarwal et al., 2021a). On tabular data sets, interpretable GAMs with few interactions (Caruana et al., 2015) can often achieve competitive accuracy, which has led to an active line of research on these models (Wang et al., 2022; Lengerich et al., 2022). From a statistical perspective, the decomposition of a function as a GAM is underdetermined, which has led to the development of additional uniqueness criteria such as functional ANOVA (Hooker, 2007; Lengerich et al., 2020).

Explainable Machine Learning. Shapley Values are one of many different feature attribution methods (Ribeiro et al., 2016; Sundararajan et al., 2017; Kommiya Mothilal et al., 2021) about which there is a large literature (Lee et al., 2019; Garreau and von Luxburg, 2020; Slack et al., 2021; Covert et al., 2021; Krishna et al., 2022; Han et al., 2022) and much debate (Lipton, 2018; Rudin, 2019; Bordt et al., 2022). Considerable debate also exists around the question whether there is an accuracy-explainability trade-off or a cost of us-

ing interpretable models (Rudin, 2019; Moshkovitz et al., 2020). Apart from GAMs, there are many other interpretable models such as rule lists (Wang and Rudin, 2015) and sparse decision trees (Lin et al., 2020). Since our work is exclusively focused on Shapley Values and GAMs, we do not offer a comprehensive review of the literature on explainable machine learning. This can be found in many other places (Molnar, 2020; Samek et al., 2021; Holzinger et al., 2022; Rudin et al., 2022).

3 BACKGROUND AND NOTATION

We consider data points $x \in \mathbb{R}^d$ with d features, and a function $f : \mathbb{R}^d \to \mathbb{R}$ whose behavior we want to explain. We consider the **local post-hoc explanation** setting with **feature attributions**: For a point $x \in \mathbb{R}^d$, our goal is to explain which input features (or combinations thereof) were most influential in determining the "decision" $f(x)$. In order to do so, we assign real numbers to input features and their combinations. The higher the absolute value of this number, the more influential the feature is considered to be (for an illustration consider Figure 1).

In what follows, we denote $[n] = \{1, \dots, n\}$ and use subsets of coordinates $S = \{s_1, \dots, s_n\} \subset [d]$ to index both data points $x_S = (x_{s_1}, \dots, x_{s_n})$ and collections of functions $f_S(x_S) = f_{x_{s_1}, \dots, x_{s_n}}(x_{s_1}, \dots, x_{s_n})$ where we assume the ordering $s_1 < \dots < s_n$.

3.1 Value Functions and Shapley Values

For a data point $x \in \mathbb{R}^d$, a subset of coordinates $S \subset [d]$, and a function f, the *value function* $v(x, S)$ is supposed to quantify how much the features that are present in S contribute towards the prediction $f(x)$. Two important value functions are the *observational SHAP* value function Lundberg and Lee (2017)

$$v(x, S) = \mathbb{E}_{z \sim \mathcal{D}} \left[f(z) \,|\, x_S \right] \tag{1}$$

and the *interventional SHAP* value function (Chen et al., 2020; Janzing et al., 2020)

$$v(x, S) = \mathbb{E}_{z \sim \mathcal{D}} \left[f(z) \,|\, do(x_S) \right]. \tag{2}$$

Shapley Values, denoted by $\Phi_i(x)$, are obtained from the value function via the well-known Shapley formula (Shapley, 1953). We first introduce the Shapley Interaction Index (Grabisch and Roubens, 1999), given by $\Delta_S(x) =$

$$\sum_{T \subset [d] \setminus S} \frac{(d - |T| - |S|)!|T|!}{(d - |S| + 1)!} \sum_{L \subset S} (-1)^{|S|-|L|} v(x, L \cup T). \tag{3}$$

The Shapley Value $\Phi_i(x)$ of feature i at x is then simply given by $\Delta_i(x)$. Importantly, different value functions give rise to different Shapley Values, so that there effectively exists a multiplicity of possible Shapley Values, depending on our choice of value function (Sundararajan and Najmi, 2020). The popular KernelSHAP algorithm (Lundberg and Lee, 2017) approximates Shapley Values with respect to the interventional SHAP value function. The corresponding attributions are also known as the *SHAP feature attributions*. The following regularity condition, satisfied by both (1) and (2), will guarantee that the value function gives rise to a well-defined functional decomposition of the function that we attempt to explain.

Definition 1 (Subset-Compliant Value Function)**.** *We say that $v(x, S)$ is a subset-compliant value function for $f : \mathbb{R}^d \to \mathbb{R}$ if $v(x, [d]) = f(x)$ and if the value $v(x, S)$ depends only on those coordinates of x that are indexed by S. For a subset-compliant value function, we also write $v(x, S) = v(x_S, S)$.*

3.2 Generalized Additive Models

We employ the following definition of a generalized additive model (GAM) of order n.

Definition 2 (Generalized Additive Model of order n)**.** *We say that $f : \mathbb{R}^d \to \mathbb{R}$ is a generalized additive model of order n if f can be written in the form*

$$f(x) = \sum_{S \subset [d],\, |S| \leq n} f_S(x_S) \tag{4}$$

In words, the function f can be described as a simple sum with interaction terms of at most n variables at a time. The individual functions f_S are called component functions of f. GAMs with few interactions ($n = 1, 2, 3$) are often considered interpretable and called Glassbox-GAMs (Lou et al., 2012; Caruana et al., 2015). The reason for this is that the feature-wise shape functions $f_1, \dots, f_d$ can be easily visualized, see for example Figure 4.

If we allow for interactions of arbitrary order, that is $n = d$, then every function can be written as a GAM. However, it is a well-known fact that representing an arbitrary function according to (4) is under-determined: Many such representations might be possible for the same function. Any such representation is called a functional decomposition of f. This non-identifiability has led to the development of additional criteria on the decomposition, such as functional ANOVA, that resolve the identification problem (Hooker, 2007; Lengerich et al., 2020).

4 FROM SHAPLEY VALUES TO GENERALIZED ADDITIVE MODELS

We now introduce n-Shapley Values, a parametric family of local-post hoc explanation algorithms that extends Shapley Values (Lundberg and Lee, 2017) and Shapley Interaction Values (Lundberg et al., 2020). We then show that every subset-compliant value function implicitly provides a functional decomposition of the function that we attempt to

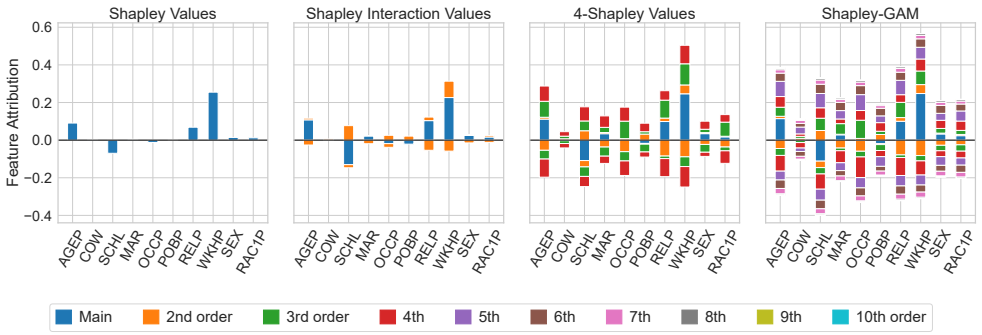


Figure 1: n-Shapley Values generate a sequence of explanations of increasing complexity, ranging from the original Shapley Values to the Shapley-GAM. From left to right: Shapley Values ($n = 1$), Shapley Interaction Values ($n = 2$), 4-Shapley Values ($n = 4$) and the Shapley-GAM ($n = d$). In each plot, we distributed the higher-order interaction effects uniformly onto all involved features (as justified by Theorem 6). Taking into account the signs of the attributions, the different contributions to each of the bars sum to the Shapley Value of that feature (Equation (13)). Taking the overall sum over all bars for all features recovers the prediction $f(x)$. See Appendix Section B for more details regarding this visualization. In this example, the function f is a random forest on the Folktables Income classification task, the data point is the first observation in our test set, and we used the value function of interventional SHAP.

explain. Due to its connection with Shapley Values, we denominate this decompositions the Shapley-GAM. We then show that for $n = d$, n-Shapley Values are equal to this decomposition.

4.1 n-Shapley Values

The definition of n-Shapley Values relates to the function f that we want to explain implicitly via the value function.

Definition 3 (n-Shapley Values)**.** *Fix a function $f : \mathbb{R}^d \to \mathbb{R}$. Let $v(x, S)$ be a value function for f. n-Shapley Values Φ_S^n provide an an attribution to all groups of at most n features at a time, that is for all sets $S \subset [d]$ with $|S| \leq n$. We define them recursively, starting from the original Shapley Values at $n = 1$ up to $n = d$, by*

$$\Phi_S^n = \begin{cases} \Delta_S & \text{if } |S| = n \\ \Phi_S^{n-1} + B_{n-|S|} \sum\limits_{\substack{K \subset [d] \setminus S \\ |K|+|S|=n}} \Delta_{S \cup K} & \text{if } |S| < n. \end{cases} \tag{5}$$

The coefficients B_n that balance the different terms are the Bernoulli numbers (see Appendix A). All terms except the Bernoulli numbers additionaly depend on the point x.

While this definition might seem rather abstract, n-Shapley Values are actually a straightforward extension of Shapley Interaction Values (Lundberg et al., 2020). These correspond to the case $n = 2$. The original Shapley Values correspond to the case $n = 1$. Similar to the original Shapley Values, n-Shapley Values are additive and always sum to the function value $f(x)$ (when summed over all subsets $S \subset [d]$) of size $\leq n$).[1] The overall intuition behind the recursive definition of n-Shapley Values is that starting from the original Shapley Values at $n = 1$, we successively add higher-order variable interactions to the explanations.

n-Shapley Values give rise to a sequence of explanations of increasing complexity, ranging from the original Shapley Values up to a functional decomposition of the function that we attempt to explain (see Theorem 4 below). Figure 1 depicts such a sequence of explanations for a random forest on the Folktables Income classification task (Ding et al., 2021). To visualize the n-Shapley Values, we evenly distribute all higher-order interactions onto the involved features. As we detail in Appendix B, this technique is justified by the recursive relationship between n-Shapley Values of different order. Note that n-Shapley Values of higher order are different from those of lower order only if the function that we attempt to explain actually contains higher-order variable interactions (this intuition will be made precise in Section 6). For this reason, n-Shapley Values can be used as a tool to assess the amount of variable interaction that is present in a given black-box predictor. For the random forest, we can see from the rightmost part of Figure 1 that it relies on very high degrees of variable interaction (for a quantitative analysis, see Section 7).

[1] The proof of Proposition 12 in the Appendix shows that the Bernoulli numbers are exactly the coefficients that balance equation (5) in this way.

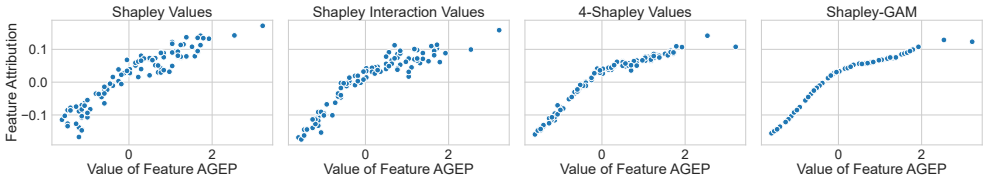


Figure 2: As $n \to d$, the n-Shapley Values provide increasingly precise representations of the component functions f_S of the Shapley-GAM. This figure depicts partial dependence plots of Φ^1_{AGEP} (Shapley Values, $n = 1$), Φ^2_{AGEP} (Shapley Interaction Values, $n = 2$), Φ^4_{AGEP} (4-Shapley Values, $n = 4$) and Φ^{10}_{AGEP} (Shapley-GAM, $n = d$). The leftmost partial dependence plot is the usual plot that is often used in order to visualize Shapley Values (Lundberg et al., 2020) (the plot depicts the original Shapley Values for the observations in the test set). It takes the often observed form where the Shapley Values are scattered around an overall functional relationship. Theorem 4 and Theorem 6 make this intuition precise by specifying how the Shapley Values are related to the component functions of the Shapley-GAM. The middle and right plots illustrate that as we move towards higher-order explanations, interaction effects can be appropriately represented. As a consequence, the partial dependence plots of individual feature attributions approach the component functions of the Shapley-GAM. In this example, the function f is a kNN classifier on the Folktables Income classification task. Appendix Figure K.8 depicts the partial dependence plots of all other features.

4.2 The Shapley-GAM

The following Theorem 4 shows two things. First, a subset-compliant value function gives rise to a well-defined functional decomposition. Second, d-Shapley Values are equal to this decomposition. The transformation of the value function that defines the decomposition is well-known as the Harsanyi Dividend (Harsanyi, 1982) or Möbius transform.

Theorem 4 (d-Shapley Values provide a functional decomposition of f)**.** *Fix a function $f : \mathbb{R}^d \to \mathbb{R}$. Let $v(x, S)$ be a subset-compliant value function for f. Then the d-Shapley Values represent the function f as a specific GAM that we denominate the Shapley-GAM. It is given by*

$$f(x) = \sum_{S \subset [d]} f_S(x_S) \tag{6}$$

with component functions

$$f_\emptyset = v(\emptyset) \quad \textit{and} \quad f_S(x_S) = \Phi^d_S(x) \tag{7}$$

where

$$\Phi^d_S(x) = \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L). \tag{8}$$

For intuition about Theorem 4, consider Figure 2. It is a well-known fact that the Shapley Value of feature i not only depends on the value of that feature, but also on the values of the other features of x (compare the leftmost partial dependence plot in Figure 2). The reason for this is that Shapley Values subsume higher-order variable interactions into the attributions of individual features (according to formula (11), as we will see below). Now, as we successively increase n, more and more variable interactions are appropriately represented in the explanations. This means that they no longer have to be subsumed into lower-order effects, which implies in turn that the lower-order components of the explanations become more distinct (middle parts of Figure 2). For $n = d$, all possible variable interactions can be represented in the explanations, which implies that d-Shapley Values become well-defined functions of the respective features (rightmost plot in Figure 2).

n-Shapley Values depend on the value function, and so does the associated functional decomposition. For the observational and interventional SHAP value functions, the functional decompositions are given as follows.

Corollary 5 (Observational and Interventional SHAP)**.** *For the observational SHAP value function* (1)*, the component functions of the Shapley-GAM are given by $f_\emptyset = \mathbb{E}[f]$,*

$$\begin{aligned} f_i(x_i) &= \mathbb{E}[f|x_i] - \mathbb{E}[f] \\ f_{i,j}(x) &= \mathbb{E}[f|x_i, x_j] - \mathbb{E}[f|x_i] - \mathbb{E}[f|x_j] + \mathbb{E}[f] \\ f_S(x_S) &= \sum_{L \subset S} (-1)^{|S|-|L|} \mathbb{E}[f|x_L]. \end{aligned} \tag{9}$$

For the interventional SHAP value function, the component functions are given by the same expression, but with the conditional expectations replaced by the causal do-operator.

As will see below (Theorem 7), there is actually a one-to-one relationship between subset-compliant value functions and different functional decompositions of f.

5 FROM GENERALIZED ADDITIVE MODELS TO SHAPLEY VALUES

In the previous section, we have seen that Shapley Values give rise to a functional decomposition of the original func-

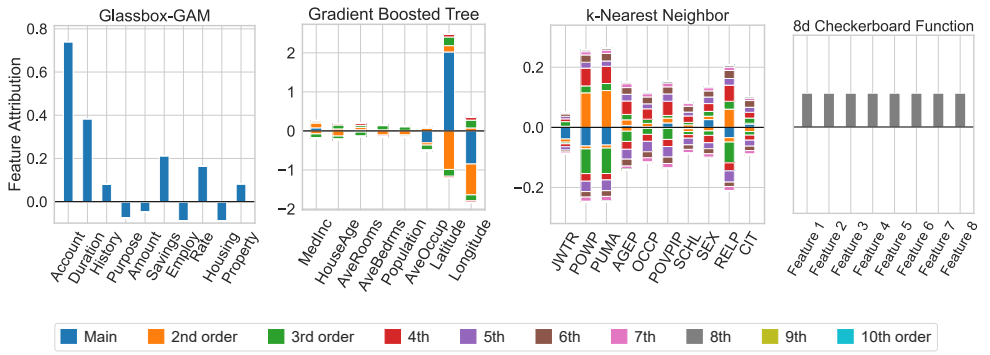


Figure 3: Visualizing the Shapley-GAM of interventional SHAP. Figures depict d-Shapley Values, visualized as in Figure 1. Different functions on different data sets require a different degree of variable interaction. (Left) A GAM without variable interactions on the German Credit data set. (Middle Left) A gradient boosted tree on the California Housing data set. (Middle Right) A kNN classifier on the Folktables Travel data set. (Right) The 8-dimensional checkerboard function (14). Additional figures for more data points and classifiers can be found in Appendix K.

tion (via the associated value function). In this section, we show that the original Shapley Values as well as n-Shapley Values of any order are linear combinations of the component functions of this decomposition. This provides a novel motivation for Shapley Values that does not require value functions or the Shapley formula. This alternative motivation of Shapley Values is equivalent to the original motivation via value functions: For every functional decomposition of f, there is a corresponding subset-compliant value function v such that the Shapley Values derived from the decomposition and v are equal (and vice-versa).

5.1 Shapley Values from the Shapley-GAM

Theorem 6 specifies the way in which the different component functions of the Shapley-GAM give rise to n-Shapley Values.

Theorem 6 (n-Shapley Values from the Shapley-GAM)**.** *Let $f(x) = \sum_{S\subset[d]} f_S(x_S)$ be the decomposition of f provided by the Shapley-GAM, and let $\Phi^n_S(x)$ be the n-Shapley Values of f. Then, it holds that*

$$\Phi^n_S = f_S + \sum_{\substack{K\subset[d]\setminus S \\ n+1\leq|S|+|K|}} C_{n-|S|,|K|}\, f_{S\cup K} \tag{10}$$

with coefficients $C_{n,m} = \sum_{k=0}^{n} \binom{n}{k} \frac{B_k}{1+m-k}$. Specifically, the Shapley Value of feature i is given by

$$\Phi^1_i = f_i + \cdots + \frac{1}{k+1} \sum_{S\subset[d]\setminus\{i\},|S|=k} f_{S\cup\{i\}} + \cdots + \frac{1}{d} f_{[d]} \tag{11}$$

where all terms additionally depend on the point x.

Theorem 6 specifies how higher-order variable interactions that are present in f are subsumed into lower-order explanations. In the case of the original Shapley Values, this is particularly intuitive: Higher-order effects are evenly distributed among the involved features.[2] Theorem 6 also specifies what information about the function f is and is not contained in Shapley Values. We see that different functions f can give rise to the same n-Shapley Values as long as $n < d$ (Grabisch, 2016). We also see that it is impossible to tell from individual Shapley Values whether the model consists of main effects or complex variable interactions. Furthermore, a feature can have zero attribution although it appears in multiple interaction effects with different signs.

For a bit more intuition about the Shapley-GAM, Figure 3 illustrates the Shapley-GAM of interventional SHAP for different functions. A main point is that different predictors require a different degree of variable interaction in order to be represented as a GAM. By definition, a Glassbox-GAM (leftmost part of Figure 3) does not require any variable interaction. The other extreme is the k-dimensional checkerboard function (14) (rightmost part of Figure 3), which only consists of interaction terms of order k. Many learned functions such gradient boosted trees (Figure 3, middle left) and the k-Nearest Neighbor (kNN) classifier (Figure 3, middle right) lie in between. Overall, there is a significant amount of variation between different methods and problems. This is also illustrated in many additional figures in Appendix K. For a quantitative analysis, see Section 7.

[2] For individual value functions, equation (11) is known in the literature on economic game theory (Grabisch, 1997)[Theorem 1]. Variants of it were independently re-discovered in Keevers (2020), Herren and Hahn (2022) and Hiabu et al. (2023).

5.2 From Functional Decompositions to Subset-Compliant Value Functions

We have show that every subset-compliant value function corresponds to a functional decomposition of f. We now show that the reverse is also true, that is every functional decomposition of f corresponds to a subset-compliant value function. The transformation that defines the value function is also known as the Zeta transform.

Theorem 7 (From Generalized Additive Models to Value Functions)**.** *Let $f(x) = \sum_{S\subset[d]} g_S(x)$ be any functional decomposition of f. Define the subset-compliant value function*

$$v(x, S) = \sum_{L\subset S} g_L(x). \tag{12}$$

Then the functional decomposition g_S is the Shapley-GAM with respect to the value function (12)*.*

Taken together, Theorem 4 and Theorem 7 establish a bijection between subset-compliant value functions and functional decompositions of f. In a sense, this implies that every functional decomposition implicitly corresponds to a notion of feature attribution via its associated value function and the Shapley formula (or, more directly, via equation (11) which is just the same).

6 RECOVERY

In this section, we connect Shapley Values with interpretable models by showing that n-Shapley Values, as well as the Shapley Taylor- and Faith-Shap interaction indices, recover GAMs. In order for this to be the case, the order of the explanation has to be at least as large as the order of the GAM.

Theorem 8 (Shapley-based Explanations Recover GAMs)**.** *Let f be a generalized additive model of order n. Assume that either*

(a) the value function is given by observational SHAP and the individual features are independent random variables, or

(b) the value function is given by interventional SHAP.

Then, n-Shapley Values, as well as the Shapley Taylor- and Faith-Shap interaction indices of order n, recover a representation of f as a GAM. In fact, all the interaction indices are equal to each other and given by

$$\Phi^n_S(x) = f_S(x_S)$$

where f_S are the component functions of the Shapley-GAM.

Theorem 8 implies that the SHAP feature attributions recover GAMs without variable interactions and that Shapley Interaction Values recover GAMs with interactions of at most two variables at a time. Unlike our previous results, Theorem 8 depends on the choice of the value function. This is because the recovery property holds if (1) the interaction index can be written like in equation (10), and (2) the Shapley-GAM is a GAM or order n — and the second point depends on the value function.

As it turns out, the independence assumption in part (a) of Theorem 8 is indeed necessary (see Appendix D). This is interesting insofar as it establishes the usefulness of the interventional SHAP value function from a purely statistical perspective, that is without any causal motivation (for a discussion about the differences between observational and interventional SHAP, see also Chen et al. (2020)).

Figure 4 (Top) illustrates the recovery result for a GAM without variable interactions. For this example, we explicitly resort to the default implementation of the Kernel SHAP algorithm, in order to see whether there is any significant approximation error (Kernel SHAP approximates the Shapley Values of the interventional SHAP value function). The top part of Figure 4 depicts the shape curve of the feature POWPUMA in the GAM (blue curve), as well as the associated Kernel SHAP values (red dots). The Kernel SHAP values lie almost exactly on the shape curve of the GAM, which means that the recovery property holds fairly precisely, at least in this simple example.

7 IS THERE AN ACCURACY-COMPLEXITY TRADE-OFF?

In the previous sections, we have outlined the connections between Shapley Values and GAMs on a theoretical level. In this section, as well as in the next section, we turn to more practical concerns. In this section, we investigate the number of variable interactions that are present in various standard classifiers. In order to do so, we rely on a number of low-dimensional data sets on which we can reliably estimate the Shapley-GAM decompositions of the different learned predictors (compare Section 8). It is interesting to compare this against the accuracy: Because models with more variable interactions can represent strictly more functions than models with less variable interactions, it is natural to suspect that more accurate classifiers might exhibit higher degrees of variable interaction (Dziugaite et al., 2020).

We suggest to measure the extent of variable interaction that is present in a given classifier with the following quantity

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\sum_{S\subset[d]} |S|\cdot|f_S(x_S)|\right] \Big/ \mathop{\mathbb{E}}_{x\sim\mathcal{D}}\left[\sum_{S\subset[d]} |f_S(x_S)|\right]. \tag{13}$$

where f_S are the component functions of the Shapley-GAM decomposition of f, using interventional SHAP.

Figure 4 (Middle) illustrates the relationship between the predictive accuracy and our measure (13) for different pre-

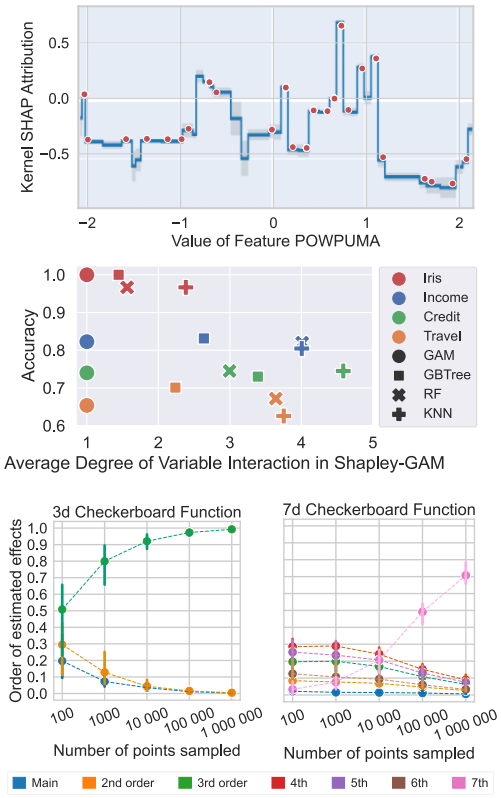


Figure 4: **Top**: Shapley Values recover GAMs without variable interactions (Theorem 8). To create this figure, we first trained a GAM on the Folktables Travel data set using the InterpretML package (Nori et al., 2019). We then computed the Kernel SHAP values for the decision function of the GAM using the `shap` package (Lundberg and Lee, 2017). For the feature POWPUMA, the Figure depicts the ground-truth variable effect in the GAM in blue, and the associated Kernel SHAP values for data points from the test set as red dots. We see that the red dots lie on the blue line, that is Kernel SHAP recovers the component function of the GAM. **Middle**: The average degree of variable interaction (13) in the Shapley-GAM of interventional SHAP for various standard classifiers. The figure depicts predictive accuracy versus the average degree of variable interaction. **Bottom**: Estimating higher-order variable interactions requires precise evaluations of the value function. A simple way to study this is by estimating the k-dimensional checkerboard function (14). *Left:* 3-way variable interactions can be precisely estimated. *Right:* 7-way variable interactions can be reliably detected, but precise estimation requires prohibitively many samples.

dictors f. The figure depicts four different kinds of classifiers: A Glassbox-GAM without variable interactions (Nori et al., 2019), a gradient boosted tree (Chen and Guestrin, 2016), a random forest, and a kNN classifier (Pedregosa et al., 2011). We compare these classifiers on four different data sets: Folktables Travel and Income (Ding et al., 2021), Iris, and German Credit. Details on the data sets and training procedures are in Appendix J.

As far as accuracy is concerned, we see from Figure 4 that GAMs without variable interactions perform fairly well against the more complicated classifiers — a fact that has often been observed in the literature (Caruana et al., 2015; Agarwal et al., 2021a). On the more complex data sets, however, there is usually a model with variable interactions and slightly better accuracy[3] As far as the degree of variable interaction is concerned, we see that there is a large amount of variation in between the different classifier.

Especially interesting is the kNN classifier. It tends to perform worse in terms of accuracy than the interpretable GAM, but exhibits very high degree of variable interaction. Observe that the kNN classifier can also be considered interpretable (by explaining the workings of the classifier and providing the k data points that are responsible for the classification). Therefore, this example shows that a high degree of variable interaction in the Shapley-GAM does not imply that a function is hard to explain per se.

This simple empirical investigation suggests that the relation between accuracy and the average degree of variable interaction in the Shapley-GAM is nuanced: While some degree of interaction seems necessary in order to achieve competitive accuracy, some classifiers seem to exhibit more interaction than that. In some cases, the correlation might even be negative (as for the kNN classifier).

8 COMPUTATION AND ESTIMATION

We now turn to the practical question of computing n-Shapley Values. In this work, we take the trivial approach and simply evaluate the value function for all possible subsets $S \subset [d]$, then combine the respective terms according to Definition 3. A Python package to compute n-Shapley Values, as well as the Shapley Taylor- and Faith-Shap interaction indices, is available `https://github.com/tml-tuebingen/nshap`. Even for the original Shapley Values, it is well-known that the number of required evaluations of the value function grows exponentially in the number of features. For this reason, there exist efficient approximations such as Kernel SHAP, as well as efficient implementations for certain function classes such as tree based models (Lundberg and Lee, 2017). We hold that

[3]The InterpretML package (Nori et al., 2019) allows to include interactions between pairs of variables which reportedly allows to be on par with black-box models on many data sets. Compare also (Lou et al., 2012).

such computationally efficient approximations are also be possible for n-Shapley Values.

Instead of focusing on the well-known computational aspect of the problem, we want to focus on the estimation aspect which seems much less studied. Note that n-Shapley Values are a statistic that is subject to sampling variation. The same is true for our visualizations (as in Figure 1), which are summary statistics of n-Shapley Values. This is because both the observational and the interventional SHAP value function require to estimate an expectation.

We now asses with a simple empirical analysis up to which order interaction effects can be estimated in practice. We consider the k-dimensional checkerboard function $B_k : [0,1]^d \to \{0,1\}$ given by

$$B_k(x_1, \dots, x_d) = \begin{cases} 0 & \text{if } \sum_{i=0}^{k} \lfloor (\lambda \cdot x_i) \rfloor \mod 2 = 0 \\ 1 & \text{otherwise} \end{cases} \tag{14}$$

where $\lambda > 2$ parameterizes the number of checkers along each axis. If data points are uniformly distributed in the unit cube $[0,1]^d$, then the Shapely-GAM of interventional SHAP of B_k is given by the single k-th order interaction effect $f_{x_1,\dots,x_k}(x_1, \dots, x_k) = B_k(x_1, \dots, x_k, 0, \dots, 0)$. The question now is how precisely we have to estimate the expectation $\mathbb{E}_{z \sim \mathcal{D}}[f(z) \,|\, do(x_S)]$ if we want to precisely estimate a kth-order interaction effect.

The bottom part of Figure 4 depicts the result of estimating 10-Shapley Values when the underlying function is the 3- or 7-dimensional checkerboard function, respectively. The x-axis depicts the number of samples used to estimate the value function, ranging from 100 to 1 000 000. The y-axis depicts the order of the estimated effects, with confidence bands that account for 5 randomly sampled data sets. From the figure, we observe that if the number of samples is small in relation to the magnitude of the interaction effect, then the estimation results in spurious lower-order effects. For $k = 3$, these effects vanish with sufficiently many samples, which means that the checkerboard function is precisely estimated. For $k = 7$, the presence of the higher-order interaction effect can be reliably detected, but not precisely estimated given reasonably many samples.

In this simple analysis, we see that interaction effects of order larger that 2 can be precisely estimated given sufficiently many samples. We also see that functions with high-order interactions are difficult to estimate and can result in artifacts. Figures for all interaction orders $k = 2, \dots, 10$ and a discussion of the precision of the depicted visualizations of n-Shapley Values can be found in Appendix C.

9 DISCUSSION

This work provides a functionally-grounded characterization of Shapley Values as they are being used in explainable machine learning (Doshi-Velez and Kim, 2017). Explainable machine learning is often believed to be an important component in societal applications of machine learning (Wachter et al., 2017; Kaminski and Urban, 2021; Kästner et al., 2021). At the same time, current approaches face a lot of criticism, for example because they are non-robust or unable to provide the desired level of model understanding (as well as for a variety of other concerns) (Lipton, 2018; Kumar et al., 2020; Slack et al., 2020; Bordt et al., 2022). In this situation, we believe that a precise understanding of the mechanics of popular explainability methods, such as the one presented in this work, is a good first step toward an informed discussion of what we can and cannot achieve.

Some of our results stand in contrast to conventional wisdom around Shapley Values, and offer a novel perspective on local-post hoc explanation algorithms. For example, we have seen that Shapley Values depend on the coordinates of the point that we attempt to explain, but not on the local neighbourhood of that point — the recovery example with the step function in Figure 4 suggests that this is also the case for the approximations of the Shapley Value that are used in practice. We have further seen that the original Shapley Values are able to faithfully explain non-linear functions, as long as the non-linearity is restricted to the specific form permitted by GAMs. As such, our results highlight the differences between Shapley Values and other feature attribution methods, for example those that are related to the gradient (Garreau and von Luxburg, 2020; Agarwal et al., 2021b), and those that perform local function approximation (Han et al., 2022).

The demonstrated connections between value functions and functional decompositions effectively link the literature on feature attributions with the tools developed in the statistics literature on functional decompositions (Hooker, 2007; Lengerich et al., 2020). This raises the question of whether criteria for functional decompositions can be useful to understand feature attributions. Here, two concurrent works made significant contributions: Hiabu et al. (2023) show that the value function of interventional SHAP can be motivated with a causal assumption on the associated functional decomposition. Herren and Hahn (2022) outline connections between observational SHAP and functional ANOVA.

While our work gives a functionally-grounded analysis of Shapley Values for any function, as well as recovery guarantees for Shapley Values and GAMs, we do not claim that Shapley Values are an appropriate post-hoc explanation method for any function (Kumar et al., 2021; Tan et al., 2022). Instead, the purpose of our work is to highlight the connections between a post-hoc explanation method and a class of interpretable models. Overall, however, we believe that many properties of Shapley Values have the potential to be more clearly understood using our perspective of functional decompositions.

Acknowledgements

This work was done in part while Sebastian was visiting the Simons Institute for the Theory of Computing. Sebastian would like to thank Rich Caruana, Gyorgy Turan, Michal Moshkovitz and Tosca Lechner for many fruitful discussions about variable interactions. The authors would also like to thank Markus Scheuer and René Gy for linking Lemma 10 to the literature on Bernoulli numbers, and the anonymous reviewers whose comments helped to improve this paper. This work has been supported by the German Research Foundation through the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 number 390727645), the BMBF Tübingen AI Center (FKZ: 01IS18039A), and the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. Neural additive models: Interpretable machine learning with neural nets. In *NeurIPS*, 2021a.

S. Agarwal, S. Jabbari, C. Agarwal, S. Upadhyay, S. Wu, and H. Lakkaraju. Towards the unification and robustness of perturbation and gradient based explanations. In *ICML*, 2021b.

E. Albini, J. Long, D. Dervovic, and D. Magazzeni. Counterfactual shapley additive explanations. In *ACM FAccT*, 2022.

S. Bordt, M. Finck, E. Raidl, and U. von Luxburg. Post-hoc explanations fail to achieve their purpose in adversarial contexts. In *ACM FAccT*, 2022.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2015.

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

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016.

I. Covert, S. Lundberg, and S. Lee. Explaining by removing: A unified framework for model explanation. *JMLR*, 2021.

F. Ding, M. Hardt, J. Miller, and L. Schmidt. Retiring adult: New datasets for fair machine learning. In *NeurIPS*, 2021.

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

G. K. Dziugaite, S. Ben-David, and D. M. Roy. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*, 2020.

D. Garreau and U. von Luxburg. Explaining the explainer: A first theoretical analysis of lime. In *AISTATS*, 2020.

H. Gould and J. Quaintance. Bernoulli numbers and a new binomial transform identity. *J. Integer Seq.*, 2014.

M. Grabisch. K-order additive discrete fuzzy measures and their representation. *Fuzzy sets and systems*, 1997.

M. Grabisch. Bases and transforms of set functions. In *On Logical, Algebraic, and Probabilistic Aspects of Fuzzy Set Theory*. Springer, 2016.

M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 1999.

R. Gy. Combinatorial identity involving bernoulli numbers. Mathematics Stack Exchange, 2022. URL `https://math.stackexchange.com/q/4520567`.

T. Han, S. Srinivas, and H. Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *arXiv preprint arXiv:2206.01254*, 2022.

J. C. Harsanyi. A simplified bargaining model for the n-person cooperative game. In *Papers in game theory*. Springer, 1982.

T. Hastie and R. Tibshirani. *Generalized Additive Models. Chapman Hall & CRC*. 1990.

A. Herren and P. R. Hahn. Statistical aspects of SHAP: Functional ANOVA for model interpretation. *arXiv preprint arXiv:2208.09970*, 2022.

T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *NeurIPS*, 2020.

M. Hiabu, J. T. Meyer, and M. N. Wright. Unifying local and global model explanations by functional decomposition of low dimensional structures. In *AISTATS*, 2023.

A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek. xxai-beyond explainable artificial intelligence. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*. Springer, 2022.

G. Hooker. Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *Journal of Computational and Graphical Statistics*, 2007.

D. Janzing, L. Minorics, and P. Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *AISTATS*, 2020.

N. Jethani, M. Sudarshan, I. C. Covert, S.-I. Lee, and R. Ranganath. Fastshap: Real-time shapley value estimation. In *ICLR*, 2021.

M. Kaminski and J. Urban. The right to contest ai. *Columbia Law Review*, 2021.

L. Kästner, M. Langer, V. Lazar, A. Schomäcker, T. Speith, and S. Sterz. On the relation of trust and explainability: Why to engineer for trustworthiness. In *IEEE 29th International Requirements Engineering Conference Workshops (REW)*, 2021.

T. L. Keevers. A power series expansion of feature importance. *Technical report*, 2020.

R. Kommiya Mothilal, D. Mahajan, C. Tan, and A. Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2021.

S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

I. Kumar, C. Scheidegger, S. Venkatasubramanian, and S. Friedler. Shapley residuals: Quantifying the limits of the shapley value for explanations. *NeurIPS*, 2021.

I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with shapley-value-based explanations as feature importance measures. In *ICML*, 2020.

H. Lakkaraju, J. Adebayo, and S. Singh. Explaining machine learning predictions: State-of-the-art, challenges, opportunities. Tutorial at NeurIPS, 2020.

E. Lee, D. Braines, M. Stiffler, A. Hudler, and D. Harborne. Developing the sensitivity of lime for better machine learning explanation. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, 2019.

B. Lengerich, S. Tan, C.-H. Chang, G. Hooker, and R. Caruana. Purifying interaction effects with the functional anova: An efficient algorithm for recovering identifiable additive models. In *AISTATS*, 2020.

B. J. Lengerich, R. Caruana, M. E. Nunnally, and M. Kellis. Death by round numbers and sharp thresholds: How to avoid dangerous ai ehr recommendations. *medRxiv*, 2022.

J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer. Generalized and scalable optimal sparse decision trees. In *ICML*, 2020.

Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018.

Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2012.

Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2013.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *NeurIPS*, 2017.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2020.

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

M. Moshkovitz, S. Dasgupta, C. Rashtchian, and N. Frost. Explainable k-means and k-medians clustering. In *ICML*, 2020.

H. Nori, S. Jenkins, P. Koch, and R. Caruana. Interpretml: A unified framework for machine learning interpretability. *arXiv preprint arXiv:1909.09223*, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 2011.

M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you? explaining the predictions of any classifier. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 2022.

W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

L. Shapley. A value for n-person games., 1953.

D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

D. Slack, A. Hilgard, S. Singh, and H. Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *NeurIPS*, 2021.

M. Sundararajan and A. Najmi. The many shapley values for model explanation. In *ICML*, 2020.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

M. Sundararajan, K. Dhamdhere, and A. Agarwal. The shapley taylor interaction index. In *ICML*, 2020.

Y. S. Tan, A. Agarwal, and B. Yu. A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds. In *AISTATS*, 2022.

C.-P. Tsai, C.-K. Yeh, and P. Ravikumar. Faith-shap: The faithful shapley interaction index. *arXiv preprint arXiv:2203.00870*, 2022.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

F. Wang and C. Rudin. Falling rule lists. In *AISTATS*, 2015.

Z. J. Wang, A. Kale, H. Nori, P. Stella, M. E. Nunnally, D. H. Chau, M. Vorvoreanu, J. Wortman Vaughan, and R. Caruana. Interpretability, then what? editing machine learning models to reflect human knowledge and values. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

A n-Shapley Values

This section details the properties of n-Shapley Values.

A.1 Bernoulli numbers

The Bernoulli numbers[1] B_n are defined by $B_0 = 1$ and

$$\sum_{k=0}^{n} \binom{n+1}{k} B_k = 0 \quad \forall n \geq 1. \tag{15}$$

In this paper, the Bernoulli numbers arise as the coefficients that make n-Shapley Values sum to the prediction (Proposition 12). In fact, equation (15) arises directly from the proof of Proposition 12. The Bernoulli numbers can be computed recursively by re-writing into (15)

$$B_n = \frac{-1}{n+1} \sum_{k=0}^{n-1} B_k \binom{n+1}{k} \quad \forall n \geq 1. \tag{16}$$

In a certain sense, the entire combinatorics around n-Shapley Values relies on the properties of the Bernoulli numbers. In particular, the proofs of Theorem 4 and Theorem 6 rely on the following two Lemmas.

Lemma 9. *For all $n \geq 1$, it holds that*

$$\sum_{k=1}^{n} \frac{B_k}{n-k+1} \binom{n}{k} = \frac{-1}{n+1}. \tag{17}$$

Proof. We re-arrange the sum to get

$$\sum_{k=1}^{n} \frac{B_k}{n-k+1} \binom{n}{k} = \frac{1}{n+1} \sum_{k=0}^{n} \binom{n+1}{k} B_k - \frac{B_0}{n+1} = \frac{-1}{n+1} \tag{18}$$

where the second equality follows from (15). □

Lemma 10. *For all $n, m \geq 0$, it holds that*

$$\sum_{k=0}^{n} \sum_{l=0}^{m} \binom{n}{k} \binom{m}{l} \frac{(n-k)!(m-l)!}{(n+m-k-l+1)!} (-1)^l B_{k+l} = \begin{cases} 1 & \text{if } n = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{19}$$

Lemma 10 follows from standard results for the Bernoulli numbers (Gould and Quaintance, 2014)[Theorem 2]. A proof is contained in Appendix I.

A.2 Additivity and Efficiency

From the recursive definition of the n-Shapley Values in Definition 3, a straightforward calculation shows that

$$\Phi_S^n(x) = \sum_{k=0}^{n-|S|} \sum_{K \subset [d] \setminus S, \, |K|=k} B_k \, \Delta_{S \cup K}(x) \tag{20}$$

which is an alternative non-recursive definition of n-Shapley Values.

[1] An introduction and discussion about Bernoulli numbers can be found, for example, in the corresponding Wikipedia article at `https://en.wikipedia.org/wiki/Bernoulli_number`.

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
B_n	1	$\frac{-1}{2}$	$\frac{1}{6}$	0	$\frac{-1}{30}$	0	$\frac{1}{42}$	0	$\frac{-1}{30}$	0	$\frac{5}{66}$	0	$\frac{-691}{2730}$	0	$\frac{7}{6}$	0	$\frac{-3617}{510}$	0	$\frac{43867}{798}$	0

Table A.1: The first 20 Bernoulli numbers.

Proposition 11 (Additivity)**.** *For all* $1 \leq n \leq d$ *and all* $f, g : \mathbb{R}^n \to \mathbb{R}$*, we have*

$$\Phi_S^n(x; f + g) = \Phi_S^n(x; f) + \Phi_S^n(x; g). \tag{21}$$

Proof. By definition, Φ_S^n is linear in Δ_S, and Δ_S is linear in the value function v. Therefore, the linearity of Φ_S^n in f follows from the linearity of v in f, i.e. from the fact that $v_{f+g}(x, S) = v_f(x, S) + v_g(x, S)$. □

Proposition 12 (Efficiency)**.** *For all* $1 \leq n \leq d$*, it holds that*

$$\sum_{\substack{S \subset [d] \\ 1 \leq |S| \leq n}} \Phi_S^n(x) = v([d]) - v(\emptyset). \tag{22}$$

Proof. For $n = 1$, the statement follows from the efficiency of the original Shapley Values. We assume that the statement holds for $n - 1$ and re-arrange the sum

$$\begin{aligned}
\sum_{\substack{S \subset [d] \\ 1 \leq |S| \leq n}} \Phi_S^n(x) &= \sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \Phi_S^n(x) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Phi_S^n(x) \\
&= \sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \left(\Phi_S^{n-1}(x) + B_{n-|S|} \sum_{\substack{K \subset [d] \setminus S \\ |K| + |S| = n}} \Delta_{S \cup K}(x) \right) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Delta_S(x) \\
&= \sum_{\substack{S \subset [d] \\ 1 \leq |S| \leq n-1}} \Phi_S^{n-1}(x) + \sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \sum_{\substack{K \subset [d] \setminus S \\ |K| + |S| = n}} B_{n-|S|} \, \Delta_{S \cup K}(x) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Delta_S(x).
\end{aligned} \tag{23}$$

Notice that the first term is equivalent to $v([d]) - v(\emptyset)$ by the induction hypothesis. It remains to show that

$$\sum_{\substack{S \subset [d] \\ 1 \leq |S| < n}} \sum_{\substack{K \subset [d] \setminus S \\ |K| + |S| = n}} B_{n-|S|} \, \Delta_{S \cup K}(x) + \sum_{\substack{S \subset [d] \\ |S| = n}} \Delta_S(x) = 0. \tag{24}$$

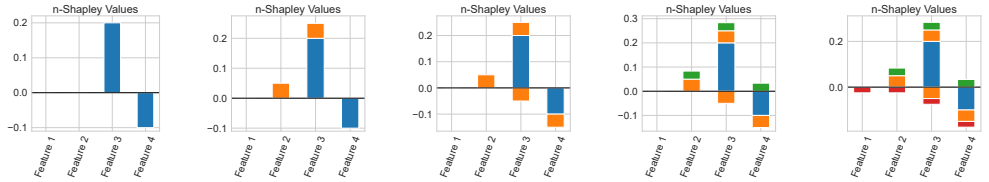
Notice that both sums are over sets of length n. In the first sum, each sets occurs multiple times. In the second sum, each set occurs exactly once. By counting the occurrences of each set in the first sum we see that (24) holds if

$$\sum_{s=1}^{n-1} B_{n-s} \binom{n}{s} + 1 = 0. \tag{25}$$

If we set $B_0 = 1$, this holds if and only if

$$\sum_{k=0}^{n-1} B_k \binom{n}{k} = 0, \tag{26}$$

which is the defining property of the Bernoulli numbers (15). In summary, we see that the Bernoulli numbers are the coefficients that balance the terms in the first sum in equation (24). □



(a) Example 1 (b) Example 2 (c) Example 3 (d) Example 4 (e) Example 5

Figure B.1: Examples that illustrate the proposed visualization technique for n-Shapley Values.

A.3 Relationship Between n-Shapley Values of Different Order

The following proposition is a straightforward extension of Theorem 6.

Proposition 13 (Relationship Between n-Shapley Values of Different Order)**.** *For $m \leq n$, let Φ_S^m and Φ_S^n be the m- and n-Shapley Values, respectively. Then, the m-Shapley Values can be computed from the n-Shapley Values by*

$$\Phi_S^m(x) = \Phi_S^n + \sum_{\substack{K \subset [d] \setminus S, \\ m-|S| < |K| \leq n-|S|}} \beta_{m-|S|,|K|} \, \Phi_{S \cup K}^n(x). \tag{27}$$

Specifically, it holds that

$$\Phi_i^1 = \Phi_i^n + \frac{1}{2} \sum_{j \neq i} \Phi_{i,j}^n + \cdots + \frac{1}{n} \sum_{\substack{K \subset [d] \setminus \{i\} \\ |K| = n-1}} \Phi_{K \cup i}^n \tag{28}$$

which is the basis for the visualizations in the paper.

Proof. The proposition follows from the counting argument used in the proof of Theorem 6. □

B Visualizing n-Shapley Values

Due to the large number of terms involved in n-Shapley Values of higher order, visualizing these explanations is difficult. However, Proposition 13 (which is really a variant of Theorem 6) states that higher-order variable interactions in n-Shapley Values are related to the original Shapley Values via a simple lump-sum formula. This gives rise to the idea of simply visualizing, for each feature, the respective components of the sum.

To illustrate this idea, let us consider a simple example. Let us begin with four different features and the usual Shapley Values. Say the first two features have attribution zero, the third feature has attribution 0.2, and the fourth feature has attribution -0.1. These Shapley Values can be visualized as usual, depicted in Figure B.1a. Now, let us add a second-order interaction effect, say $\Phi_{2,3}^2 = 0.1$. Because this interaction effect would ultimately be added to the attributions of feature 2 and feature 3 with a factor of $\frac{1}{2}$, let us simply add two corresponding bars to the attributions of these features, with the color indicating that it is a second-order effect. From the resulting Figure B.1b, it can then be seen that we have two main effects and a single positive interaction effect between features 2 and 3. If there were another interaction effect, say $\Phi_{3,4}^2 = -0.1$, we would proceed in the same way, taking care of the sign. From the resulting Figure B.1c, it can be seen that there are two main effects and a number of second-order interactions. With higher-order interactions we proceed accordingly, as illustrated for $\Phi_{2,3,4}^3 = 0.1$ (Figure B.1d) and $\Phi_{1,2,3,4}^4 = -0.1$ (Figure B.1d).

Note that while this form of visualization faithfully depicts the relative magnitude of the different variable interactions, it is in general not possible to tell from the figures which variables interact with each other, for example when there are a number of different second-order effects.

C Estimating n-Shapley Values

Here we collect some additional details regarding the estimation of n-Shapley Values. We note that the discussion here is not exhaustive. Our objective is to (1) raise awareness for the fact that computing n-Shapley Values incurs an estimation

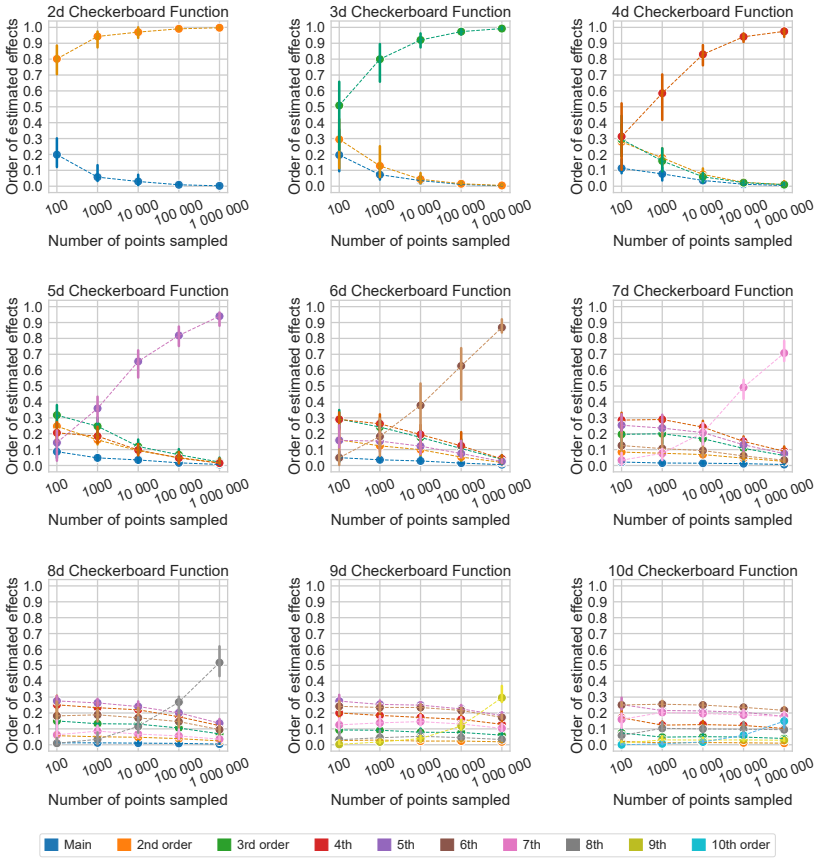


Figure B.2: Estimating higher-order variable interactions requires precise evaluations of the value function. A simple way to study this is by estimating the k-dimensional checkerboard function (14). Compare Figure 4 in the main paper.

problem, and (2) ensure that the results presented in the main paper are precisely estimated and not statistical artifacts.

Figure B.2 depicts the result of estimating the k-dimensional checkerboard function (14) for all values $k = 2, \ldots, 10$ (compare Section 8 in the main paper). As already discussed in the main paper, we can see from the figure that estimation becomes gradually harder as we increase the order of interaction.

In Figure C.3, we assess the degree up to which our visualizations are effected by the presence of spurious interaction effect of intermediate order, as observed when estimating a checkerboard function with too few samples. The figure visualizes the Shapley-GAM decomposition of a kNN classifier on the Folktables Travel data set, estimated with 500, 5000 and 133549 samples per evaluation of the value function, respectively. By comparing the left and middle part of Figure C.3 (estimation with 500 and 5000 samples, respectively), we see that 500 samples are to few and result in the presence of spurious interaction effects, for example of of order 4 and 5. This can be seen from the fact that some of these effects vanish as we increase the number of samples. By comparing the middle and right part of Figure C.3 (estimation with 5000 and 133549 samples, respectively), we see that estimation with 5000 samples is already quite precise for this kNN classifier. This can be seen from the fact that significantly increasing the number of samples does not have any significant effect on the

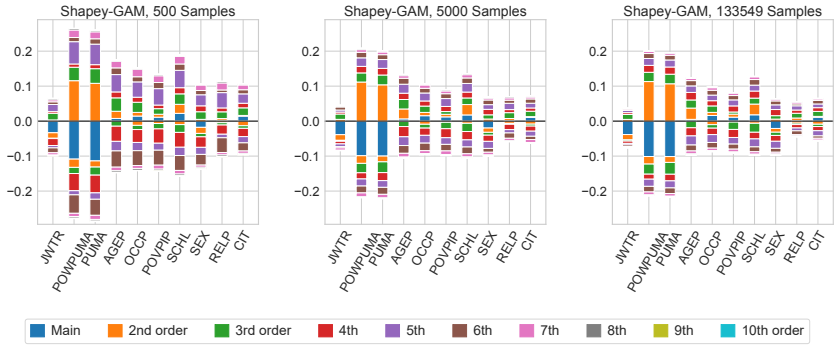


Figure C.3: Estimating higher-order interactions with too few samples can result in spurious interaction effects of intermediate order. These effects are also visible in our visualizations. **Left:** Estimation with 500 samples per evaluation of the value function results in spurious interaction effects. **Middle:** This can be seen from the fact that parts of the estimated effects vanish if we increase the number of samples to 5000 per evaluation of the value function. **Right:** Using all 133549 observations in the training data per evaluation of the value function, we get almost the same visualization as for 5000 samples. The function in this example is a kNN classifier and the data set is the Folktables Travel data set.

visualization.[2]

Table K.2 depicts the individual terms that underlie the visualization in Figure C.3. From Table K.2, we see that main effects are precisely estimated even with 500 samples. However, many relatively small higher-order coefficients are not very precisely estimated even for $N = 5000$. Note that the latter point is not in contrast to the fact that Figure C.3 is precisely estimated for $N = 5000$. Figure C.3 depicts summary statistics that are more precisely estimated than the individual components.

D The Statistical Independence Assumption for Observational SHAP is Necessary

In this section we give a simple example to demonstrate that the assumption of independent random variables for the observational SHAP value function in Theorem 8 is indeed necessary.

Consider the GAM of order 1

$$f(x_1, x_2) = x_1 + x_2.$$

Assume that x_1 and x_2 are correlated normal random variables

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, & \rho \\ \rho, & 1 \end{pmatrix} \right)$$

with $0 \leq \rho \leq 1$. We have

$$\mathbb{E}[x_2|x_1] = \rho x_1.$$

A simple calculation shows that the Shapley-GAM of observational SHAP is given by

$$f_\emptyset = 0, \quad f_1(x_1) = (1+\rho)x_1, \quad f_2(x_2) = (1+\rho)x_2, \quad f(x_1, x_2) = -\rho(x_1 + x_2).$$

According to Theorem 6, the observational SHAP values are then given by

$$\Phi_1 = (1 + \frac{\rho}{2})x_1 - \frac{\rho}{2}x_2, \qquad \Phi_2 = (1 + \frac{\rho}{2})x_2 - \frac{\rho}{2}x_1.$$

Clearly, recovery does not hold: Despite the fact that the underlying function is a GAM of order 1, the Shapley-GAM is a GAM of order 2. The Shapley Values also depend on both coordinates – hence they are not well-defined functions of the individual coordinates.

[2]This could of course be discussed much more rigorously.

In contrast, the Shapley-GAM of the interventional SHAP value function is given by

$$f_\emptyset = 0, \quad f_1(x_1) = x_1, \quad f_2(x_2) = x_2.$$

Moreover, the interventional SHAP values are given by

$$\Phi_1 = x_1, \qquad \Phi_2 = x_2,$$

that is recovery holds with the interventional SHAP value function (as guaranteed by Theorem 8).

E Proof of Theorem 4

Proof of Theorem 4. We are going to show that

$$\Phi_S^d(x) = \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L). \tag{29}$$

Note that the RHS evaluates the value function v only for sets $L \subset S$. From the assumption that the value function is subset-compliant, it follows that the RHS is a well-defined function of x_S. According to Proposition 12 (efficiency), the d-Shapley Values sum to $v(x) - v(\emptyset)$ which implies the Theorem.

To show (29), we consider the non-recursive definition of n-Shapley Values 20 and then substitute the definition of $\Delta_S(x)$ from Definition 3.

$$\begin{aligned}
\Phi_S^d(x) &= \sum_{k=0}^{d-|S|} \sum_{K \subset [d]\setminus S, |K|=k} B_k \, \Delta_{S\cup K}(x) \\
&= \sum_{k=0}^{d-|S|} \sum_{K \subset [d]\setminus S, |K|=k} B_k \sum_{T \subset [d]\setminus(S\cup K)} \frac{(d-|T|-|S|-|K|)!|T|!}{(d-|S|-|K|+1)!} \sum_{L \subset S\cup K} (-1)^{|S|+|K|-|L|} v(x,\, L\cup T). \qquad (30) \\
&= \sum_{K \subset [d]\setminus S} \sum_{T \subset [d]\setminus(S\cup K)} B_{|K|} \frac{(d-|T|-|S|-|K|)!|T|!}{(d-|S|-|K|+1)!} \sum_{L \subset S\cup K} (-1)^{|S|+|K|-|L|} v(x,\, L\cup T).
\end{aligned}$$

Where the last equation follows from the realization that we are summing over all possible subsets of $[d] \setminus S$.

In equation (30), we are summing over the value of the same sets multiple times. Let us fix a set $M = L \cup T$ and count how often it occurs in the sum. First note that $v(x, M)$ occurs exactly once for every set K, namely by choosing $T = M \setminus (S \cup K)$ and $L = M \cap (S \cup K)$. Since the coefficients do not only depend on the size of K, but also on $|T|$ and $|L|$, let us partition the set $K = K_1 \cup K_2 = \{K \cap M\} \cup \{K \setminus M\}$. Let $n_1 = |M \setminus S|$ and $n_2 = |[d] \setminus (S \cup M)|$ denote the maximum sizes of both partitions. With this counting argument, we arrive at

$$(-1)^{|S|-|M|} \sum_{K_1 \subset M\setminus S} \sum_{K_2 \subset [d]\setminus(S\cup M)} B_{|K_1|+|K_2|} \frac{(n_2-|K_2|)!(n_1-|K_1|)!}{(n_1+n_2-|K_1|-|K_2|+1)!} (-1)^{|K_2|} \tag{31}$$

occurrences of the term $v(x, M)$. Notice that equation (31) is equal to

$$(-1)^{|S|-|M|} \sum_{k_1=0}^{n_1} \sum_{k_2=0}^{n_2} \binom{n_1}{k_1} \binom{n_2}{k_2} \frac{(n_2-k_2)!(n_1-k1)!}{(n_1+n_2-k_1-k_2+1)!} (-1)^{k_2} B_{k_1+k_2} \tag{32}$$

The desired result now follows from the properties of the Bernoulli numbers. In particular, since $M \subset S \iff n_1 = 0$, we see from Lemma 10 that (32) equals $(-1)^{|S|-|M|}$ if $M \subset S$ and 0 otherwise. Comparing the terms for all possible sets $M \subset [d]$, we see that (30) equals (29).

Note that if we fix the point x, then the Shapley-GAM at x is equivalent to the Moebious transform of the measure $v(x, \cdot)$. From this perspective, Theorem 4 can be seen as an application of Theorem 2 in Grabisch (1997). □

F Proof of Theorem 6

Proof of Theorem 6. According to Theorem 4, the d-Shapley Values can be written as

$$\Phi_S^d(x) = f_S(x) \tag{33}$$

where $f_S(x)$ are the component functions of the Shapley-GAM. Hence, the d-Shapley Values are a linear combination of the component functions of the Shapley-GAM. From the recursive definition of the n-Shapley Values, we see that

$$\Phi_S^n(x) = \Phi_S^{n+1}(x) - B_{1+n-|S|} \sum_{K \subset [d] \setminus S, |K|+|S|=n+1} \Phi_{S \cup K}^{n+1}(x) \tag{34}$$

that is the n-Shapley Values are a linear combination of the terms involved in the $n+1$-Shapley Values. By induction, we see that the n-Shapley Values are linear combinations of the component functions of the Shapley-GAM.

It remains to determine the coefficients $C_{n,m}$. We present a counting argument that is based on the recurrence relation (34). In this counting argument, we first determine the coefficients $D_{n,m}$ where the first index corresponds to the distance between $|S|$ and the order of the Shapley Values, and the second index corresponds to the different between the size of the interaction effect and the order of the Shapley Values. Suppose that we are computing n-Shapley Values. If we use equation (34) to proceed recursively from d-Shapley Values to n-Shapley Values, then the first time that the component function $f_{S \cup K}$ is being added to Φ_S^m is during the computation of the $(|S| + |K| - 1)$-Shapley Values. According to equation (34), the linear coefficient will simply be $D_{|K|-1,1} = -B_{|K|}$. The second time that the component function $f_{S \cup K}$ is being added to Φ_S^m is during the computation of the $(|S| + |K| - 2)$-Shapley Values. This is because we have previously added $-B_1 f_{S \cup K}$ to all the terms of order $|S| + |K| - 1$ that are a subset of $S \cup K$. There are $\binom{|K|}{1}$ such terms, and we are now adding all of them to f_S, using the coefficient $-B_{|K|-1}$. This means that we arrive at a total coefficient of

$$D_{|K|-2,2} = -B_{|K|} + B_{|K|-1}\binom{|K|}{1}B_1. \tag{35}$$

By a similar argument we arrive at a coefficient of

$$D_{|K|-3,3} = -B_{|K|} + B_{|K|-1}\binom{|K|}{1}B_1 - B_{|K|-2}\binom{|K|}{2}B_2 - B_{|K|-2}\binom{|K|}{2}B_1\binom{2}{1}B_1. \tag{36}$$

for the $(|S| + |K| - 3)$-Shapley Values. In general, that is when we compute n-Shapley Values, the component function $f_{S \cup K}$ is being added to Φ_S^n once for every possible pathway that goes from a set of order $n+1$ to the set $S \cup K$ by successively adding different numbers of elements. For $k \geq 1$, let

$$P_k = \left\{ (p_1, \dots, p_k) \in \mathbb{N}_{\geq 0}^k \;\middle|\; \sum_{i=1}^k p_i = k \quad \text{and} \quad p_i = 0 \implies (p_j = 0 \, \forall j > i) \right\} \tag{37}$$

be the set of pathways of length k. This means that we have $P_1 = \big\{(1)\big\}$,

$$\begin{aligned} P_2 &= \big\{(2,0),(1,1)\big\}, \\ P_3 &= \big\{(3,0,0),(2,1,0),(1,2,0),(1,1,1)\big\}, \\ P_4 &= \big\{(4,0,0,0),(3,1,0,0),(2,2,0,0),(2,1,1,0), \\ &\qquad (1,3,0,0),(1,2,1,0),(1,1,2,0),(1,1,1,1)\big\} \end{aligned} \tag{38}$$

and so on. By accounting for the coefficients B_k and the signs along each path, the coefficients can be written as

$$D_{n,m} = \sum_{(p_1,\dots,p_m) \in P_m} (-1)^{\sum_{i=1}^m \text{sign}(p_i)} \binom{n+m}{n+p_1} B_{n+p_1} \prod_{i=2}^m B_{p_i} \binom{m - \sum_{j=1}^{i-1} p_j}{p_i} \tag{39}$$

From this, we derive the special case

$$
\begin{aligned}
D_{0,m} &= \sum_{(p_1,\ldots,p_m)\in P_m} (-1)^{\sum_{i=1}^m \text{sign}(p_i)} \binom{m}{i_1} B_{p_1} \prod_{i=2}^{m} B_{p_i} \binom{m - \sum_{j=1}^{i-1} p_j}{p_i} \\
&= \sum_{(p_1,\ldots,p_m)\in P_m} (-1)^{\sum_{i=1}^m \text{sign}(p_i)} \prod_{i=1}^{m} B_{p_i} \binom{m - \sum_{j=1}^{i-1} p_j}{p_i} \\
&= -B_m - \sum_{p_1=1}^{m-1} a_{p_1} \binom{m}{p_1} \sum_{(\hat{p}_1,\ldots,\hat{p}_{m-p_1})\in P_{m-p_1}} (-1)^{\sum_{i=1}^{m-p_1} \text{sign}(p_i)} \prod_{j=1}^{m-p_1} B_{\hat{p}_j} \binom{m - i_1 - \sum_{s=1}^{j-1} \hat{p}_s}{\hat{p}_j} \\
&= -B_m - \sum_{p_1=1}^{m-1} a_{p_1} \binom{m}{p_1} \beta_{0,m-p_1} \\
&= -B_m - \sum_{p_1=1}^{m-1} a_{p_1} \binom{m}{p_1} \frac{1}{m-p_1+1} \\
&= -\sum_{k=1}^{m} \frac{B_k}{m-k+1} \binom{m}{k} \\
&= \frac{1}{m+1}
\end{aligned}
\tag{40}
$$

where the last equality is due to Lemma 9. Now, this implies that

$$
\Delta_S(x) = \Phi_S^{|S|}(x) = f_S(x) + \sum_{K\subset[d]\setminus S,\ |K|\geq 1} D_{0,|K|}\, f_{S\cup K}(x) = \sum_{K\subset[d]\setminus S} \frac{1}{1+|K|} f_{S\cup K}(x) \tag{41}
$$

which is a version of Theorem 1 in Grabisch (1997). Using (41) and the explicit formula for n-Shapley Values (20), we get

$$
\begin{aligned}
\Phi_S^n(x) &= \sum_{k=0}^{n-|S|} \sum_{K\subset[d]\setminus S,\,|K|=k} B_k\, \Delta_{S\cup K}(x) \\
&= \sum_{k=0}^{n-|S|} \sum_{K\subset[d]\setminus S,\,|K|=k} B_k \sum_{T\subset[d]\setminus(S\cup K)} \frac{1}{1+|T|} f_{S\cup K\cup T}(x)
\end{aligned}
\tag{42}
$$

From which we see that the component function $f_{S\cup\tilde{K}}$ is being added to $\Phi_S^n(x)$ exactly

$$
C_{n-|S|,|\tilde{K}|} = \sum_{k=0}^{n-|S|} \binom{n-|S|}{k} \frac{B_k}{1+|\tilde{K}|-k} \tag{43}
$$

times which concludes the proof. □

G Proof of Theorem 7

Proof of Theorem 7. According to Theorem 4, the Shapley-GAM decomposition is given by

$$
f_S(x) = \sum_{L\subset S} (-1)^{|S|-|L|} v(x_L, L). \tag{44}
$$

By substituting the definition of the value function (12)

$$\begin{aligned} f_S(x) &= \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L) \\ &= \sum_{L \subset S} (-1)^{|S|-|L|} \sum_{T \subset L} g_T(x) \\ &= \sum_{L \subset S} \sum_{T \subset L} g_T(x) (-1)^{|S|-|L|} \\ &= \sum_{T \subset S} g_T(x) \sum_{L \subset S \setminus T} (-1)^{|S|-|L|-|T|} \\ &= g_S(x) \end{aligned} \tag{45}$$

Where we have re-arranged the sum to count the number of occurrences of the set T, and then used the fact that inner sum averages to zero except for $T = S$. □

H Proof of Theorem 8

We show a slightly more general result than what is stated in the main paper. In fact, we show that recovery holds for all interaction indices that can be written as

$$I_S^n(x) = f_S(x) + \sum_{\substack{K \subset [d] \setminus S \\ n+1 \leq |S|+|K|}} C_{n,|S|,|K|} \, f_{S \cup K}(x) \qquad \forall S \subseteq [d], |S| \leq n \tag{46}$$

where $f_S(x)$ are the component functions of the Shapley-GAM and $C_{n,|S|,|K|} \in \mathbb{R}$ are coefficients that depend on the interaction index. n-Shapley Values can be written like this according to Theorem 6. For the Faith-Shap interaction index, this representation is given in Theorem 19 in Tsai et al. (2022)

$$\text{Faith-Shap}_S^n(x) = f_S(x) + \sum_{\substack{K \subset [d] \setminus S \\ n+1 \leq |S|+|K|}} (-1)^{n-|S|} \frac{|S|}{n+|S|} \frac{\binom{n}{|S|}\binom{|S|+|K|-1}{n}}{\binom{|S|+|K|+n-1}{n+|S|}} \, f_{S \cup K}(x) \qquad \forall |S| \leq n. \tag{47}$$

Also the Shapley Taylor interaction index (Sundararajan et al., 2020) can, due to its symmetry, be written as

$$\text{Shapley-Taylor}_S^n(x) = \begin{cases} f_S(x) & \text{if } |S| < n \\ f_S(x) + \sum_{\substack{K \subset [d] \setminus S \\ n+1 \leq |S|+|K|}} \frac{1}{\binom{|S|+|K|}{|K|}} \, f_{S \cup K}(x) & \text{if } |S| = n. \end{cases} \tag{48}$$

Proof of Theorem 8. We assume that the function f can be written as a GAM of order n, that is

$$f(x) = \sum_{S \subset [d],\, |S| \leq n} g_S(x_S). \tag{49}$$

Notice that this GAM is not necessarily the Shapley-GAM, but just some way to write the function f as a GAM. Let f_S be the component functions of the Shapley-GAM. Now, n-Shapley Values, the Faith-Shap interaction index, as well as the Shapley Taylor interaction index, can be written as a linear combination of the component functions of the Shapley-GAM

$$I_S^n(x) = f_S(x_S) + \sum_{K \subset [d] \setminus S,\ |S|+|K| > n} C_{n-|S|,|K|} \, f_{S \cup K}(x_{S \cup K}) \tag{50}$$

where the specific linear coefficients $C_{n,m}$ depend on the interaction index (Theorem 6, equation (47), equation (48)). According to equation (50), the interaction index equals $f_S(x_S)$ plus some weighted components of the Shapley-GAM of order greater than n. As a consequence, it remains to show is that the Shapley-GAM is a GAM of order n (then the second sum vanishes and we arrive at $I_S^n(x) = f_S(x_S)$ which is what we want to show).

It remains to show that the Shapley-GAM is a GAM of order n. According to Theorem 4, the component functions of the Shapley-GAM are given by

$$f_S(x) = \sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L). \tag{51}$$

We want to show that the component functions of degree greater than n vanish. Let us first consider observational SHAP. Here we have

$$\begin{aligned}
\sum_{L \subset S} (-1)^{|S|-|L|} v(x_L, L) &= \sum_{L \subset S} (-1)^{|S|-|L|} \mathbb{E}[f(x)|x_L] \\
&= \sum_{L \subset S} (-1)^{|S|-|L|} \mathbb{E}\left[\sum_{T \subset [d],\, |T| \leq n} g_T(x_T) \Big| x_L \right] \\
&= \sum_{L \subset S} (-1)^{|S|-|L|} \sum_{T \subset [d],\, |T| \leq n} \mathbb{E}\left[g_T(x_T) | x_L \right] \\
&= \sum_{T \subset [d],\, |T| \leq n} \sum_{L \subset S} (-1)^{|S|-|L|} \mathbb{E}\left[g_T(x_T) | x_L \right]
\end{aligned} \tag{52}$$

Consider the inner sum. If $|S| > n$, we can always pick an element $i \in S \setminus T$ and write

$$\sum_{L \subset S \setminus \{i\}} (-1)^{|S|-|L|} \Big(\mathbb{E}\left[g_T(x_T) | x_L \right] - \mathbb{E}\left[g_T(x_T) | x_{L \cup \{i\}} \right] \Big) \tag{53}$$

If the input features are independent, then $g_T(x_T)$ and x_i are independent, from which we get by the properties of the conditional expectation that

$$\mathbb{E}\left[g_T(x_T) | x_{L \cup \{i\}} \right] = \mathbb{E}\left[g_T(x_T) | x_L \right] \tag{54}$$

It follows that the inner sum is zero for all sets T, and that the component functions of the Shapley-GAM of degree greater than n are equal to zero, too.

Let us now consider interventional SHAP. Just as for observational SHAP, we arrive at equation (53) using the linearity of the expectation operator. Hence, we require that

$$\mathbb{E}\left[g_T(x_T) | do(x_{L \cup \{i\}}) \right] = \mathbb{E}\left[g_T(x_T) | do(x_L) \right] \tag{55}$$

which follows from the properties of the causal do-operator. Intuitively, since g_T does not depend on the value of feature i, intervening on that feature has no effect. □

I Proof of Lemma 10

Proof. Let us first consider the case $n = 0$. For $n = 0$ and $m = 0$, we have

$$\binom{0}{0} \binom{0}{0} \frac{(0-0)!(0-0)!}{(0+0-0-0+1)!} (-1)^0 B_0 = 1. \tag{56}$$

For $n = 0$ and $m \geq 1$, we have

$$\begin{aligned}
\sum_{l=0}^{m} \binom{m}{l} \frac{1}{(m-l+1)} (-1)^l B_l &= \frac{1}{m+1} \sum_{l=0}^{m} \binom{m+1}{l} (-1)^l B_l \\
&= \frac{-2}{m+1} \binom{m+1}{1} B_1 + \sum_{l=0}^{m} \binom{m+1}{l} \\
&= -2B_1 + 0 = 1.
\end{aligned} \tag{57}$$

where we used (15) and the fact that the odd Bernoulli numbers vanish except for $n = 1$. For $m = 0$ and $n \geq 1$, we also have from (15)

$$\sum_{k=0}^{n} \binom{n}{k} \frac{1}{(n-k+1)} (-1)^0 B_k = \frac{1}{n+1} \sum_{k=0}^{n} \binom{n+1}{k} B_k = 0. \tag{58}$$

It remains to show the general case $n, m \geq 1$. According to a derivation by Gy (2022), the problem in this case is equivalent to

$$(-1)^n \sum_{l=0}^{m} \frac{B_{n+l+1}}{n+l+1} \binom{m}{l} + (-1)^m \sum_{k=0}^{n} \frac{B_{m+k+1}}{m+k+1} \binom{n}{k} = -\frac{1}{(n+m+1)\binom{n+m}{m}} \tag{59}$$

Now, Theorem 2 in Gould and Quaintance (2014) with $s = 1$ states that for any sequence of numbers $(a_n)_{n\geq 0}$, it holds that

$$\sum_{k=0}^{m} \binom{m}{k} \frac{a_{n+k+1}}{n+k+1} = \sum_{k=0}^{n} (-1)^{n-k} \binom{n}{k} \frac{b_{m+k+1}}{m+k+1} + \frac{(-1)^{n+1} a_0}{(m+n+1)\binom{m+n}{n}} \tag{60}$$

where the sequence $(b_n)_{n\geq 0}$ is the binomial transform of the sequence $(a_n)_{n\geq 0}$, given by

$$b_n = \sum_{k=0}^{n} \binom{n}{k} a_k. \tag{61}$$

Setting $a_n = B_n$, we have from (15) that the binomial transform of the Bernoulli numbers is simply

$$b_n = \sum_{k=0}^{n} \binom{n}{k} B_k = (-1)^n B_n \tag{62}$$

where the factor $(-1)^n$ takes care of the special case $n = 1$. Using (60) with $a_n = B_n$ and $b_n = (-1)^n B_n$, we get

$$(-1)^n \sum_{k=0}^{m} \binom{m}{k} \frac{B_{n+k+1}}{n+k+1} = -\sum_{k=0}^{n} (-1)^m \binom{n}{k} \frac{B_{m+k+1}}{m+k+1} - \frac{1}{(m+n+1)\binom{m+n}{n}} \tag{63}$$

where we multiplied both sides with $(-1)^n$. This is the same as (59) which concludes the proof. □

J Datasets and Models

In our experiments, we use the following data sets and models.

J.1 Datasets

Folktables Income. Folktables is a Python package that provides access to data sets derived from recent US Censuses `https://github.com/zykls/folktables`. We used this package to obtain the data from the 2016 Census in California. The machine learning problem is the ACSIncome prediction task, that is to predict whether an individual's income is above $50,000, based on 10 personal characteristics (Ding et al., 2021). The data set contains of 152 149 observations.

Folktables Travel Time. Folktables is a Python package that provides access to data sets derived from recent US Censuses `https://github.com/zykls/folktables`. We used this package to obtain the data from the 2016 Census in California. The machine learning problem is the ACSTravelTime prediction task, that is to predict whether an individual has to commute to work longer than 20 minutes, based on 10 personal characteristics (Ding et al., 2021). The data set contains 133 549 observations.

German Credit. The German Credit Data set is a data set with 20 different features on individual's credit history and personal characteristic. The machine learning problem is to predict credit risk in binary form. We obtained the data set from the UCI machine learning repository and reduced the number of features to 10 without any observed drop in accuracy. The data set contains 1000 observations.

California Housing. The California Housing data set was derived from the 1990 U.S. census. The regression problem is to predict the median house value, based on 8 characteristics. We obtained the data set form the `scikit-learn` library. The data set contains 20 640 observations.

Iris. The Iris data set is a simple flower data set. The machine learning problem is to classify whether the flower is of a particular kind or not, based on 4 different features. We obtained the data set form the `scikit-learn` library. The data set contains 150 observations.

J.2 Models

Glassbox-GAM. We train the Glassbox-GAMs with the `interpretML` library (Nori et al., 2019) and default parameters (no interactions).

Gradient Boosted Tree. We use the `xgboost` library (Chen and Guestrin, 2016) and train with 100 trees per model. This setting allows to achieve competitive accuracy for gradient boosted trees.

Random Forest. We use the `scikit-learn` library (Pedregosa et al., 2011) and train with 100 trees per forest. This setting allows to achieve competitive accuracy for random forests.

k-Nearest Neighbor. We use the `scikit-learn` library (Pedregosa et al., 2011). The hyperparameter k was chosen with cross-validation to be 30, 80, 25, 10, 1 for the data sets as listed above.

K Additional Plots and Figures

K.1 Folktables Income

K.1.1 Glassbox-GAM

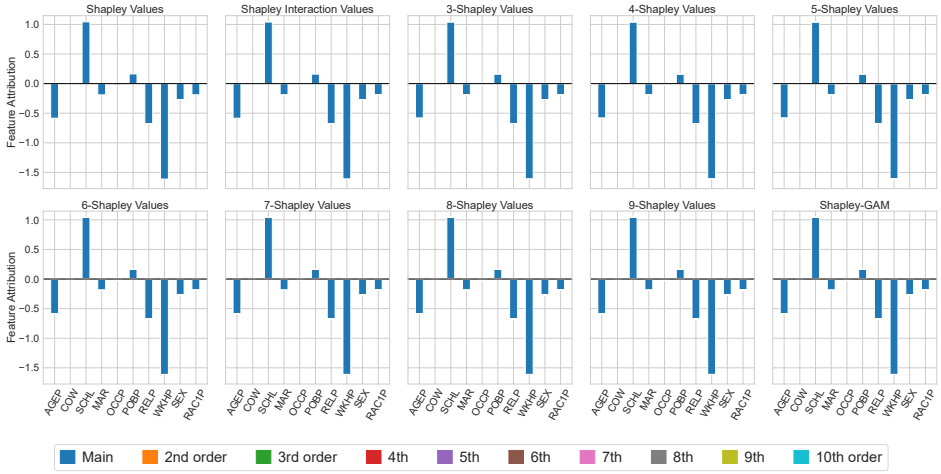


Figure K.4: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the Folktables Income data set.

K.1.2 Gradient Boosted Tree

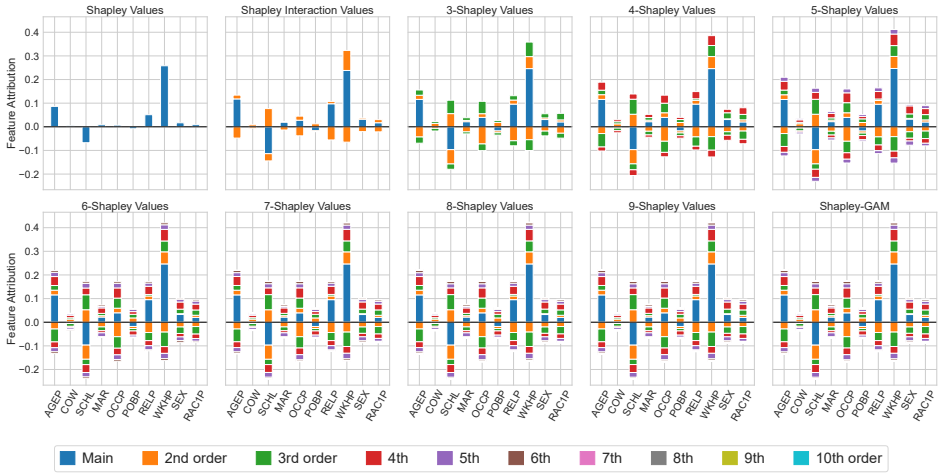


Figure K.5: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the Folktables Income data set.

K.1.3 Random Forest

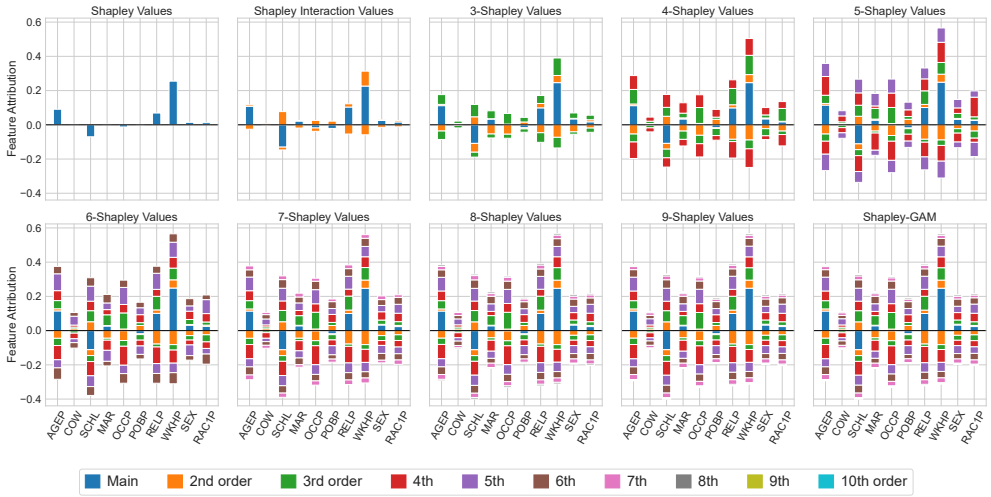


Figure K.6: n-Shapley Values for a Random Forest and the first observation in our test set of the Folktables Income data set.

K.1.4 k-Nearest Neighbor

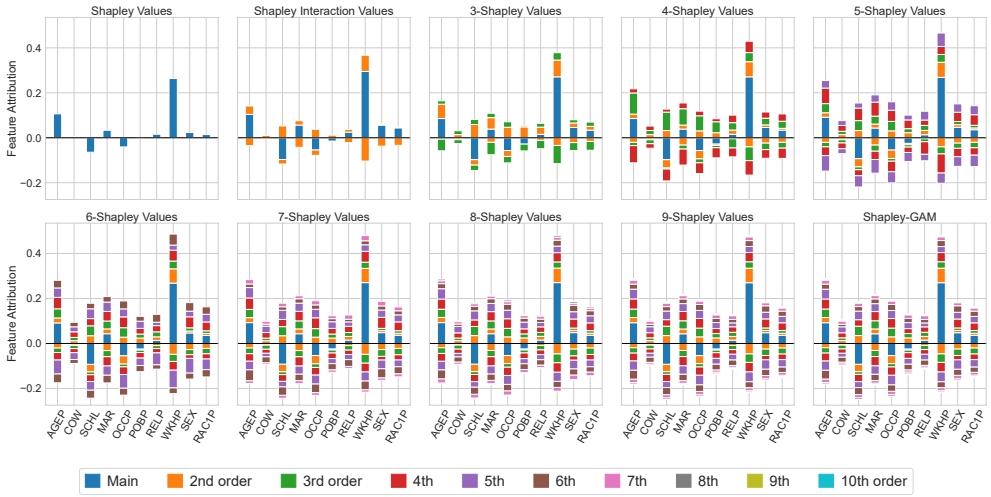


Figure K.7: n-Shapley Values for a kNN classifier and the first observation in our test set of the Folktables Income data set.

Figure K.8: Partial dependence plots for the kNN classifier on the Folktables Income data set (compare Figure 2 in the main paper). Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.

K.2 Folktables Travel

K.2.1 Glassbox-GAM

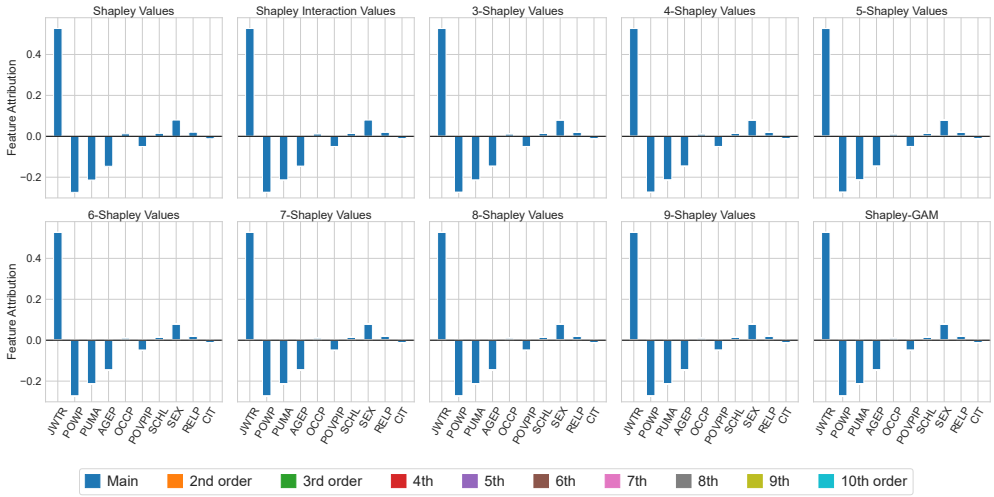


Figure K.9: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the Folktables Travel data set.

K.2.2 Gradient Boosted Tree

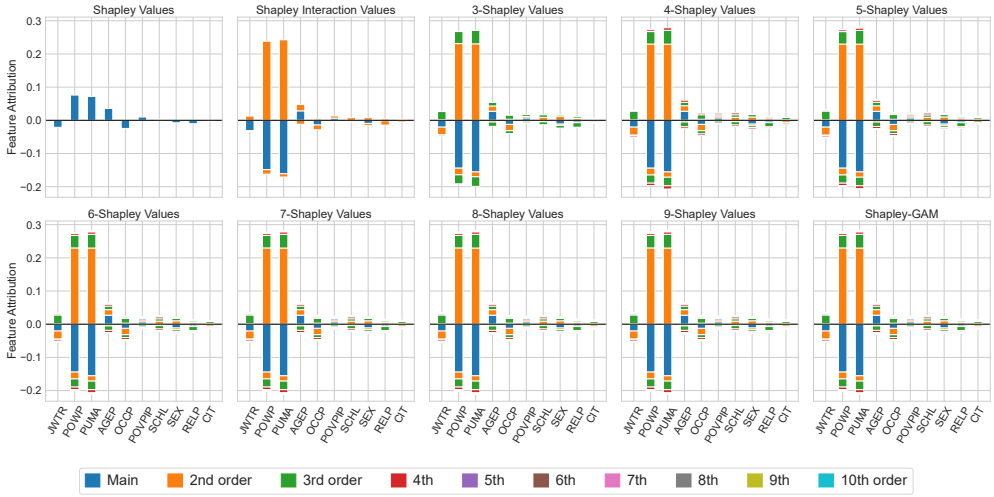


Figure K.10: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the Folktables Travel data set.

K.2.3 Random Forest

Figure K.11: n-Shapley Values for a Random Forest and the first observation in our test set of the Folktables Travel data set.

K.2.4 k-Nearest Neighbor

Figure K.12: n-Shapley Values for a kNN classifier and the first observation in our test set of the Folktables Travel data set.

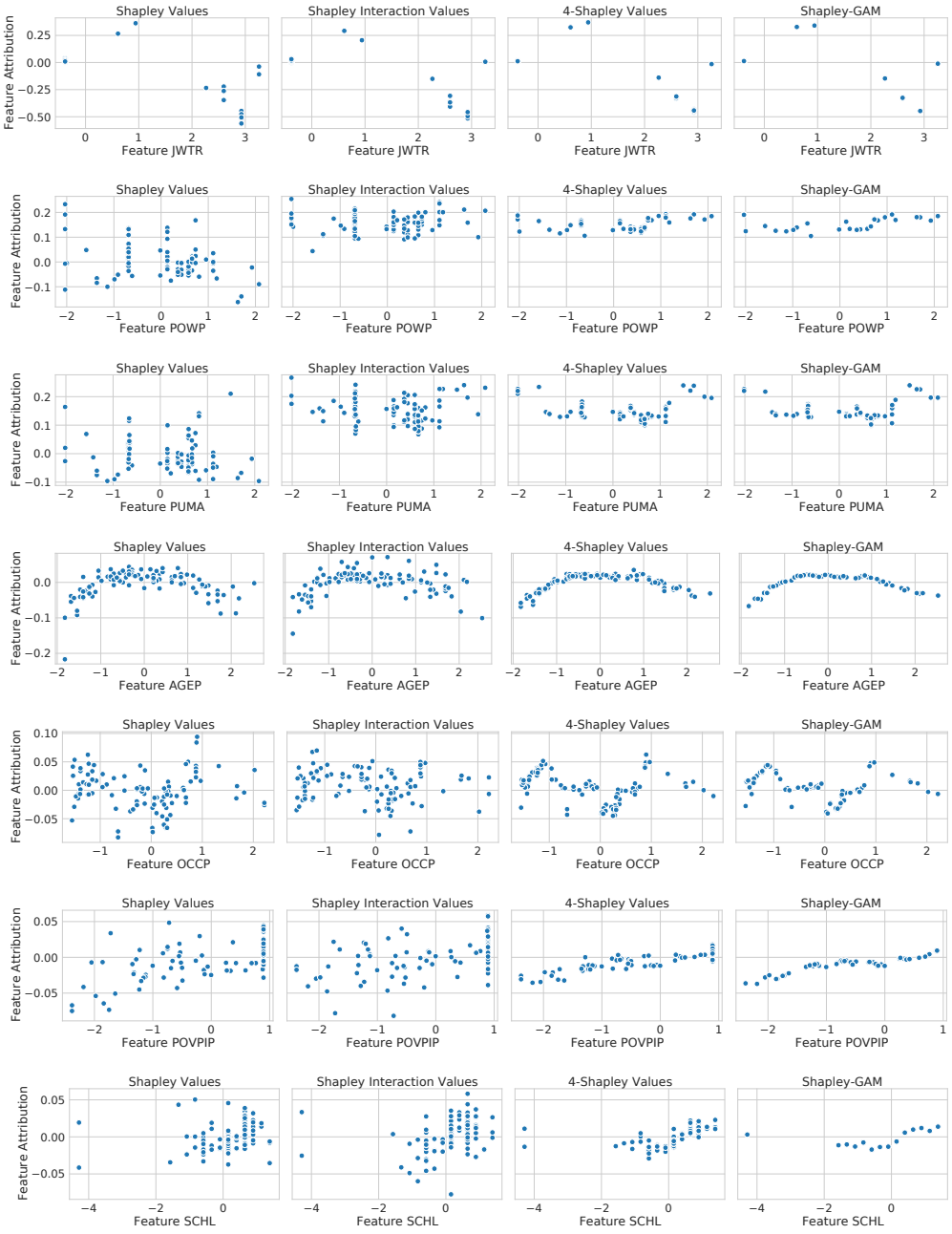


Figure K.13: Partial dependence plots for the random forest on the Folktables Travel data set. Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.

K.3 German Credit

K.3.1 Glassbox-GAM

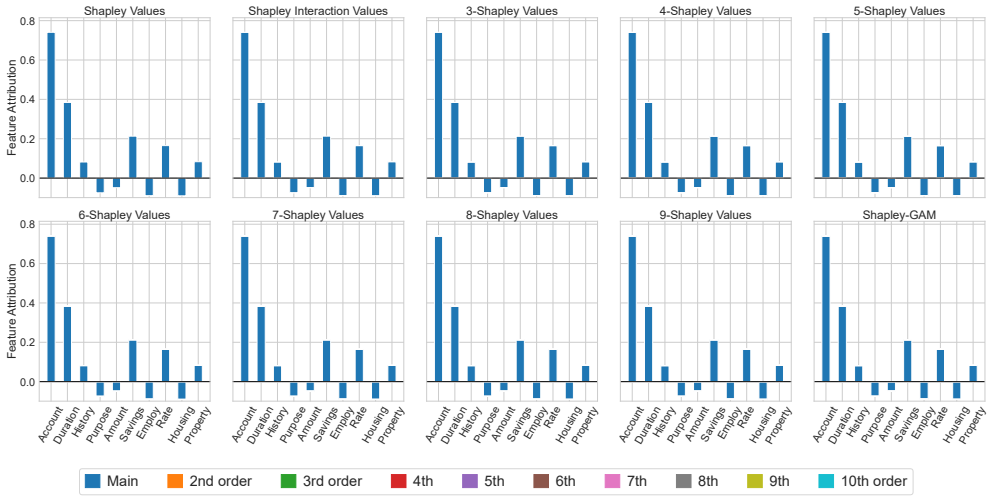


Figure K.14: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the German Credit data set.

K.3.2 Gradient Boosted Tree

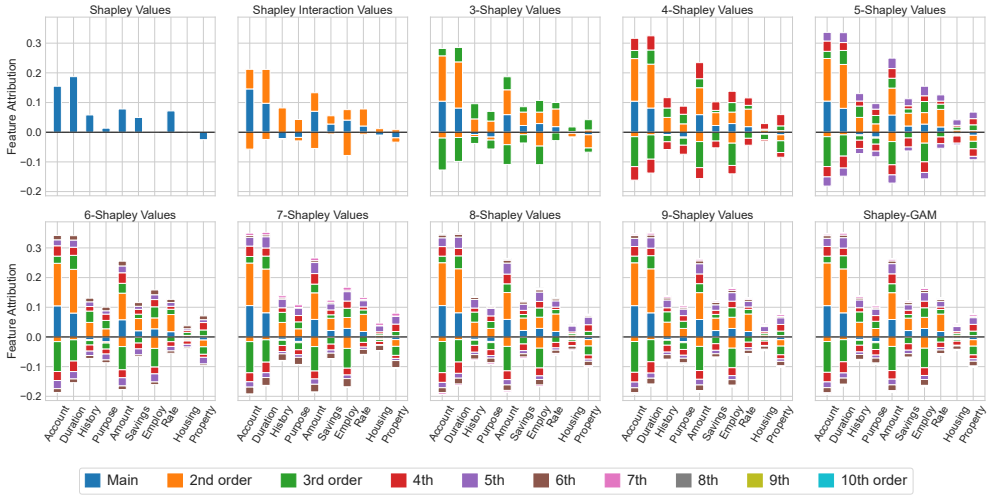


Figure K.15: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the German Credit data set.

K.3.3 Random Forest

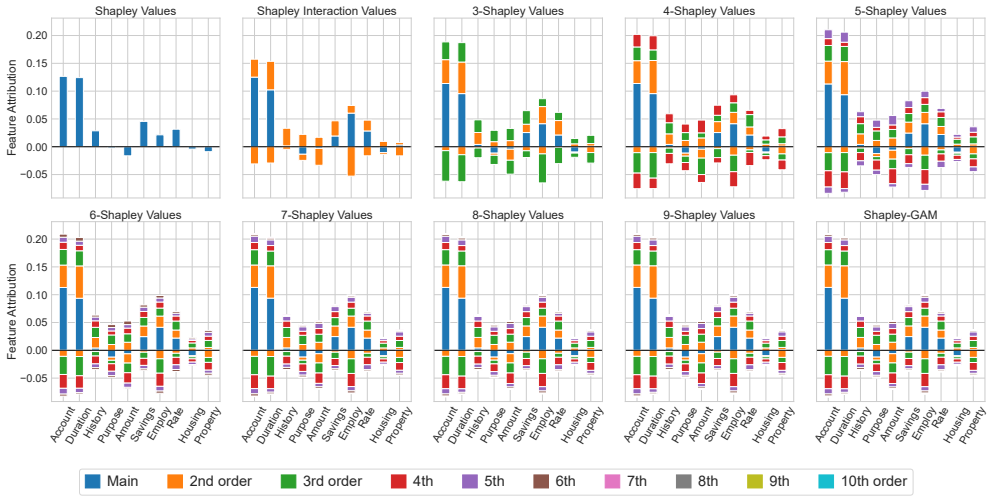


Figure K.16: n-Shapley Values for a Random Forest and the first observation in our test set of the German Credit data set.

K.3.4 k-Nearest Neighbor

Figure K.17: n-Shapley Values for a kNN classifier and the first observation in our test set of the German Credit data set.

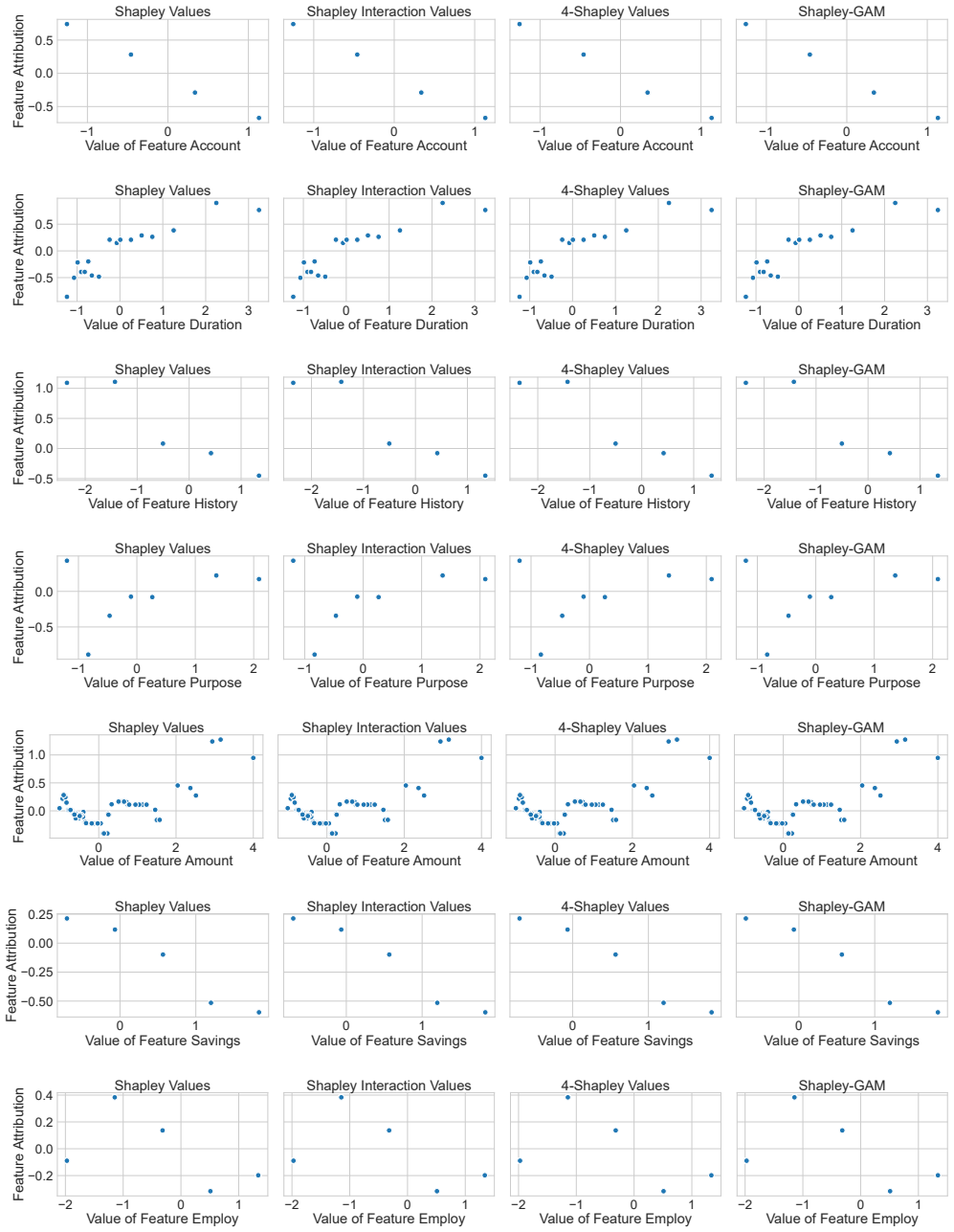


Figure K.18: Partial dependence plots for the Glassbox-GAM without interaction terms on the German Credit data set. Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.

K.4 California Housing

K.4.1 Glassbox-GAM

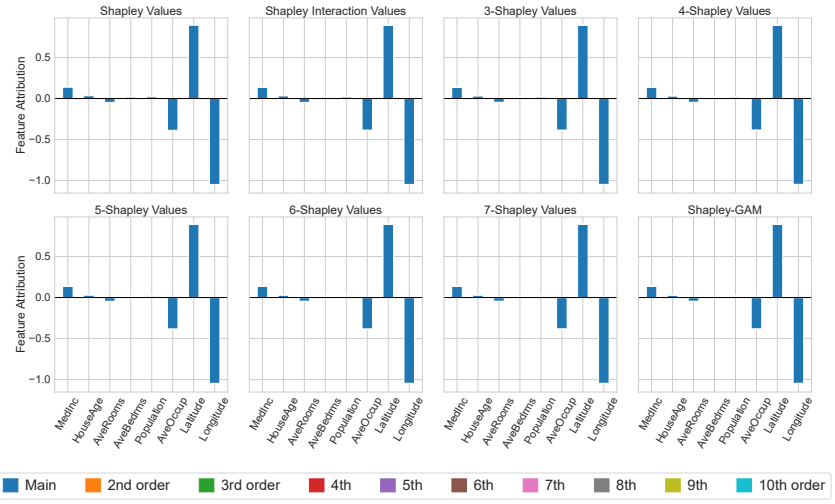


Figure K.19: n-Shapley Values for a Glassbox-GAM and the first observation in our test set of the California Housing data set.

K.4.2 Gradient Boosted Tree

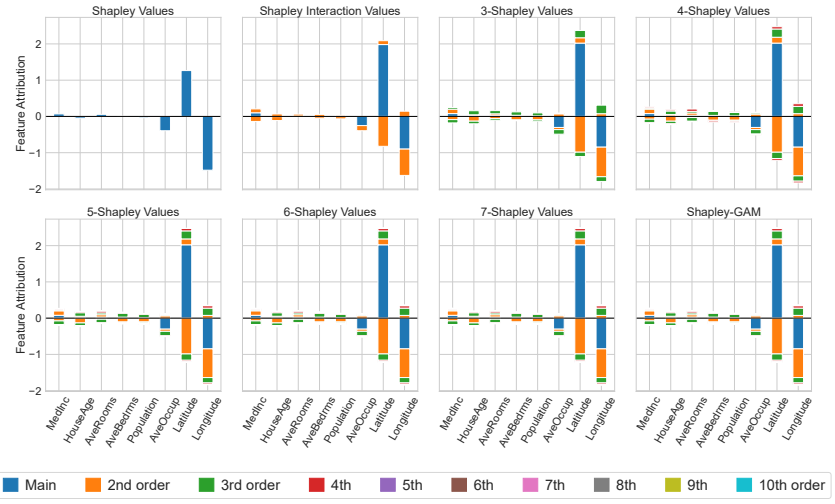


Figure K.20: n-Shapley Values for a Gradient Boosted Tree and the first observation in our test set of the California Housing data set.

K.4.3 Random Forest

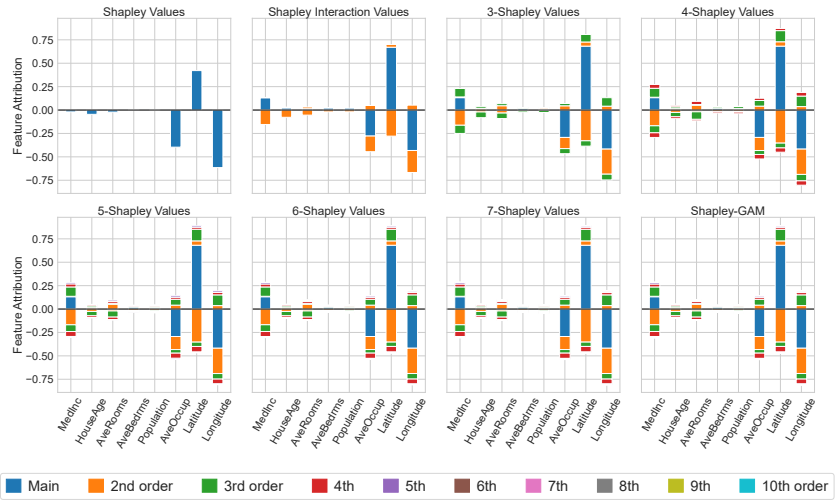


Figure K.21: n-Shapley Values for a Random Forest and the first observation in our test set of the California Housing data set.

K.4.4 k-Nearest Neighbor

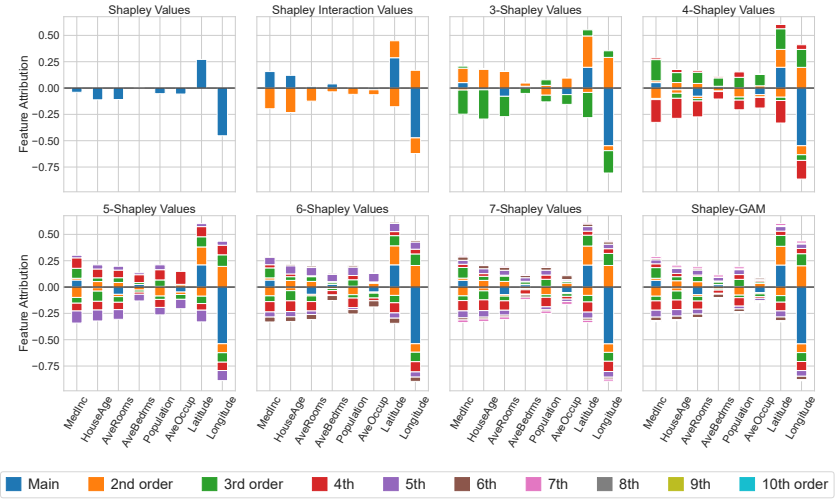


Figure K.22: n-Shapley Values for a kNN classifier and the first observation in our test set of the California Housing data set.

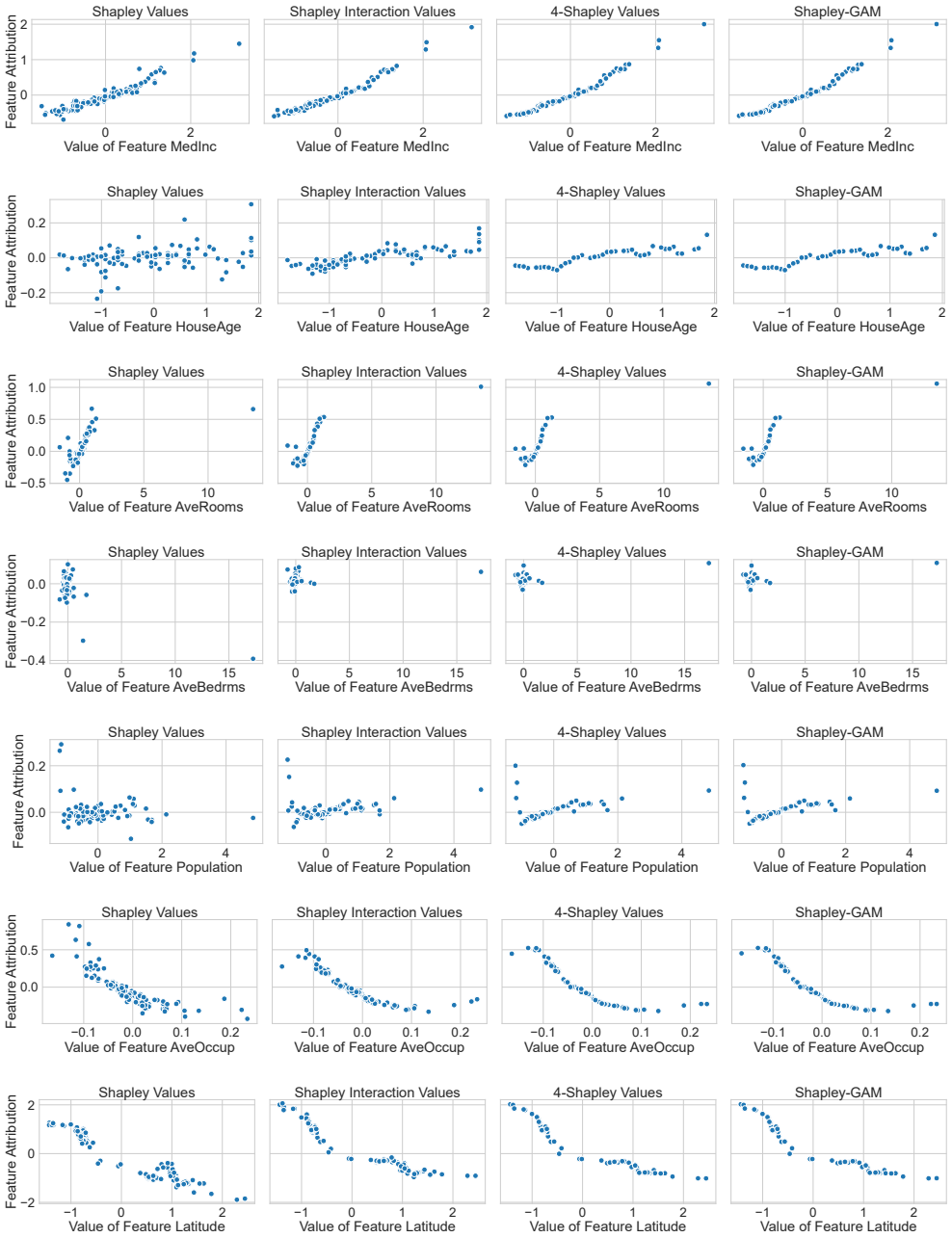


Figure K.23: Partial dependence plots for the gradient boosted tree on the California Housing data set. Depicted are the partial dependence plots of Φ_i^n for $n = \{1, 2, 4, 10\}$ and 7 different features.

Subset S	N=500	N=5000	N=133 549
(0,)	0.1128	0.1144	0.1165
(1,)	0.0005	-0.0006	0.0022
(2,)	-0.1248	-0.1117	-0.1098
(3,)	0.0227	0.0283	0.0281
(4,)	-0.0041	-0.0020	-0.0018
(5,)	-0.0123	-0.0189	-0.0200
(6,)	0.0845	0.1003	0.0982
(7,)	0.2357	0.2478	0.2505
(8,)	0.0280	0.0329	0.0347
(9,)	0.0197	0.0238	0.0241
(0, 1)	-0.0023	-0.0020	-0.0032
(0, 2)	0.0059	0.0005	-0.0025
(0, 3)	-0.0146	-0.0128	-0.0126
(0, 4)	0.0089	0.0102	0.0102
(0, 5)	0.0038	0.0140	0.0141
(0, 6)	-0.0244	-0.0242	-0.0213
(0, 7)	-0.0452	-0.0416	-0.0426
(0, 8)	-0.0032	-0.0028	-0.0038
(0, 9)	-0.0043	-0.0055	-0.0071
(1, 2)	-0.0012	-0.0009	-0.0022
(1, 3)	-0.0004	0.0003	0.0007
(1, 4)	-0.0006	0.0011	-0.0011
(1, 5)	-0.0060	-0.0000	0.0004
(1, 6)	0.0027	0.0024	0.0014
(1, 7)	0.0093	0.0102	0.0079
(1, 8)	-0.0017	0.0020	0.0007
(1, 9)	0.0048	0.0032	0.0027
(2, 3)	0.0029	-0.0006	-0.0016
(2, 4)	-0.0419	-0.0534	-0.0547
(2, 5)	-0.0128	-0.0095	-0.0115
(2, 6)	0.0389	0.0286	0.0290
(2, 7)	0.0752	0.0695	0.0677
(2, 8)	-0.0031	-0.0044	-0.0070
(2, 9)	0.0151	0.0039	0.0031
(3, 4)	-0.0112	-0.0093	-0.0091
(3, 5)	0.0006	0.0058	0.0055
(3, 6)	-0.0068	-0.0116	-0.0099
(3, 7)	-0.0286	-0.0298	-0.0304
(3, 8)	-0.0135	-0.0165	-0.0181
(3, 9)	0.0038	-0.0036	-0.0041
(4, 5)	-0.0016	0.0069	0.0071
(4, 6)	-0.0279	-0.0295	-0.0298
(4, 7)	-0.0100	-0.0070	-0.0079
(4, 8)	-0.0019	-0.0037	-0.0043
(4, 9)	-0.0091	-0.0116	-0.0122
(5, 6)	0.0026	0.0083	0.0079
(5, 7)	0.0084	0.0152	0.0157
(5, 8)	-0.0000	0.0055	0.0045
(5, 9)	0.0015	0.0044	0.0041
(6, 7)	-0.0551	-0.0603	-0.0581
(6, 8)	-0.0132	-0.0174	-0.0182
(6, 9)	-0.0053	-0.0140	-0.0126
(7, 8)	-0.0125	-0.0102	-0.0127
(7, 9)	-0.0102	-0.0151	-0.0161
(8, 9)	0.0052	0.0014	-0.0004
(0, 1, 2)	-0.0058	-0.0026	-0.0014
(0, 1, 3)	0.0018	0.0028	0.0020
(0, 1, 4)	-0.0000	-0.0030	-0.0021
(0, 1, 5)	0.0070	-0.0005	-0.0013
(0, 1, 6)	0.0060	0.0024	0.0030
(0, 1, 7)	-0.0039	-0.0024	-0.0015
(0, 1, 8)	0.0073	0.0007	0.0014
(0, 1, 9)	-0.0003	-0.0006	-0.0009
(0, 2, 3)	0.0038	0.0031	0.0030
(0, 2, 4)	-0.0274	-0.0141	-0.0079
(0, 2, 5)	0.0088	0.0062	0.0081
(0, 2, 6)	-0.0042	0.0006	-0.0006
(0, 2, 7)	0.0233	0.0242	0.0275
(0, 2, 8)	0.0043	0.0023	0.0055
(0, 2, 9)	-0.0298	-0.0249	-0.0216
(0, 3, 4)	0.0149	0.0078	0.0091
(0, 3, 5)	0.0019	-0.0023	-0.0014
...	...	...	...

Subset S	N=500	N=5000	N=133 549
...	...	...	...
(2, 7, 8, 9)	0.0043	0.0009	0.0005
(3, 4, 5, 6)	-0.0101	-0.0143	-0.0135
(3, 4, 5, 7)	0.0045	-0.0030	-0.0049
(3, 4, 5, 8)	-0.0020	-0.0053	-0.0047
(3, 4, 5, 9)	-0.0064	-0.0049	-0.0054
(3, 4, 6, 7)	0.0097	0.0076	0.0079
(3, 4, 6, 8)	-0.0058	-0.0058	-0.0047
(3, 4, 6, 9)	0.0032	0.0018	0.0017
(3, 4, 7, 8)	0.0007	-0.0011	-0.0011
(3, 4, 7, 9)	0.0041	0.0003	0.0004
(3, 4, 8, 9)	0.0006	0.0013	0.0021
(3, 5, 6, 7)	0.0052	0.0059	0.0071
(3, 5, 6, 8)	-0.0024	-0.0011	-0.0000
(3, 5, 6, 9)	-0.0044	-0.0023	-0.0019
(3, 5, 7, 8)	-0.0023	-0.0014	-0.0011
(3, 5, 7, 9)	-0.0007	-0.0031	-0.0024
(3, 5, 8, 9)	-0.0010	-0.0007	-0.0005
(3, 6, 7, 8)	0.0035	0.0027	0.0034
(3, 6, 7, 9)	-0.0034	-0.0052	-0.0045
(3, 6, 8, 9)	-0.0019	-0.0011	-0.0004
(3, 7, 8, 9)	0.0018	0.0014	0.0003
(4, 5, 6, 7)	-0.0052	-0.0020	-0.0037
(4, 5, 6, 8)	-0.0025	-0.0001	0.0007
(4, 5, 6, 9)	-0.0019	0.0005	0.0004
(4, 5, 7, 8)	-0.0027	0.0009	0.0016
(4, 5, 7, 9)	-0.0017	0.0005	0.0010
(4, 5, 8, 9)	-0.0004	-0.0004	0.0000
(4, 6, 7, 8)	-0.0000	-0.0003	0.0011
(4, 6, 7, 9)	0.0017	0.0005	0.0006
(4, 6, 8, 9)	-0.0005	-0.0005	0.0005
(4, 7, 8, 9)	0.0007	-0.0000	-0.0001
(5, 6, 7, 8)	-0.0041	-0.0024	-0.0012
(5, 6, 7, 9)	-0.0038	-0.0046	-0.0039
(5, 6, 8, 9)	0.0013	-0.0009	-0.0007
(5, 7, 8, 9)	-0.0003	0.0003	0.0004
(6, 7, 8, 9)	0.0022	0.0005	0.0003
(0, 1, 2, 3, 4)	0.0042	0.0010	0.0003
(0, 1, 2, 3, 5)	0.0004	0.0010	0.0012
(0, 1, 2, 3, 6)	0.0018	0.0004	0.0002
(0, 1, 2, 3, 7)	0.0014	0.0006	0.0012
(0, 1, 2, 3, 8)	0.0007	-0.0004	-0.0001
(0, 1, 2, 3, 9)	0.0006	0.0012	0.0015
(0, 1, 2, 4, 5)	0.0051	0.0013	0.0013
(0, 1, 2, 4, 6)	0.0016	0.0011	0.0010
(0, 1, 2, 4, 7)	0.0005	-0.0011	-0.0009
(0, 1, 2, 4, 8)	0.0022	0.0005	0.0008
(0, 1, 2, 4, 9)	0.0026	-0.0002	-0.0000
(0, 1, 2, 5, 6)	0.0025	0.0023	0.0038
(0, 1, 2, 5, 7)	0.0013	-0.0001	0.0003
(0, 1, 2, 5, 8)	0.0012	0.0011	0.0017
(0, 1, 2, 5, 9)	0.0017	-0.0008	-0.0006
(0, 1, 2, 6, 7)	0.0005	-0.0010	-0.0008
(0, 1, 2, 6, 8)	-0.0003	-0.0008	-0.0001
(0, 1, 2, 6, 9)	0.0008	-0.0001	-0.0003
(0, 1, 2, 7, 8)	0.0003	-0.0010	-0.0005
(0, 1, 2, 7, 9)	0.0004	-0.0008	-0.0006
(0, 1, 2, 8, 9)	-0.0004	-0.0006	-0.0003
(0, 1, 3, 4, 5)	0.0013	0.0002	-0.0007
(0, 1, 3, 4, 6)	0.0017	0.0009	-0.0001
(0, 1, 3, 4, 7)	0.0028	0.0010	0.0010
(0, 1, 3, 4, 8)	0.0032	-0.0004	-0.0000
(0, 1, 3, 4, 9)	-0.0005	-0.0006	-0.0001
(0, 1, 3, 5, 6)	-0.0012	0.0000	0.0003
(0, 1, 3, 5, 7)	0.0018	-0.0006	-0.0003
(0, 1, 3, 5, 8)	0.0003	-0.0001	0.0000
(0, 1, 3, 5, 9)	-0.0012	-0.0000	0.0002
(0, 1, 3, 6, 7)	0.0011	0.0002	0.0009
(0, 1, 3, 6, 8)	0.0020	-0.0004	-0.0002
(0, 1, 3, 6, 9)	-0.0000	0.0007	0.0004
(0, 1, 3, 7, 8)	0.0015	-0.0003	0.0001
(0, 1, 3, 7, 9)	-0.0003	-0.0010	-0.0004
...	...	...	...

Subset S	N=500	N=5000	N=133 549
...	...	...	...
(1, 2, 5, 6, 7, 8, 9)	0.0029	-0.0003	-0.0009
(1, 3, 4, 5, 6, 7, 8)	0.0005	-0.0032	-0.0035
(1, 3, 4, 5, 6, 7, 9)	0.0061	0.0051	0.0049
(1, 3, 4, 5, 6, 8, 9)	0.0062	0.0014	-0.0009
(1, 3, 4, 5, 7, 8, 9)	0.0002	0.0002	0.0009
(1, 3, 4, 6, 7, 8, 9)	0.0015	0.0015	0.0008
(1, 3, 5, 6, 7, 8, 9)	0.0002	-0.0026	-0.0004
(1, 4, 5, 6, 7, 8, 9)	0.0025	0.0026	0.0016
(2, 3, 4, 5, 6, 7, 8)	-0.0038	0.0007	-0.0002
(2, 3, 4, 5, 6, 7, 9)	0.0039	0.0042	0.0036
(2, 3, 4, 5, 6, 8, 9)	0.0059	0.0022	0.0013
(2, 3, 4, 5, 7, 8, 9)	-0.0042	-0.0016	-0.0010
(2, 3, 4, 6, 7, 8, 9)	-0.0007	0.0013	0.0008
(2, 3, 5, 6, 7, 8, 9)	-0.0046	-0.0029	-0.0015
(2, 4, 5, 6, 7, 8, 9)	0.0012	0.0018	0.0008
(3, 4, 5, 6, 7, 8, 9)	0.0014	0.0009	0.0011
(0, 1, 2, 3, 4, 5, 6, 7)	-0.0021	-0.0027	-0.0019
(0, 1, 2, 3, 4, 5, 6, 8)	0.0037	0.0021	0.0014
(0, 1, 2, 3, 4, 5, 6, 9)	0.0018	-0.0006	-0.0015
(0, 1, 2, 3, 4, 5, 7, 8)	0.0002	0.0002	-0.0001
(0, 1, 2, 3, 4, 5, 7, 9)	-0.0002	-0.0006	-0.0012
(0, 1, 2, 3, 4, 5, 8, 9)	0.0015	0.0018	-0.0001
(0, 1, 2, 3, 4, 6, 7, 8)	0.0005	0.0010	-0.0003
(0, 1, 2, 3, 4, 6, 7, 9)	-0.0004	0.0013	0.0003
(0, 1, 2, 3, 4, 6, 8, 9)	0.0025	0.0014	0.0005
(0, 1, 2, 3, 4, 7, 8, 9)	-0.0013	0.0001	-0.0003
(0, 1, 2, 3, 5, 6, 7, 8)	0.0037	0.0016	-0.0005
(0, 1, 2, 3, 5, 6, 7, 9)	0.0009	0.0008	-0.0009
(0, 1, 2, 3, 5, 6, 8, 9)	0.0018	0.0009	-0.0002
(0, 1, 2, 3, 5, 7, 8, 9)	0.0014	0.0010	-0.0002
(0, 1, 2, 3, 6, 7, 8, 9)	0.0000	0.0006	0.0001
(0, 1, 2, 4, 5, 6, 7, 8)	0.0030	0.0017	0.0002
(0, 1, 2, 4, 5, 6, 7, 9)	-0.0009	-0.0002	0.0000
(0, 1, 2, 4, 5, 6, 8, 9)	0.0052	0.0014	0.0004
(0, 1, 2, 4, 5, 7, 8, 9)	-0.0010	0.0006	-0.0001
(0, 1, 2, 4, 6, 7, 8, 9)	-0.0013	0.0003	-0.0000
(0, 1, 2, 5, 6, 7, 8, 9)	0.0007	0.0003	-0.0004
(0, 1, 3, 4, 5, 6, 7, 8)	-0.0013	0.0008	-0.0006
(0, 1, 3, 4, 5, 6, 7, 9)	0.0003	0.0017	0.0006
(0, 1, 3, 4, 5, 6, 8, 9)	0.0010	0.0005	-0.0001
(0, 1, 3, 4, 5, 7, 8, 9)	-0.0006	0.0007	-0.0000
(0, 1, 3, 4, 6, 7, 8, 9)	-0.0007	0.0005	0.0002
(0, 1, 3, 5, 6, 7, 8, 9)	-0.0001	0.0008	0.0002
(0, 1, 4, 5, 6, 7, 8, 9)	-0.0002	0.0010	0.0001
(0, 2, 3, 4, 5, 6, 7, 8)	0.0006	0.0001	-0.0007
(0, 2, 3, 4, 5, 6, 7, 9)	-0.0005	0.0015	0.0003
(0, 2, 3, 4, 5, 6, 8, 9)	0.0012	0.0004	0.0001
(0, 2, 3, 4, 5, 7, 8, 9)	-0.0005	0.0002	-0.0001
(0, 2, 3, 4, 6, 7, 8, 9)	-0.0010	0.0002	0.0002
(0, 2, 3, 5, 6, 7, 8, 9)	0.0009	0.0004	-0.0000
(0, 2, 4, 5, 6, 7, 8, 9)	-0.0007	0.0007	0.0001
(0, 3, 4, 5, 6, 7, 8, 9)	-0.0010	0.0008	0.0006
(1, 2, 3, 4, 5, 6, 7, 8)	-0.0131	-0.0081	-0.0069
(1, 2, 3, 4, 5, 6, 7, 9)	-0.0018	0.0002	0.0013
(1, 2, 3, 4, 5, 6, 8, 9)	-0.0073	-0.0006	0.0015
(1, 2, 3, 4, 5, 7, 8, 9)	0.0039	0.0040	0.0042
(1, 2, 3, 4, 6, 7, 8, 9)	0.0011	0.0000	0.0014
(1, 2, 3, 5, 6, 7, 8, 9)	0.0018	0.0036	0.0014
(1, 2, 4, 5, 6, 7, 8, 9)	-0.0021	-0.0014	0.0005
(1, 3, 4, 5, 6, 7, 8, 9)	-0.0036	-0.0048	-0.0048
(2, 3, 4, 5, 6, 7, 8, 9)	-0.0021	-0.0039	-0.0038
(0, 1, 2, 3, 4, 5, 6, 7, 8)	-0.0023	-0.0018	-0.0002
(0, 1, 2, 3, 4, 5, 6, 7, 9)	-0.0008	-0.0013	0.0003
(0, 1, 2, 3, 4, 5, 6, 8, 9)	-0.0063	-0.0024	-0.0003
(0, 1, 2, 3, 4, 5, 7, 8, 9)	0.0012	0.0000	0.0010
(0, 1, 2, 3, 4, 6, 7, 8, 9)	0.0012	-0.0003	0.0002
(0, 1, 2, 3, 5, 6, 7, 8, 9)	-0.0017	-0.0008	0.0005
(0, 1, 2, 4, 5, 6, 7, 8, 9)	0.0003	-0.0004	0.0004
(0, 1, 3, 4, 5, 6, 7, 8, 9)	-0.0000	-0.0019	-0.0009
(0, 2, 3, 4, 5, 6, 7, 8, 9)	-0.0003	-0.0017	-0.0010
(1, 2, 3, 4, 5, 6, 7, 8, 9)	-0.0041	-0.0008	-0.0022
(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)	0.0005	0.0012	-0.0002

Table K.2: The individual terms of the Shapley-GAM decomposition of a kNN classifier on the Folktables Travel data set. The table depicts a number of selected terms of the full decomposition, estimated with 500, 5000 and 133549 samples per evaluation of the value function. The depicted terms are visualized in Figure C.3. From the table, we see that many relatively small higher-order coefficients are not very precisely estimated for $N = 5000$, whereas the overall sums (visualized in Figure C.3) are.

Chapter 3

Discussion

The first newspaper report of a machine that could *"walk, talk, see, write, reproduce itself and be conscious of its existence"* is now 65 years old (The New York Times, 1958). While the early days of computing in the 1940s and 1950s did indeed see important conceptual contributions, the desire to build machines that could solve real-world tasks like complex gameplay and image recognition remained elusive for the biggest part of the 20th century.

The last 15 years, however, have seen incredible advances in artificial intelligence, up to the point where seasoned scholars started to question whether research should continue at the current pace (Bengio, 2023). Long-standing open problems have been solved to a degree that seemed implausible just 10 years ago. While current computer programs and robots are arguably not conscious, they are indeed increasingly able to walk, talk, see, and write. In some respect, machine learning research today is therefore back to the debates that arose at its very beginning. What would be a valid test for general intelligence? What does it take to build truly intelligent machines? But also: How do today's systems work, exactly?

Apart from these scientific debates, the impact of machine learning on society is increasingly being seen as problematic. Notably, this is not an abstract concern but is evidenced by a large number of systems that have been shown to be demonstrably problematic (Barocas et al., 2019). At a high level, one of the main concerns about the usage of artificial intelligence systems in social contexts is that these systems are, by design, incredibly intransparent. This is problematic insofar as these systems are usually being deployed by institutions that are already relatively powerful, leading to concerns that artificial intelligence might adversely affect the balance of power in society (Acemoglu and Johnson, 2023).

In this current moment, the nascent field of explainable machine learning attempts to make model behavior transparent by providing human-understandable "explanations" for the behavior of complex machine learning systems. This approach has led to demonstrable success cases, especially in the area of model debugging and improvement. It has led, for example, to the discovery that models trained on medical images tend to rely on doctors' annotations instead of the underlying medical conditions.

What is more, there is also evidence that explainable machine learning can be useful for scientific discovery (Janizek et al., 2023).

With regard to the ambitious objectives of making complex systems generally transparent, or providing explanations that are useful for domain experts and regulators, explainable machine learning has so far been largely unsuccessful. At the time of the writing of this thesis, there exists little to no evidence that demonstrates the usefulness of post-hoc explanations in challenging application scenarios with human domain experts. A notable exception to this is the approach of building interpretable models.

In summary, it is now becoming increasingly clear that faithfully explaining complex machine learning systems is not an easy task. This applies both to models in computer vision, as well as to high-dimensional applications with tabular data. While it has not been easy to make progress on the problem of model interpretability, the issue of the opaqueness of currently available models remains nevertheless pressing. As such, there are still many questions worth exploring in explainable machine learning.

Chapter 4

Bibliography

D. Acemoglu and S. Johnson. *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*. Public Affairs, 2023.

J. Adebayo. *Towards Effective Tools for Debugging Machine Learning Models*. PhD thesis, Massachusetts Institute of Technology, 2022.

J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

J. Adebayo, M. Muelly, I. Liccardi, and B. Kim. Debugging tests for model explanations. In *Advances in Neural Information Processing Systems*, 2020.

J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. ProPublica. *See https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing*, 2016.

S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 2015.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019.

Y. Bengio. Slowing down development of AI systems passing the Turing test. *See https://yoshuabengio.org/2023/04/05/slowing-down-development-of-ai-systems-passing-the-turing-test/*, 2023.

L. Breiman. Random forests. *Machine Learning*, 2001a.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 2001b.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability, and Transparency*, 2018.

R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *IEEE Winter Conference on Applications of Computer Vision*, 2018.

C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin. This looks like that: deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, 2019.

B. Christian. *The alignment problem: Machine learning and human values*. WW Norton & Company, 2020.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, 2019.

D. Donoho. 50 Years of Data Science. *Journal of Computational and Graphical Statistics*, 2017.

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

G. K. Dziugaite, S. Ben-David, and D. M. Roy. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*, 2020.

D. Garreau and U. von Luxburg. Explaining the Explainer: A First Theoretical Analysis of LIME. In *International Conference on Artificial Intelligence and Statistics*, 2020.

S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, 2019.

J. D. Janizek, A. B. Dincer, S. Celik, H. Chen, W. Chen, K. Naxerova, and S.-I. Lee. Uncovering expression signatures of synergistic drug responses via ensembles of explainable machine-learning models. *Nature Biomedical Engineering*, 2023.

D. Janzing, L. Minorics, and P. Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, 2020.

C. Kang. OpenAI's Sam Altman Urges A.I. Regulation in Senate Hearing, The New York Times. *See https://www.nytimes.com/2023/05/16/technology/openai-altman-artificial-intelligence-regulation.html*, 2023.

B. Kim. *Interactive and interpretable machine learning models for human machine collaboration*. PhD thesis, Massachusetts Institute of Technology, 2015.

B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, 2018.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm ProPublica. *See https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm*, 2016.

T. Lattimore and C. Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

B. J. Lengerich, R. Caruana, M. E. Nunnally, and M. Kellis. Death by round numbers and sharp thresholds: How to avoid dangerous ai ehr recommendations. *medRxiv*, 2022.

J. Lin, C. Zhong, D. Hu, C. Rudin, and M. Seltzer. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, 2020.

Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 2018.

S. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S. I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2020.

T. McGrath, A. Kapishnikov, N. Tomašev, A. Pearce, M. Wattenberg, D. Hassabis, B. Kim, U. Paquet, and V. Kramnik. Acquisition of Chess Knowledge in AlphaZero. *Proceedings of the National Academy of Sciences*, 2022.

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

H. Narasimhan, W. Jitkrittum, A. K. Menon, A. Rawat, and S. Kumar. Post-hoc estimators for learning to defer to an expert. In *Advances in Neural Information Processing Systems*, 2022.

OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining*, 2016.

M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, 2018.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 2019.

C. Rudin, C. Wang, and B. Coker. The age of secrecy and unfairness in recidivism prediction. *Harvard Data Science Review*, 2020.

C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 2022.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

S. J. Russell. *Artificial Intelligence: A Modern Approach*. Pearson Education, Inc., 2010.

W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 2021.

R. Sayres, A. Taly, E. Rahimy, K. Blumer, D. Coz, N. Hammel, J. Krause, A. Narayanaswamy, Z. Rastegar, D. Wu, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*, 2019.

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, 2017.

K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *AAAI/ACM Conference on AI, Ethics, and Society*, 2020.

M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, 2017.

The New York Times. New navy device learns by doing. *See https://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html*, 1958.

P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. Soyer, I. Zalaudek, and H. Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 2020.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 2017.

D. H. Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 1996.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *13th European Conference on Computer Vision*, 2014.

D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.