Explaining Black Box Algorithms:

Epistemological Challenges and Machine Learning Solutions

Thesis submitted in partial fulfilment of the requirement for the degree of
DPhil in Information, Communication and the Social Sciences
at the Oxford Internet Institute, University of Oxford

David Watson
Oxford Internet Institute, Green Templeton College

Supervisors: Prof. Luciano Floridi & Dr. Matthew Kusner
January 2021

71,249 words

Abstract

This dissertation seeks to clarify and resolve a number of fundamental issues surrounding algorithmic explainability. What constitutes a satisfactory explanation of a supervised learning model or prediction? What are the basic units of explanation and how do they vary across agents and contexts? Can reliable methods be designed to generate model-agnostic algorithmic explanations? I tackle these questions over the course of eight chapters, examining existing work in interpretable machine learning (iML), developing a novel theoretical framework for comparing and developing iML solutions, and ultimately implementing a number of new algorithms that deliver global and local explanations with statistical guarantees. At each turn, I emphasise three crucial desiderata: algorithmic explanations must be causal, pragmatic, and severely tested.

In Chapter 1, I introduce the topic through real world examples that vividly demonstrate the ethical and epistemological imperative to better understand the behaviour of black box models. A literature review follows in Chapters 2 and 3, where I situate the project at the intersection of critical data studies, philosophy of information, and computational statistics. In Chapter 4, I examine conceptual challenges for iML that result in misleading, counterintuitive explanations. In Chapter 5, I propose a formal framework for iML – *the explanation game* – in which players collaborate to find the best solution(s) to explanatory questions through a gradual procedure of iterative refinements. In Chapter 6, I introduce a novel test of conditional independence that doubles as a flexible measure of global variable importance. In Chapter 7, I combine feature attributions and counterfactuals into a single method that retains and extends the axiomatic guarantees of Shapley values while rationalising results for agents with well-defined preferences and beliefs. I conclude in Chapter 8 with a review of my results and a discussion of their significance for data scientists, policymakers, and end users.

Contents

Abbreviations

ACM Association for Computing Machinery
AI Artificial intelligence
ATE Average treatment effect
ATT Adaptive thresholding test
BCM Bayesian case model
BCN Bayesian causal network
CART Classification and regression tree
CATE Conditional average treatment effect
CDS Critical data studies
CE Cross-entropy
CI Conditional independence
COMPAS Correctional offender management profiling for alternative sanctions
CORELS Certifiably optimal rule lists
CLP Context, level of abstraction, and purpose parameters
CPI Conditional predictive impact
CRAN Comprehensive R Archive Network
CRT Conditional randomisation test
CTT Correctness theory of truth
DAG Directed acyclic graph
ERM Empirical risk minimisation
FAT Fairness, accountability, and transparency
FDR False discovery rate
GAN Generative adversarial network
GDPR General Data Protection Regulation
GLM Generalised linear model
GUI Graphical user interface
ICE Individual conditional expectation
iML Interpretable machine learning
i.i.d. Independent and identically distributed
IP Intellectual property
ITE Individual treatment effect
LIME Local interpretable model-agnostic explanations
LoA Level of abstraction
LOCO Leave one covariate out
LORE Local rule-based explanations

LRP Layer-wise relevance propagation
LRT Likelihood ratio test
MAE Mean absolute error
ML Machine learning
MMD Maximum mean discrepancy
MNIST Modified National Institute of Standards and Technology
MMCE Mean misclassification error
MSE Mean square error
NTA Network theory of account
NeurIPS Neural Information Processing Systems
OLS Ordinary least squares
PaP Permute and predict
PDP Partial dependence plot
PS Prototype selection
PyPI Python package index
QII Quantitative input influence
RCT Randomised control trial
SCOT Social construction of technology
SHAP Shapley additive explanations
STS Science and technology studies
RKHS Reproducing kernel Hilbert space
TMLE Targeted maximum likelihood estimation
VI Variable importance
VIM Variable importance measure
VC Vapnik-Chervonenkis
WLS Weighted least squares
xAI Explainable artificial intelligence

List of Tables

List of Figures

Acknowledgements

I must begin by thanking my parents, Ian Watson and Susana Epstein (née Silva). Both were the first in their families to go to high school. They faced challenges in their youth that I cannot quite fathom – broken homes, deceased parents, political repression – only to immigrate to the United States as adults and find each other in a New York University graduate programme. They raised me in a house full of love, siblings, books, and, above all, an unwavering commitment to education. Dinnertime was a stage to debate the putative merits of liberalism, the intelligibility of a collective unconscious, the origins of realism in Western drama – all over a homemade *tortilla española*, with the Yankees playing in the background. I had the good fortune of growing up a stranger to poverty and hunger, never doubting for a moment my own safety or security. I appreciate that these are rare privileges in human history and hardly guaranteed in the modern world. For that and so much more, I owe my parents a debt of gratitude that can never be repaid. Thank you, mom and dad.

My siblings – Violeta, Gaspar, and Carmen – are an inspirational trio of educators, with expertise ranging from film history to Spanish literature and sexual health. Their example was formative in my early years, and, though we have spread out across the globe since growing up in a crowded Harlem apartment (with just a single bathroom for the six of us!), their imprint upon my mannerisms, tastes, and sense of humour persists. They are very literally a part of me, and shall forever be. Moreover, they are directly responsible for my eight nieces and nephews – Mora, Joaquín, Vittorio, Aleli, Federico, Mila, Junot, and Penelope – who have taught me lessons about humanity that cannot be learned in any classroom. I love you all deeply, and cannot thank you enough.

My thesis advisor, Luciano Floridi, has served as my primary academic mentor since Michaelmas 2014, when I began my MSc. I still remember the excitement I felt reading his 2011 monograph, *The Philosophy of Information*, as a frustrated Editorial Assistant at HarperCollins just a couple of years out of my undergraduate degree. I found myself variously discovering truths I had always dimly intuited and devising counterarguments I was sure would force his capitulation. I was probably wrong on both counts, but the enthusiasm was undeniable. Since then, we have collaborated on numerous projects. Some of my fondest Oxford memories include meetings in Luciano's office, debating works by our favourite (and least favourite) philosophers, trading jokes and recommended readings. As I transition from doctoral candidate to early career academic, I am grateful and fortunate to have Luciano as not just my graduate thesis advisor (twice over!) but as a colleague and friend.

I spent the second year of my DPhil on a doctoral enrichment studentship at the Alan Turing Institute in London. I met a number of great scholars there, whose influence on my thinking is peppered throughout this thesis. I am especially grateful to Matthew Kusner, my

second thesis advisor, who kindly agreed to continue working with me as he moved from the University of Oxford's Computer Science Department to University College London (UCL)'s Centre for Artificial Intelligence. From his new position at UCL, Matt introduced me to his colleague Ricardo Silva, whose causality research group has become a second academic home for me. Other Turing Institute Research Fellows to whom I am grateful for guidance include David Leslie, Rajen Shah, Chris Russell, and Chris Holmes.

I owe an enormous thank you to Michael Barnes, my former supervisor at Queen Mary University of London's Centre for Translational Bioinformatics (C4TB). Mike hired me right out of my MSc, as a young and woefully underqualified data scientist. I learned more doing applied projects at his lab than I did in years of classroom study. Mike imbued me with a profound appreciation for the complexity and wonder of molecular biology, and introduced me to many of the machine learning tools that have kept me busy ever since. His support was crucial both in my original doctoral application and throughout my DPhil. I hope to continue working with Mike and the rest of the C4TB team for years to come.

I am thankful to my department, the Oxford Internet Institute (OII), where I have met a diverse array of faculty and students who have expanded and challenged my views on a wide range of topics. From digital governance to ethical design, privacy to anti-trust regulation, the OII is at the forefront of an emerging interdisciplinary research agenda for twenty-first century social science, and it has been my distinct pleasure to rub elbows with experts from all sectors of the institute. Among the many individuals there to whom I owe thanks, I want to single out Carl Öhman, who was invariably the first reader (and most withering critic) of my draft chapters. I am also grateful to Nahema Marchal for restoring my faith in quantitative political science, and to all my friends at the Digital Ethics Lab – Corinne Cath, Josh Cowls, Julia Słupska, Nikita Agarwal, Vincent Wang, Jessica Morley, Jakob Mökander, and Andreas Tsamadas, among others. The OII faculty have also been a tremendous resource throughout my graduate studies. I have received especially helpful guidance and encouragement from Ralph Schroeder, Brent Mittelstadt, Sandra Wachter, and Gina Neff.

A quartet of devoted assessors ushered me through the (occasionally byzantine) process of transfer and confirmation. My thanks to Taha Yasseri, Marco Scutari, Mariarosaria Taddeo, and Robin Evans for their time and commitment. Their thoughtful comments have significantly improved the quality of this manuscript.

I have had the good fortune to collaborate on several projects during the course of this thesis with researchers outside my department, including Marvin Wright at the Leibniz Institute in Bremen, David Kinney at the Santa Fe Institute, and Limor Gultchin at the Oxford Department of Computer Science. Their novel perspectives have enriched my own in lasting, unforeseen ways.

Finally, I want to thank my tireless partner, Brita Lee Cooper. Without her love and support, this project never would have been possible. Brita has been on this journey with me one way or another since my MSc years, hearing out my ideas (with occasionally unconcealed scepticism) and always encouraging me to do my best. She has familiarised herself with a considerable amount of useless jargon on my account, to say nothing of the draft chapters she has reviewed and college dinners she has attended. She has made me more coffees and sandwiches than I care to count, and shown me a baffling amount of affection that I confess I have not always deserved. It is a testament to our mutual love and stubbornness that we close this chapter of our lives as it began: together, overindulging on cheese, planning vacations we cannot afford.

I am sure that there are many important people I have failed to acknowledge here. To them, I offer a sincere if somewhat disappointing blanket apology. It takes a village to raise a child, I am told; it took at least a hamlet to write this thesis. Needless to say, any errors that remain in this manuscript are mine alone. I am eternally grateful for all the assistance I have received and wisdom I have gleaned during my doctoral research. I hope to pay it forward and then some.

Introduction

On 15 March, 2016, the computer program AlphaGo (Silver et al., 2016) defeated 18-time world champion Lee Sedol in the fifth and final game of the Google DeepMind Challenge Match, a weeklong tournament that pitted the world's greatest digital and analogue Go players against each other. The coup de grâce was superfluous. AlphaGo had already won three of the previous four games, a decisive and historic victory that many experts predicted was decades away. The ancient Chinese boardgame has long been considered a major test of artificial intelligence (AI), given its exponential complexity and the strategic foresight required to sustain or defend against coordinated attacks. A frequently cited combinatoric result states that there are more possible positions in a single game of Go than there are atoms in the observable universe (Tromp & Farnebäck, 2007), a fact relayed with some combination of relish and terror by news outlets covering the event.[1]

In May of that same year, investigative journalists at ProPublica published a controversial report entitled "Machine Bias" (Angwin et al., 2016). The authors focussed on a proprietary algorithm known as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), developed by Northpointe (now Equivant), a private firm whose slogan is "Software for Justice". COMPAS is a pretrial risk assessment tool that assigns scores to defendants with the intent of predicting their likelihood of reoffending within two years. It is currently used in nine US states. Poring over data from some 12,000 individuals arrested in Broward County, Florida between 2013 and 2014, ProPublica found evidence of systematic racial discrimination in the algorithm's risk assessments. COMPAS was twice as likely to mislabel Black subjects as high risk compared to whites, while whites were far more likely to be mislabelled as low risk. Researchers at Northpointe defended their model, arguing that equality of error rates is a misleading metric of algorithmic fairness, and that COMPAS meets a different (incompatible) fairness criterion known as predictive parity (Dieterich, Mendoza, & Brennan, 2016). In 2017, the US Supreme Court declined to hear the case of *Loomis v. Wisconsin*, in which a plaintiff alleged that COMPAS violated his right to due process by labelling him high risk based on a proprietary (and potentially discriminatory) dataset and algorithm that could not be challenged in court. The plaintiff, Eric Loomis, continues to serve his six-year prison sentence.

These two examples – AlphaGo and COMPAS – epitomise the opportunities and challenges posed by recent advances in AI research. Machine learning (ML) algorithms, which automatically detect and exploit patterns in data, have made enormous progress on a wide range of tasks in just the last few years. ML models are now as good as or better than humans

[1] Similar results have been reported for chess (Shannon, 1950), although there is some ambiguity in this case between "possible chess games" and "sensible chess games". See (Levy, 1988).

at diagnosing breast cancer (McKinney et al., 2020), discovering powerful antibacterial compounds (Stokes et al., 2020), and navigating complex virtual environments (Vinyals et al., 2019), to cite just a few random examples from within the last year. The trend is expected to continue in the near and long term, as hardware becomes increasingly efficient and datasets grow ever larger, providing engineers with all the ingredients they need to create more sophisticated tools for signal detection and processing.

However, these achievements have not been matched by commensurate advances in *algorithmic explainability* – the emerging subfield of ML devoted to helping users better understand computational predictions and/or models. On the contrary, many of the top performing algorithms, such as deep neural networks, are essentially black boxes. These dazzlingly complex inference engines are optimised for predictive accuracy, not user intelligibility. Such models cannot be easily explained or understood, even by the engineers who design them, leading some to question the utility of ML for sensitive applications where risks are high and user trust is essential (Lipton, 2017; Rudin, 2019). How can we trust what we cannot understand?

Algorithmic explainability raises profound conceptual and technical challenges that we are only just now beginning to confront. This dissertation seeks to clarify and resolve some of the most pressing issues in this nascent discourse. The project has major implications for practitioners who model complex systems; policymakers tasked with regulating the use of algorithms; and, ultimately, end users who will benefit from a better understanding of how and why an automated system has made some decision of consequence.

The thesis is split into three parts that (very) loosely track Aristotle's tripartite division of human activity: *theoria*, *praxis*, and *poiesis*. Much has been written about the proper interpretation of these terms (see, e.g., Parry, 2020). I intend here the simplest possible translation: thinking, doing, and making. In Part I, I review a wealth of literature from the social sciences, philosophy, and computational statistics to survey the landscape of current debates and ground the ensuing analysis (thinking). In Part II, I critically examine existing proposals through an epistemological approach, pointing out crucial failures and offering a novel framework in which to compare and design algorithmic explanations (doing). In Part III, I implement a number of flexible methods for explaining models and predictions with statistical guarantees (making). A more thorough explication of my research questions and thesis outline can be found in §1.4, once relevant background concepts have been introduced and defined.

The remainder of this chapter is structured as follows. In §1.1, I offer a brief historical overview of algorithmic explainability and highlight three goals that motivate research in this area. A typology follows in §1.2. I review the basic principles of supervised learning in §1.3. In §1.4, I articulate my research questions and outline the proceeding chapters. I address several points regarding data ethics and reproducibility in §1.5, before concluding in §1.6.

§1.1 A Field Emerges

Explainability is fast becoming a top priority in statistical research, where it is often abbreviated as xAI (explainable Artificial Intelligence) or iML (interpretable Machine Learning). I adopt the latter initialism here to emphasise my focus on supervised learning algorithms (formally defined in §1.3), as opposed to other, more generic AI applications. Following a number of workshops on algorithmic fairness, accountability, and transparency at major international computer science conferences, the Association for Computing Machinery (ACM) launched an annual interdisciplinary conference specifically devoted to these themes in 2018. The burgeoning field of iML gained a degree of legal urgency with the passage of the latest European Union General Data Protection Regulation (GDPR). This legislation, which went into effect in May 2018, sparked a global debate over the so-called "right to explanation" in cases where individuals are subject to automated decisions. Although there is disagreement among legal scholars as to the proper interpretation of the relevant GDPR articles (Selbst & Powles, 2017; Wachter, Mittelstadt, & Floridi, 2017), the issue is now firmly on the agenda of sociologists, philosophers, and computer scientists in both academia and industry – to say nothing of the policymakers tasked with synthesising their research into actionable proposals. Relevant literature from these areas is surveyed in Chapters 2 and 3.

As I shall argue more comprehensively in the following pages, there are three main reasons why we typically seek to explain the predictions of black box algorithms: to audit, to validate, and to discover. The first goal arises most frequently in high-stakes settings such as criminal justice and job applications, where it is important to ensure that computational models do not discriminate against historically disadvantaged groups. Testing the reliance of algorithms like COMPAS on protected attributes like race requires sophisticated explainability techniques. Cases such as these tend to be of particular interest to critical data studies scholars, who emphasise the social, economic, and political implications of using algorithms to accelerate or automate sensitive decisions.

Validation, by contrast, is more of an epistemological than an ethical goal. Even high-performing algorithms are prone to overfit their training data, resulting in unexpected behaviour on test sets. That is why many iML tools are designed to give engineers more systematic ways to probe a model's internal logic. For instance, the authors of the LIME algorithm (Ribeiro, Singh, & Guestrin, 2016), a sort of locus classicus for iML research, deliberately trained a faulty image classifier to distinguish between huskies and wolves using a biased dataset in which all and only the wolves appeared against a snowy background. When the model (predictably) mislabelled a test photo of a husky in the snow, LIME explained the error by selecting the portions of the image that drove the prediction – the background snow, rather than the foreground canine. See Fig. 1.

(a) Husky classified as wolf (b) Explanation

Figure 1.1. An image classifier mislabels a husky as a wolf due to the snow in the background. Training data deliberately included snow in all and only the wolf images. From (Ribeiro et al., 2016, p. 1143).

A final goal of iML, often overshadowed by the previous two, is to discover new properties of the target system. Imagine if AlphaGo could not only beat world champions at Go, but teach us how to improve our own game by explaining strategic decisions as it plays. What if similar approaches could be applied to successful algorithms in chemistry or particle physics? The prospect is not as fanciful as it may seem. Mining the insights of prognostic clustering methods is already common practice in computational biology (John et al., 2019; John et al., 2020). When statistical models outperform human experts, there is reason to believe they have learned some valuable information that we do not yet know. The scientific potential of advances in iML could therefore be substantial.

§1.2 A Typology of Algorithmic Explainability

Algorithmic explanations come in many flavours. The following typology is adapted from Molnar's (2020) textbook guide to iML, which provides a helpful overview of technical approaches and current state of the art. Roughly put, there are three key dichotomies that orient iML research: intrinsic vs. post-hoc, model-specific vs. model-agnostic, and global vs. local. A final consideration is what type of output the method generates. Each point is considered in turn.

Intrinsic vs. post-hoc. An intrinsically explainable algorithm is one that raises no intelligibility issues in the first place. Canonical examples include (sparse) linear regression and (short) rule lists. Some have argued that such models are the only ones that should be allowed in high-risk settings (Rudin, 2019). Unfortunately, many interesting real-world problems cannot be adequately modelled with intrinsically explainable algorithms. In such cases, we turn to post-hoc tools, which take some target model f as input and attempt to explain its predictions, at least near some region of interest.

Model-specific vs. model-agnostic. Model-specific iML solutions take advantage of the assumptions and architectures upon which particular algorithms are built to generate fast and

accurate explanations. For example, much work in iML has been specifically devoted to deep neural networks (Bach et al., 2015; Montavon et al., 2017; Shrikumar, Greenside, & Kundaje, 2017; Sundararajan, Taly, & Yan, 2017), an especially complex class of algorithms with unique explanatory affordances and constraints. Model-agnostic tools, on the other hand, strive for more general applicability. Treating the target function *f* as a black box, they attempt to explain its predictions with few or no assumptions about the data generating process. Model-agnostic approaches are especially useful in cases where the target model is protected by intellectual property (IP) laws, while model-specific methods are generally more efficient and reliable when *f*'s architecture is known.

Global vs. local. A global explanation helps the user understand the behaviour of the target model *f* across all regions of the feature space. This is generally difficult to achieve when *f* is complex. A local explanation, by contrast, is only meant to apply to the area near some particular point of interest. For instance, a properly specified linear regression is globally explainable in the sense that the model formula holds with equal probability for any randomly selected datapoint. However, a local linear approximation to some nonlinear *f* will fit best near the target point, and does not in general tell us anything about how the model behaves in remote regions of the feature space.

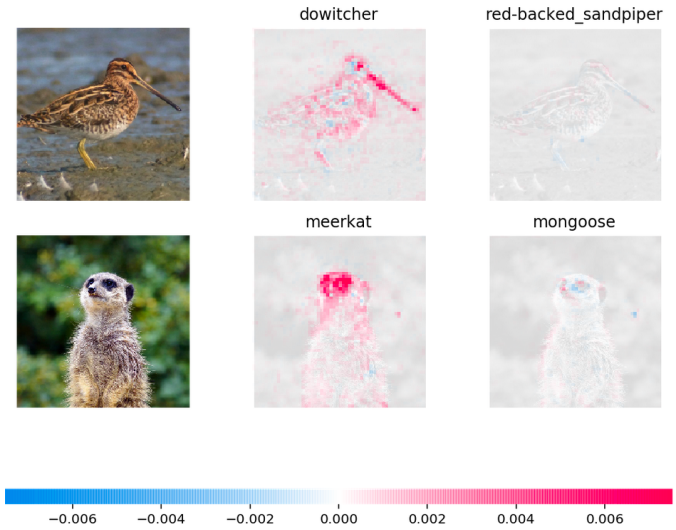

Figure 1.2. Saliency maps generated by SHAP (Lundberg & Lee, 2017), visualising the activation potentials of pixels for true (second column) and false (third column) labels.

A final axis of variation for iML tools is their output class. Typically, these methods explain predictions through some combination of images, statistics, and/or examples. Visual explanations are especially well-suited to image classifiers. Familiar techniques include superpixels, as in Fig. 1, or saliency heatmaps, as in Fig. 2. Other common visual approaches include plots that illustrate the partial dependence or individual conditional expectation of variables, which can inform users about potential causal effects and feature interactions (see Fig. 3). Statistical outputs, by contrast, may include rule lists, tables, or numbers quantifying

explanatory value in some predefined way. Finally, exemplary methods report informative datapoints, either in the form of prototypes (which typify a given class) or counterfactuals (which represent the most similar sample with a sufficiently different predicted outcome).

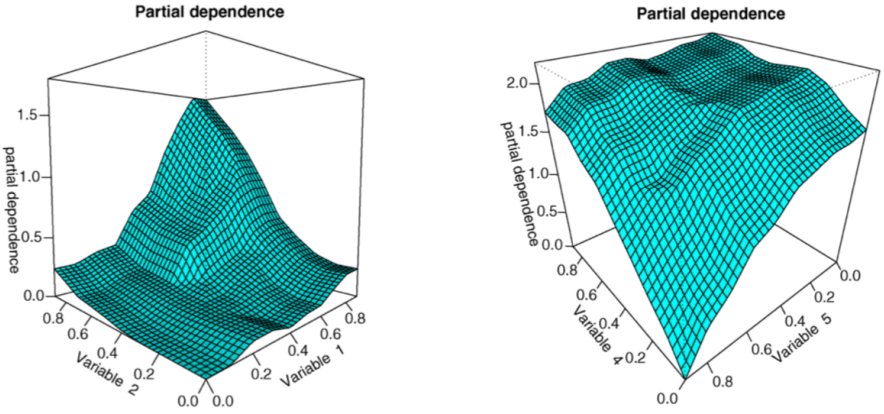

Figure 1.3. Three-dimensional partial dependence plots illustrating interaction effects between variables when integrating over the empirical range of covariates. See (Friedman, 2001).

Examples of all these iML types and tools will be critically analysed in Parts I and II. In Part III, I advance post-hoc, model-agnostic proposals at global and local resolutions, respectively. The outputs in both cases are primarily statistical, although I offer some accompanying plots to help visualise results.

§1.3 Supervised Learning

My focus in this dissertation is limited to supervised learning algorithms,[2] which may be informally defined as any method for finding predictive patterns in data. More precisely, supervised learners start with a training set of n data pairs $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$. The vector $\boldsymbol{x}_i$ denotes a point in p-dimensional space, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, with coordinate x_{ij} corresponding to the i^{th} value of feature X_j. Samples are presumed to be independently and identically distributed (i.i.d.) instances of some fixed but unknown joint distribution $P(\boldsymbol{X}, Y)$. The goal in supervised learning is to infer a function f: $\boldsymbol{X} \rightarrow Y$ that maps $\boldsymbol{x}$-vectors to y-outcomes. Performance is evaluated via some loss function that quantifies model error on an independent test set sampled from the same distribution.

It should be noted that I take an unusually permissive view of what sorts of models ought to count as instances of supervised learning. Following Vapnik (1995; 1998), the founder of statistical learning theory, I see no principled distinction between traditional techniques such as ordinary least squares linear regression and novel algorithms like gradient boosted forests. In both cases, the goal is to optimise predictive performance using data. The fact that linear regression happens to predate digital computers is a historical observation of little

[2] Unsupervised and reinforcement learning algorithms pose altogether unique explanatory challenges that are beyond the scope of this project.

mathematical significance. The fact that linear regression is based on many assumptions and gradient boosting based on relatively few is a difference of degree, not of kind. As for purported differences in model interpretability, it turns out that the arbitrary dichotomy between "old" and "new" algorithms does not map neatly onto the distinction between transparent and opaque models. The widely held dogma that linear regressions are somehow inherently simple cannot be seriously maintained by anyone who has actually fit one to a large dataset. With just a few hundred samples and a few dozen features, coefficients cannot feasibly be calculated by hand and model parameters are too numerous to simultaneously grasp or recall. Meanwhile, the recently published CORELS algorithm (Angelino et al., 2018), which relies on state-of-the-art branch-and-bound subroutines and highly efficient data structures, produces short rule lists that are specifically designed for easy interpretation. I cite these problems and examples just to say that it makes more sense to acknowledge a large, heterogenous class of supervised learning algorithms than to subdivide these models into artificial groups based on flimsy distinctions that fall apart under scrutiny.

The challenge with widening my scope to include all instances of supervised learning (broadly defined) becomes apparent in Part II, where I set about implementing model-agnostic statistics for testing the importance of feature subsets. The more restrictions we impose upon the target function class and underlying data generating process, the better we can tailor solutions to particular problems with high sensitivity and specificity. Unfortunately, however, it is unlikely that we will have access to such information in practice. IP laws protect copyrighted material including model architecture and parameters, while underlying mechanisms of data generation are notoriously difficult to identify or test. For these reasons, I follow the model-agnostic tradition in iML that seeks to build maximally general tools for computing post-hoc explanations with few or no assumptions about the target system.

§1.4 Research Questions and Project Outline

With these preliminaries out of the way, I can now articulate my research questions and outline the remainder of this thesis.

My goal in this dissertation is to bring philosophical and statistical rigour to the problem of explainability for black box models. For all the attention iML has recently received, important ambiguities remain regarding its true objectives and success criteria. This project seeks to resolve these fundamental issues over the course of eight chapters. As noted above, the thesis is split into three parts. Following a literature review in Part I, Parts II and III address conceptual and technical challenges, respectively. In Part II, the guiding question is:

RQ1. What constitutes a satisfactory explanation of a supervised learning model or prediction?

Drawing on copious literature from philosophy and the social sciences, I advance a particular vision of iML that places an explicit emphasis on causation, pragmatism, and error rate control – three considerations largely ignored by the leading post-hoc explanation tools available today. In Part III, I tackle a more practical question:

RQ2. Can reliable methods be developed for generating model-agnostic algorithmic explanations?

I respond in the affirmative, implementing a number of novel techniques for testing the global and local importance of feature subsets, thereby demonstrating how general-purpose explanations can be efficiently computed for a wide class of supervised learners and loss functions.

The remainder of this thesis is structured as follows. In Chapter 2, I survey the social scientific and philosophical literature on algorithmic fairness, accountability, and transparency, as well as more general theories of explanation. In Chapter 3, I review a range of influential iML proposals from computational statistics. In Chapter 4, I analyse three important conceptual challenges for iML that I argue are inadequately addressed by current methods, resulting in misleading and counterintuitive explanations. In Chapter 5, I propose a formal framework for iML – *the explanation game* – in which players collaborate to find the best solution(s) to explanatory questions through a gradual procedure of iterative refinements. In Chapter 6, I introduce a novel test of conditional independence that doubles as a flexible measure of global variable importance. I propose a local explanation procedure in Chapter 7, extending work in counterfactuals and Shapley values with a decision theoretic framework that allows for more flexible results. I conclude in Chapter 8 with a review of my results and a discussion of their significance for data scientists, policymakers, and end users.

The eight chapters of this thesis form an integrated whole. Chapters 2-7 have each been adapted as standalone papers that are all either published, under review, or in preparation. There is therefore some inevitable redundancy in the following pages, as relevant terms are reintroduced in service of different arguments and familiar dialectics recapitulated. I have attempted to streamline these repetitions in the present manuscript, although I am afraid they cannot be altogether avoided. I elected to write an integrated thesis due to the inherent interdisciplinarity of my doctoral project. Algorithmic explainability is a multifaceted topic that touches on longstanding debates in social science, epistemology, and statistics. No single methodological approach could possibly do justice to the breadth and complexity of this subject. Accordingly, the intended audience varies across the chapters. I am preparing a revised version of Chapter 2 for submission to *Philosophy of Technology*; a version of Chapter 3 will be submitted to *The 2020 Yearbook of the Digital Ethics Lab*; Chapter 4 is under review for a special issue of *Synthese* on recent topics in philosophy of statistics; Chapter 5 was published in *Synthese*; Chapter 6 was recently accepted by *Machine Learning*; and a revised version of Chapter 7 will be submitted to the ACM Conference on Artificial Intelligence, Ethics, and

Society (AIES). This diversity of outputs would not have been possible within the formal confines of a monograph.

Disciplinary promiscuity notwithstanding, there is a singular thematic focus to the present work. All chapters orbit around the core topic of algorithmic explainability, albeit from varying perspectives and emphasising different problematics. The same themes continually reappear throughout – especially the foundational, complementary roles played by causality, pragmatics, and testing. The chapters build on one another, most clearly within their respective parts, and together advance a unified critique of the current iML discourse, as well as a conceptual and technical toolkit for how to analyse existing proposals and design better solutions.

§1.5 Ethics and Reproducibility

This thesis does not use any personal data, as defined in Article 4(1) of the GDPR. All data analysed herein are either simulated or taken from public data repositories such as the University of California Irvine Machine Learning repository (Dua & Graff, 2017) and the gene expression omnibus database (Clough & Barrett, 2016). Real-world datasets are all either aggregated, anonymised, or otherwise distanced from their original data subjects, where applicable. More information on particular datasets, including references to original publications, can be found in the relevant chapters.

All empirical analysis was conducted in the `R` statistical computing environment. Code for reproducing all results and figures can be found in the dedicated GitHub repository: https://github.com/dswatson/dissertation.

§1.6 Conclusion

I conclude this introductory chapter with a reminder of the opportunities and challenges this topic presents.

AlphaGo's achievements may seem impressive, but they have been long since eclipsed by more recent software updates. The original algorithm was trained on some 160,000 matches by human experts, supplemented with a few million rounds of self-play (Silver et al., 2016). The following year, DeepMind unveiled AlphaGo Zero (Silver et al., 2017). This model required just one twelfth the compute power of its predecessor and took only three days to train, compared to several months for AlphaGo. More importantly, the new program never saw a single human match. Armed with nothing but the rules of the game and a well-designed loss function, AlphaGo Zero defeated AlphaGo by 100 games to 0. In a follow-up paper, DeepMind released AlphaZero (Silver et al., 2018), a generic reinforcement learning algorithm similarly unencumbered by human data. At the time of writing, AlphaZero is the reigning world champion of chess, shogi, and Go.

Meanwhile, technical progress on algorithmic fairness and explainability has been relatively slow. The most important advances in recent years have not been clever new ways to ensure just outcomes in ML – although, as chronicled in Chapter 2, many have been proposed – but impossibility results demonstrating the incompatibility of various intuitive formalisations of fairness. Explainability has arguably come somewhat further, especially in the model-specific literature. However, as I argue in Chapters 3 and 4, leading model-agnostic proposals often fail to work as advertised. If AI continues to move fast and break things, then iML will need to accelerate its effort to clean up the mess.

Part I: *Theoria*

In Defence of Sociotechnical Pragmatism

§2 Abstract

Interpretable machine learning is not an island, entire of itself. On the contrary, it is inextricably linked to a number of ongoing debates in sociology and philosophy that occasionally go back centuries. In this chapter, I present a narrative literature review in two parts. First, I examine the scholarship on fairness, accountability, and transparency in machine learning through a dialectical interplay between two competing ideologies. *Sociotechnical dogmatism* holds that society is full of inefficiencies and imperfections that can only be solved by better algorithms; *sociotechnical scepticism*, by contrast, opposes many or most instances of automation on principle. I argue that both perspectives are reductive and unhelpful, instead advocating for a pragmatic synthesis. This pragmatism re-emerges in the following section on philosophical theories of explanation, where I review a number of epistemological models designed for the natural sciences. I endorse a brand of causal pragmatism that combines elements of Floridi's levelism and Mayo's reliabilism to place a special emphasis on notions of agency and trust.

§2.1 Introduction

In March 1811, a group of workers in Nottingham, England began destroying textile machinery in protest against high unemployment and low wages. The practice quickly spread across the industrial regions of the north. By 1812, government officials were sufficiently concerned about the growing unrest that "machine breaking" had become a capital offense. The Luddite Rebellion would last until 1816, when 12,000 British Army troops – a force not much smaller than the 15,000 Wellington led into battle against Napoleon in the Peninsular War just a few years prior – suppressed the movement in a series of violent skirmishes.

Today, the word "Luddite" has become a more or less derogatory term – despite efforts by some authors to reclaim the label (Jones, 2006; Sale, 1996) – suggesting a backward and reactionary stance toward technological innovation. The negative connotations are not entirely fair. As several commentators have noted, the Luddites took no issue with technology as such, but were instead focused on ensuring better conditions for workers in a rapidly industrialising economy with no semblance of an organised labour movement. In an influential essay on the topic, Hobsbawm observes that "collective bargaining by riot was at least as effective as any other means of bringing trade union pressure, and probably *more* effective than any other means available before the era of national trade unions to such groups as weavers, seamen and coal-miners" (1952, p. 66). The episode is perhaps best understood as an early instance of a long running tension between forces intent on promoting greater automation

(primarily if not exclusively for financial gain) and those who resist this impulse (typically out of concern for the potential injustices that will result). I shall refer to these two camps as the *sociotechnical dogmatists* and *sceptics*, respectively. Neither group is entirely homogeneous, and the particular arguments advanced by their proponents inevitably vary from case to case. However, the struggle between the two is a persistent and instructive feature in the history of technological development.

In modern times, the focus of this dialectic has shifted from industrial to digital technologies – especially to paradigm shifting advances in machine learning (ML) algorithms, which are increasingly pervasive in both the public and private spheres. ML systems not only mediate our experience of news (Newman et al., 2019), entertainment (Morris, 2015), and one another (Turkle, 2017); they also guide decisions in healthcare (Topol, 2019a), criminal justice (Završnik, 2019), and cybersecurity (Taddeo, McCutcheon, & Floridi, 2019), to cite just a few prominent examples. ML algorithms have powered the rise of so-called "tech giants" – seven of the ten most valuable public companies in the world today are technology firms, with a cumulative market capitalisation of some $8 trillion – who actively fund further research and development into ML to further solidify their market advantage. This state of affairs provides ample ammunition for both the sociotechnical dogmatists, who marvel at the value created by these algorithms, and sociotechnical sceptics, who watch aghast as state and corporate interests use this technology to consolidate power and automate historical injustices.

In §2.2, I will both build up and tear down this purported dichotomy, which I argue is ill-equipped to conceptualise the opportunities and challenges posed by issues of fairness, accountability, and transparency in ML. I survey a range of critical data studies (CDS) scholarship – including legal, social, cultural, and technical literatures – tracing the dialectical interplay between sociotechnical dogmatists and sceptics from Marx up to the present day. I ultimately endorse a pragmatic synthesis that goes beyond this reductive dualism.

In §2.3, I turn to philosophical theories of explanation, where pragmatic approaches are once again uniquely able to cut through the fog of age-old debates. With a focus on twentieth century analytic theory, I examine logical, counterfactual, and interventionist accounts of explanation in the natural sciences. I embrace a hybrid brand of causal pragmatism that incorporates elements of Floridi's levelism and Mayo's reliabilism, emphasising the essential role of agency and trust. I argue that these are necessary components of any algorithmic explanation intended to make predictions more intelligible for end users of complex ML models.

§2.2 The Politics of Algorithms

There is a large and growing body of CDS literature devoted to the ethical, legal, and social implications of algorithms. This scholarship originally emerged out of a focus on issues of fairness, accountability, and transparency – notions collectively acronymised as FAT, the

erstwhile name of an annual conference on the subject organised by the Association for Computing Machinery (ACM) that began meeting in 2018.[1] The goals of the FAT ML movement are distinct yet interrelated. I argue that all three crucially rely on the development of more general and sophisticated methods for algorithmic *explanation*. Without a rigorous account of a statistical model's behaviour, we cannot evaluate whether it is fair, hold bad actors accountable, or promote greater technological transparency.

CDS is an inherently interdisciplinary undertaking. In a special issue of *Big Data & Society* devoted to the topic, Iliadis & Russo (2016) define CDS as a "nascent field…a formal attempt at naming the types of research that interrogate all forms of potentially depoliticized data science and to track the ways in which data are generated, curated, and how they permeate and exert power on all manner of forms of life" (p. 2). This is a wide remit. In what follows, I shall attempt to do justice to the methodological heterogeneity of CDS – drawing on legal, social, and epistemological debates – while focusing on the core issues of FAT ML, with special emphasis on the undergirding role of algorithmic explainability. I will therefore generally avoid the copious CDS literature on topics like data privacy, digital governance, and media studies – important though they are – except insomuch as they occasionally impinge upon the present discussion.

The remainder of this section is structured as follows. In §2.2.1, I highlight the social impact of algorithms (actual and potential) and review legal frameworks in the EU and the US intended to regulate these technologies through greater explainability. In §2.2.2, I examine a number of influential monographs and articles that have framed the popular and scholarly debate on algorithmic explainability. I devote §2.2.3 specifically to algorithmic fairness, which is arguably the most mature subfield of CDS. I situate the current discourse in a long-running dialectic between rationalist and critical tendencies, ultimately embracing a pragmatic synthesis in §2.2.4.

§2.2.1 The Sociolegal Landscape

Perhaps the most widely cited reason to explain algorithms is their large and growing social impact. Controversial applications of this technology include (but are not limited to) predicting criminal recidivism (Angwin et al., 2016), predictive policing (Perry et al., 2013), job hiring (Upadhyay & Khandelwal, 2018), credit scoring (Lessmann et al., 2015), student admissions (Dennis, 2018), clinical medicine (Topol, 2019b), military threat assessment (Nasrabadi, 2014), and cybersecurity (Taddeo, 2019). In each case, failure to properly screen training datasets for biased inputs threatens to automate injustices already present in society. For

[1] In 2019, the ACM conference was renamed FAT* (pronounced "FAT star"); in 2020, it was renamed FAccT. In what follows, I will continue to refer to the general movement for greater fairness, accountability, and transparency in ML by its more obvious (if somewhat unfortunate) acronym.

instance, studies have found that algorithmic profiling consistently shows online advertisements for higher paying jobs to men over women (Datta et al., 2015); that prominent recommender systems have erroneously suppressed content with homosexual themes as "adult" (Gillespie, 2014); and that facial recognition software is often trained on predominantly white subjects, making them inaccurate classifiers for Black and brown faces (Buolamwini & Gebru, 2018). A recent study in *Science* found evidence of significant racial bias in a widely used healthcare screening algorithm that affects millions of Americans (Obermeyer et al., 2019). Simulations suggest that rectifying the disparity would nearly triple the number of Black patients receiving medical attention.

Concern over the potential harms posed by opaque ML models is evident in the EU's 2018 General Data Protection Regulation (GDPR). Unfortunately, there is little consensus as to what the law in fact says regarding the so-called "right to explanation". Some commentators find the relevant protection in Article 22, which affords data subjects the right to contest algorithmic decisions (Goodman & Flaxman, 2017); or else in Articles 13-15, which guarantee the right to "meaningful information about the logic involved" in algorithmic decisions (Selbst & Powles, 2017). Others have challenged both readings, arguing that the text of the GDPR is too restrictive and unclear (Edwards & Veale, 2017), not to mention full of loopholes that firms can easily exploit (Wachter et al., 2017). No matter who is correct – the issue will likely remain undecided until a relevant case is brought before the European Court of Justice – there is no question that EU policymakers are beginning to seriously consider the social impact of ML, and perhaps even take preliminary steps towards regulating the industries that rely on such technologies (HLEGAI, 2019; OECD, 2019).

Lawmakers in the US, meanwhile, have not been as eager to update their regulatory frameworks. Textual guidance on matters of algorithmic bias is found primarily in Title VII of the 1964 US Civil Rights Act, the Fair Credit Reporting Act of 1970, the Equal Credit Opportunity Act of 1974, and the Equal Employment Opportunity Commission's 1978 Uniform Guidelines. As Barocas & Selbst argue in their influential *California Law Review* essay, "Big data's disparate impact" (2016), these laws may permit algorithmic discrimination if the statistical associations a model learns to exploit are sufficiently informative with respect to a target variable. For instance, it may well be true that some city's police department records show higher crime rates in minority neighbourhoods. Should a predictive policing algorithm therefore recommend a stronger police presence in these areas? Contextual specifics matter – e.g., a history of discriminatory policing in the city – but the law does not unambiguously rule out such practices. Barocas & Selbst underscore the dangers of relying on existing guidelines and case law in this new era of automated decision systems, mapping out the technical and political challenges that must be met to protect disadvantaged groups.

In a follow up article published in the *Fordham Law Review* (2018), the authors shift their focus from fairness to explainability, which they acknowledge is a prerequisite for judging an algorithm's reliance on protected attributes. Selbst & Barocas criticise current disclosure laws in the US and the EU for failing to distinguish between predictions that are *inscrutable* (i.e., incomprehensibly complex) and those that are *nonintuitive* (i.e., surprising and unobvious). Since "intuition serves as the unacknowledged bridge between a descriptive account and a normative evaluation" (p. 1086), regulations that focus exclusively on algorithmic inscrutability cannot in principle determine whether a particular algorithm's predictions are ethically defensible. For instance, explanations may fail to provide recourse, as legally required by laws on credit scoring in the US, if they do not recommend actionable steps data subjects can take to change their algorithmic predictions. This topic will be revisited in the discussion on counterfactual explanations in §3.4.

The dangers are no less acute under the GDPR, which does not regard inferences drawn from statistical models as personal data. This significantly curtails the rights of individuals to exercise control over such inferences, no matter how unreasonable or even discriminatory they may be (Wachter & Mittelstadt, 2019). The issue of unreasonable or counterintuitive associations in ML models is difficult if not impossible to parse without post-hoc explanation tools, highlighting the unexpected intersection between explainability and data privacy, understood as both an individual and a group right (Floridi, 2014; Mittelstadt, 2017). A whitepaper by an interdisciplinary team of legal scholars, software engineers, and cognitive scientists based at Harvard's Berkman Klein Center argues that new tools for explanation are required to meet the regulatory challenges ahead, and expresses optimism that ML solutions should be technically feasible (Doshi-Velez & Kortz, 2017).

§2.2.2 Framing the Debate

The proliferation of ML in various public and private sector contexts has led to a surge in literature on the topic, not just in academic journals but also in the popular press and trade publishing. A number of book-length works have garnered particular attention. One of the earliest instances was Pasquale's *The Black Box Society* (2015), in which the author highlights the dangers of unregulated algorithms in tech, finance, and government surveillance. Pasquale argues that intellectual property (IP) laws have been cynically deployed by powerful actors to promote their own interests and avoid external oversight. In *Weapons of Math Destruction*, O'Neil (2016) extends the analysis to education, advertising, and criminal justice. She demonstrates how algorithms implement pernicious feedback loops that disproportionately impact vulnerable communities. "Ill-conceived mathematical models," she writes, "are opaque, unquestioned, and unaccountable, and they operate at a scale to sort, target, or 'optimize' millions of people" (p. 17). Eubanks examines the effects of such technologies on poor

Americans in *Automating Inequality* (2018), while Noble provides an intersectional critique of Google search results in *Algorithms of Oppression* (2018). In *Artificial Unintelligence*, Broussard (2018) coins the term "technochauvinism" (akin to the aforementioned sociotechnical dogmatism) to describe the irrational preference for technological solutions over all others. Most recently, Zuboff (2019) argues that tech giants have inaugurated *The Age of Surveillance Capitalism*, in which human experience is systematically processed into behavioural data and used to develop prediction products that undermine autonomy and democracy.

These monographs track a similar trend in academic publications over the last decade, where the focus gradually shifted away from "big data" (the buzzword of the 2000s) and toward "artificial intelligence" (the buzzword of the 2010s). For instance, a 2016 special issue of *Philosophical Transactions of the Royal Society A* was devoted to the ethical impact of data science. In the introductory article, Floridi & Taddeo (2016) highlight how algorithms pose unique moral challenges independent of their informational inputs or practical implementations. Mittelstadt et al. (2016) identify six types of ethical concerns raised by algorithms: (1) inconclusive evidence; (2) inscrutable evidence; (3) misguided evidence; (4) unfair outcomes; (5) transformative effects; and (6) traceability. The first three are rooted in epistemic, data-centric issues, while items (4) and (5) are more obviously normative and algorithmic. The sixth is an "overarching concern" (p. 4), which demands that we follow the entire inferential pathway of a given prediction, from preliminary data gathering to model training and deployment. The authors argue that only through this cautious and painstaking procedure can moral responsibility be properly apportioned in complex sociotechnical systems. Of course, the goal of algorithmic traceability, which unifies ethical concerns (1)-(5), can only be realised if we have the proper tools to explain predictions and models with sufficient clarity to interrogate their epistemic and normative consequences.

A 2017 special issue of *Information Communication and Society* was devoted to the social power of algorithms. In the introductory article, Beer similarly highlights the inescapable role of explainability. "[W]e need to understand what algorithms are and what they do in order to fully grasp their influence and consequences" (2017, p. 3), he concludes. This goal is complicated by several obstacles. Burrell (2016) identifies three sources of algorithmic opacity: (1) intentional corporate or state secrecy; (2) technical illiteracy; and (3) inherent complexity. The first of these echoes Pasquale's concerns about IP law and Zuboff's critique of surveillance capitalism. The second amounts to a call for more widespread education in computer science, which will be essential both to foster informed debate about appropriate regulations and to ensure that a diverse community of programmers is engaged in designing these powerful tools. My focus, however, is on item (3) in Burrell's list.

Inherent complexity has long been used as an excuse to dismiss any critical discussion of algorithmic intelligibility. This theme emerges time and again in popular and scholarly

analyses, which almost unanimously insist upon a natural, inevitable tension between predictive accuracy and human intelligibility in ML. The typical framing runs something like this. The most impressive results from the recent surge in AI research and funding have come from so-called "black box" algorithms like deep neural networks (Goodfellow, Bengio, & Courville, 2016), support vector machines (Schölkopf & Smola, 2017), and gradient boosted forests (Schapire & Freund, 2012). These sophisticated, high-performing regression and classification techniques are notoriously difficult to comprehend, often resulting in models with unmanageable parameters and hyperparameters that do not admit of any easy interpretation. While we may feel at home with our neat linear models and simple rule lists, the expressive power of such algorithms is severely limited in comparison with more exotic architectures and learning ensembles. Complex problems require complex solutions, and we should therefore abandon all hope of ever understanding our best ML models.

Not all commentators are resigned to this brand of statistical defeatism. In a recent article published in *Nature Machine Intelligence*, Rudin (2019) argues forcefully against this narrative, which she suggests is grounded in anecdotal evidence at best and corporate secrecy at worst. She observes that science has long shown a preference for parsimony, elevating Occam's Razor from a rule of thumb to an organising principle. She rejects the purported accuracy-interpretability trade-off, summarising her message in the article's blunt, memorable title: "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." The plea is founded on an important if somewhat subtle lesson from statistical learning theory. In a seminal article, Breiman (2001b) introduced the notion of a "Rashomon set", named after the 1950 Akira Kurosawa film in which four characters give very different accounts of a samurai's untimely death in 8th century Kyoto. Just as multiple witnesses to the same crime may provide inconsistent testimony, Breiman shows that ML models can approximate the same functional relationship through divergent learning strategies. The resulting models form a so-called "Rashomon set" of predictors, which perform similarly on test data but trace unique paths from input to output, e.g. relying on different feature subsets to compute conditional expectations.

Rudin draws two lessons from this parable. First, we should not trust post-hoc explanation methods, which may appear to mimic the target function but in fact converge on similar predictions via unrelated reasoning. Second, in any Rashomon set of sufficient size and complexity, we should expect to find at least one (globally) interpretable model. If we must use ML in high-stakes applications like healthcare and criminal justice, she concludes, then there is no reason to rely on opaque algorithms like deep neural networks. She backs up her claim with numerous instances of models optimised for transparency, such as SLIM (Ustun & Rudin, 2019), which computes risk scores through mixed integer linear programming, and CORELS

(Angelino et al., 2018), which learns sparse decision sets that are provably optimal under certain conditions.

Rudin makes some compelling points, but there are two key problems with her analysis. First, there is no general guarantee that some interpretable algorithm will outperform black box competitors or even be in the Rashomon set of high-performing models for any given prediction problem. This follows directly from the no free lunch theorem (Wolpert & Macready, 1997), which states (roughly) that there is no one-size-fits-all solution in ML. Any algorithm that is optimal on one class of problems will necessarily be suboptimal on another. Of course, this cuts both ways – any particular black box algorithm is likewise guaranteed to fail on some datasets. However, to the extent that we value performance above all – as we may rationally choose to do in certain high stakes settings where, say, lives hang in the balance – we will have to be open to models of variable interpretability.

Second, as Burrell's list reminds us, the opacity of black box algorithms is not just a by-product of complex statistical techniques, but of institutional realities that are unlikely to change anytime soon. The modern digital economy places an enormous premium on data and algorithms, which may well be amongst the most valuable assets on the books for many firms – and not just those explicitly devoted to information technology. History has shown that private interests are willing to go to extraordinary lengths to protect their IP, not only from potential competitors but from regulators and indeed any form of external scrutiny (Pasquale, 2015). Even if a firm were using an interpretable model to make its predictions – say, a sparse linear regression or short rule list – the model architecture and parameters would likely be subject to strict copyright protections. Some have called for the creation of independent third-party agencies tasked with the responsibility of auditing data and code under non-disclosure agreements (Floridi et al., 2018; Wachter et al., 2017), a sort of digital ombudsman to advocate for data subjects and protect consumer rights. While I am personally in favour of such an initiative, as of the time of this writing there exists no such body under EU, UK, or US law. Until such legislation is enacted, anyone attempting to monitor the fairness, accountability, and transparency of algorithms will probably have no choice but to treat the underlying technology as a black box.

§2.2.3 Fairness and Its Discontents

Given the potential harms posed by algorithmic bias in an increasingly automated information society, it should come as no surprise that *fairness* is a primary concern for CDS scholars. But how do we define fairness? A substantial subgenre of the FAT ML literature is devoted to formalising criteria in an explicit effort to answer this question. Prominent examples include:

- *Fairness through unawareness*. A model is fair if sensitive attributes A are not included in the training data.

- *Demographic parity*. A model is fair if predictions $\hat{Y}$ are independent of sensitive attributes, i.e. $\hat{Y} \perp A$.
- *Equality of opportunity*. A model is fair if predictions are independent of sensitive attributes after conditioning on the true outcome Y, i.e. $\hat{Y} \perp A|Y$.
- *Calibration*. A model is fair if outcomes are independent of sensitive attributes after conditioning on predictions, i.e. $Y \perp A|\hat{Y}$.

These definitions have been widely studied (albeit with frustratingly inconsistent nomenclature), and many common learning algorithms satisfy at least one of them (Berk et al., 2018; Chouldechova & Roth, 2018; Corbett-Davies et al., 2018; Donini et al., 2018; Verma & Rubin, 2018). Note that in all cases, the goal is to minimise the impact of sensitive attributes on model predictions, i.e. to ensure that sensitive attributes do not generally factor into the *explanation* of algorithmic outputs. This may be tested through an auditing scheme that permits little or no access to the underlying model, so long as auditors can compile a sufficiently large and representative dataset of $(A, Y, \hat{Y})$ instances.

More information and assumptions are required to meet the somewhat specialised criteria of individual and counterfactual fairness, which nonetheless share similar overarching goals:

- *Individual fairness*. A model is fair if predictions are similar for similar individuals. That is, given some distance measure $d(\cdot, \cdot)$ and small values of (ε, δ), samples i and j separated by $d(i, j) \leq \varepsilon$ should receive similar predictions with high probability, $\mathbb{P}(\hat{Y}_i = \hat{Y}_j) \geq 1 - \delta$. See (Dwork et al., 2012) for details.
- *Counterfactual fairness*. A model is fair if predictions do not change when a counterfactual intervention alters the value of a sensitive attribute, i.e. $\mathbb{P}(\hat{Y}|A = a) = \mathbb{P}(\hat{Y}|do(A = a'))$, where $a \neq a'$. See (Kusner et al., 2017) for details.

A thorough analysis of the advantages and disadvantages of these and other fairness criteria is beyond the scope of this literature review. For a comprehensive and multifaceted discussion, see (Barocas, Hardt, & Narayanan, 2019). For my purposes, it suffices to observe that while each arguably captures some intuitive notion of fairness, the sheer multitude of proposals is somewhat disconcerting. A tutorial held at the 2018 ACM FAT* conference surveyed no fewer than 21 competing definitions of algorithmic fairness (Narayanan, 2018). Others have emerged since (e.g., Kim, Reingold, & Rothblum, 2018; Romano et al., 2020; Sharifi-Malvajerdi, Kearns, & Roth, 2019). Impossibility theorems have shown that many of the most popular formal criteria are mutually incompatible except in trivial cases (Chouldechova, 2017; Friedler, Scheidegger, & Venkatasubramanian, 2016; Kleinberg, Mullainathan, & Raghavan, 2017). These results suggest that while mathematical formulae may help to clarify the trade-offs inherent in any socially sensitive decision-making context, they cannot in principle "solve" the problems posed by algorithmic fairness.

The impulse to machine learn our way out of fundamental social problems like systematic discrimination and structural inequality has been criticised as just another instance of naïve techno-solutionism (Morozov, 2013). The techno-solutionist, often cast as a Silicon Valley entrepreneur or venture capitalist, believes that the world is full of messy and inefficient processes just waiting to be sorted and optimised. The concept is hardly novel. Weber (1930) describes a similar impulse, which he dubs *rationalisation*, and famously argues that it characterises a distinctly modern set of cultural and institutional practices, from bureaucratic administrative states to capitalist modes of production. Solutionism also has roots in *technological determinism*, which holds that human relations and social structures are driven primarily or even exclusively by material development. Determinism is often associated with Marx (1867; 1885; 1894),[2] who views bourgeois factory owners as techno-solutionists of a sort, eager to promote a narrative of inevitable industrialisation while greedily wringing every last drop of surplus value out of their workers and production lines.

The critique of techno-solutionism can be traced from its origins in Marx and Weber through the writings of Frankfurt School philosophers – notably, Horkheimer & Adorno (1947) and Habermas (1981) – who forcefully challenge so-called Enlightenment ideals. They argue that these principles, which inaugurated an era of scientific rationalism (not to mention revolutionary politics) in Europe, ultimately engender a liberal mythology every bit as dangerous as the sociopolitical order it displaced. This backlash against the received dogma of science – particularly its behaviourist and positivist tendencies – arguably reached its apex with the social constructivism of Latour & Woolgar (1979) and the "strong programme" in the sociology of science championed by Bloor (1976) and his colleagues at the University of Edinburgh. These early proponents of science and technology studies (STS) question the privileged status of scientific realism and highlight the irreducibly social origins of all knowledge claims. Scholars in the social construction of technology (SCOT) tradition extend this scepticism to technology as well (Bijker, Hughes, & Pinch, 1987), arguing that it is human relations and social structures that shape the development of new technologies – not the other way around.

It may not be immediately obvious what this intellectual history has to do with FAT ML. However, these dialectics provide an informative backdrop for ongoing CDS debates regarding the extent to which it is appropriate or desirable to deploy algorithms in high-risk settings. Some techno-solutionists explicitly endorse the so-called "end of theory" perspective on big data, which clearly echoes the behaviourism of the early twentieth century. In a notorious *Wired* cover story from 2008, the magazine's then-Editor in Chief Chris Anderson embraces the view with aplomb:

[2] As an exegetical aside, there is some dispute over the true extent to which Marx was in fact a technological determinist, at least in the uncompromising sense that the label is occasionally employed by modern authors. See (Bimber, 1990).

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (Anderson, 2008)

This brand of radical instrumentalism elevates predictive accuracy over explanatory insight as the ultimate goal of inquiry. Whereas the traditional scientific method requires carefully designed experiments to isolate cause and effect, the rise of big data and artificial intelligence has allegedly inaugurated a new era of automated discovery requiring little or no human input (Hey, Tansley, & Tolle, 2009). Understanding why the model works is as impossible as it is irrelevant. The math behind these algorithms is too sophisticated, the datasets too vast for the mortal mind to possibly comprehend.

This line of reasoning is lazy and dangerous. As the examples from §2.2.1 vividly demonstrate, blind faith in the objective accuracy of algorithms is misplaced. The mythology of data-driven omniscience runs deep in the modern psyche, but the truth is that numbers can never "speak for themselves". Datasets are invariably gathered and curated by humans, encoding our assumptions, biases, and oversights at every turn (boyd & Crawford, 2012). Modern critics of techno-solutionism, such as Hoffmann (2019), argue that the logic of computation is fundamentally unfit to address problems of social injustice, as both are founded on the same rationalist mode of hierarchical labelling and sorting. The problem formulation itself – in terms of variables and averages, optimising metrics for prespecified groups – fails to question or even acknowledge how the very categories that algorithmic fairness seeks to protect are themselves socially constructed and potentially reductive (Hanna et al., 2020). Data science inspires what McQuillan calls a "machinic neoplatonism" (2018) that privileges hidden mathematical order over lived experience. Greater "transparency" does little to resolve these issues, and may even exacerbate them by normalising neoliberal modes of agency in which end users must navigate a complex marketplace of algorithmic alternatives with limited information (Ananny & Crawford, 2016).

If these critics are right, then the effort to advance technical solutions to algorithmic injustices will only add more fuel to the fire. Rather than formalise some new and improved fairness criteria or maximise social harmony with clever objective functions, these authors encourage us to rethink our relationship with technology and one another. They caution against any impulse – solutionist, rationalist, determinist – that prioritises quantitative modes of psychosocial organisation, analysis, and engagement to the exclusion of qualitative alternatives.

§2.2.4 The Pragmatic Turn

These objections are provocative and perspicacious – but they are not the last word on this debate. In most cases, the alternative to algorithmic decision systems is human decision

systems, and it is far from obvious whether this latter option generally promotes more just outcomes. Studies have demonstrated time and again that even the most well-intentioned of people are prone to implicit biases against historically disadvantaged groups (Greenwald & Krieger, 2006). Identifying and mitigating such biases with computational methods is not just a speculative ideal. Kleinberg et al. (2017) show that replacing or supplementing judicial bail decisions with a high-performance ML model results in huge welfare gains along practically any metric of interest. Simulated results find crime reductions of nearly 25% with no change in jailing rates, or jailing rate reductions of over 40% with no increase in crime rates. All gains are achieved while reducing racial disparities. In a subsequent paper, Kleinberg et al. (2018) make a strong case that increased automation can reduce discrimination by inaugurating rigorous, objective procedures for auditing and appealing ML predictions. While it is exceedingly difficult under current laws to prove that a person has engaged in discriminatory behaviour, it is relatively straightforward to test and even correct for algorithmic bias given minimal access to the target model and/or a sufficient training dataset (provided we have settled on some formal definition of fairness).

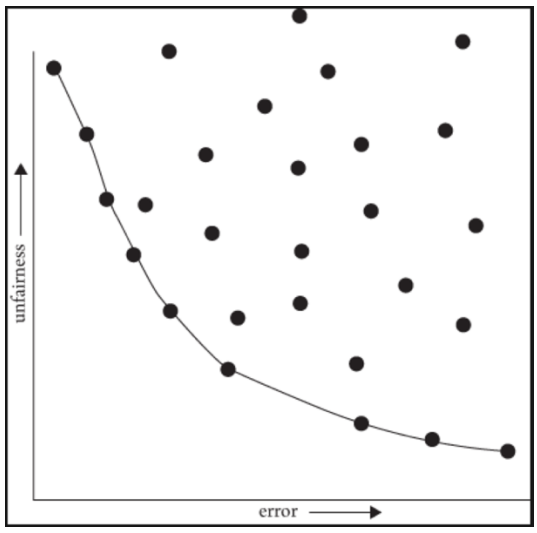

Figure 2.1. A schematic example of the Pareto frontier. Models are scored by their error and unfairness. Pareto-efficient solutions form a boundary beyond which no model can improve in either direction without incurring some loss in the other. From (Kearns & Roth, 2019, p. 127).

The Kleinberg et al. (2017) study is somewhat unusual for highlighting a case in which automation boosts both accuracy *and* fairness. Most literature in this area tends to start from the (not unreasonable) assumption that these two desiderata are in natural tension with each other. Even in such cases, there is nothing to be gained by ignoring the trade-off. We may describe the situation in terms of a *Pareto frontier*.[3] Imagine a two-dimensional space with axes for accuracy and fairness. Given some formal definition of each, we may score decisions (human or otherwise) along both axes and thereby locate them within this coordinate system

[3] The concept is named after Italian engineer and economist Vilfredo Pareto, who introduced the idea in his study of economic efficiency and income distributions. See (Pareto, 1935).

(see Fig. 2.1). We say that system *A* Pareto-dominates system *B* if and only if it is strictly better along at least one axis and no worse along the other. The Pareto frontier is constituted by the set of points that are not Pareto-dominated by any other point in this space – i.e., systems that cannot be made more accurate without becoming less fair, or made fairer without becoming less accurate. Note that there is no context-independent way to decide which point along the frontier we consider optimal, for this judgment depends upon our relative valuations of these two desiderata for this particular problem. What Kleinberg et al. (2017) effectively demonstrate is that in the case of bail decisions, human judges in their dataset are nowhere near the Pareto frontier.

In their 2019 book *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*, theoretical computer scientists Kearns & Roth argue that delicate trade-offs like this cannot be navigated without first confronting them head-on:

> While the idea of considering cold, quantitative trade-offs between accuracy and fairness might make you uncomfortable, the point is that there is simply no escaping the Pareto frontier. Machine learning engineers and policy-makers alike can be ignorant of it or refuse to look at it. But once we pick a decision-making model (which might in fact be a human decision-maker), there are only two possibilities. Either that model is *not* on the Pareto frontier, in which case it's a "bad" model…or it *is* on the frontier, in which case it implicitly commits to a numerical weighting of the relative importance of error and unfairness. Thinking about fairness in less quantitative ways does nothing to change these realities — it only obscures them. (2019, pp. 127-28)

The notion that putting numbers to problems like this somehow does violence to our underlying humanity or commits us to naïve techno-solutionism is itself reductive and wrong-headed. Quantitative methods are merely one tool among many for diagnosing and combatting social injustice, and an especially powerful one at that. Ignoring this option on philosophical grounds incurs a devastating opportunity cost that society can ill afford.

A more nuanced approach to these challenging issues – one that rejects both the dogmatic positivism of the solutionists and the self-defeating puritanism of the critical theorists – is possible. This perspective would need to be principled but flexible, acknowledging the irreducible context-dependence of particular judgments on algorithmic fairness and intelligibility without compromising its commitment to either ideal. It would have to be relational, not relativist, preserving the autonomy of individual agents to determine their own tolerance for error, complexity, and unfairness. Finally, it would have to be technically grounded, unafraid to combine quantitative and qualitative modes of reasoning to arrive at novel solutions.

I claim that the philosophy of *pragmatism* meets all these desiderata. Kleinberg et al. and Kearns & Roth provide some intuition for what a pragmatic approach might look like in the specific case of algorithmic fairness, but these principles can be extended more generally. The point will be explored from an epistemological perspective in §2.3, and put to work constructing novel iML solutions in Parts II and III.

§2.3 The Philosophy of Explanation

We have seen how all roads in the CDS discourse lead to explainability. There can be no algorithmic fairness, accountability, or transparency without some method for making ML models more interpretable. Often overlooked in this literature are certain fundamental questions about this desideratum perhaps considered too abstract to bother asking. To wit:

- What constitutes a satisfactory explanation?
- What are the basic elements of explanation?
- How do explanations advance knowledge?

These questions are squarely within the ambit of *philosophy* – the second of three disciplinary pillars undergirding this project.

Explanation is an ancient topic of philosophical interest, debate, and confusion. In the *Phaedo*, Plato writes – through the voice of his mentor, Socrates, speaking from his deathbed – that all inquiry begins with a call for explanation (Plato, 1997). Aristotle devotes large portions of his *Physics* and *Metaphysics* to expounding the doctrine of the four causes – formal, material, efficient, and final – that jointly explain the nature of things (Aristotle, 1984). More examples can no doubt be found in venerable philosophical traditions the world over. In this section, I will focus on modern analytic epistemology and philosophy of science, which provide an insightful and heterogeneous collection of formal theories and perspectives on explanation.

The remainder of §2.3 is structured as follows. I begin in §2.3.1 with a review of the deductive-nomological model, a sort of *locus classicus* for modern philosophical discourse on scientific explanation. I consider counterfactual and interventionist alternatives in §2.3.2, which resolve a number of difficulties with the deductive-nomological model but introduce new problems that are especially acute in iML. In §2.3.4, I examine the merits of epistemological pragmatism, which I argue preserves the most attractive components of previous theories while overcoming a number of major obstacles. The section concludes with a discussion of trust and testing in §2.3.5.

§2.3.1 The Deductive-Nomological Model

Contemporary philosophical analysis of scientific explanation arguably begins with Hempel, who attempts to boil causal reasoning down to its most basic formal elements (Hempel & Oppenheim, 1948; Hempel, 1965). According to his theory, the explanation for some physical event E consists of two components:

(1) a non-empty set of observation statements $S = \{s_1, \ldots, s_n\}$; and

(2) at least one law-like generalisation L, such that

$$(S \,\&\, L) \rightarrow E.$$

This account is *deductive*, insomuch as the explanandum follows logically from the explanans; and *nomological*, insomuch as it incorporates a law of nature as an essential premise of the

argument. Thus Hempel terms this the deductive-nomological (DN) model, an influential fusion of logic and science that is characteristic of the positivist tradition from which it emerged. The DN model works fairly well within the framework of classical mechanics, where the observation set S may contain, say, information about the position and momentum of some object x, which, in conjunction with Newtonian laws of motion L, entails a new proposition E that states the position of x at some future time.

This model no doubt boasts a certain formal elegance, but critics have persuasively argued that it provides neither necessary nor sufficient conditions for successful explanation. Hempel himself (1965) points out that the entailment relation purported to hold between explanans and explanandum is overly strong. Consider a probabilistic explanation of the form:

s_1: Patient a has infection x
s_2: Patient a receives treatment
L_1: 0% of untreated patients with infection x survive
L_2: 99% of treated patients with infection x survive
$\therefore$ E: Patient a survives

The conjunction of observations $S = \{s_1, s_2\}$ and statistical regularities $L = \{L_1, L_2\}$ confers high probability on the outcome E but does not logically entail it. Yet little or nothing about a's survival is left unexplained by S & L. If this critique is right, then the DN model does not provide necessary conditions for explanation.[4]

A more damning counterexample for the DN model comes from Salmon (1971, p. 34), who proposes the following DN-compliant explanation:

S: John Jones is a male who has been taking birth control pills regularly
L: All males who take birth control pills regularly fail to get pregnant
$\therefore$ E: John Jones fails to get pregnant

Intuitively, the issue with this explanation is that it violates some unstated causal relevance criterion. John Jones did not fail to get pregnant *because* he took birth control pills; he failed to get pregnant because

S: John Jones is a male; and
L: Males do not get pregnant.

His pill-taking is therefore causally irrelevant to the explanandum. If Salmon's argument is sound, then the DN model does not provide sufficient conditions for explanation.

The DN model has become something of a philosophical punching bag in the analytic tradition. Detractors certainly have a lot to work with, but we should be careful to recognise that a baby lurks somewhere in that bathwater. Hempel's categorical distinction between observation statements (i.e., particular claims) and law-like regularities (i.e., universal claims) is especially astute, as is the formal requirement that both types of propositions be involved in any successful explanation. Although his theory was designed with the natural sciences in

[4] Hempel (1965) proposes a new class of explanations called *inductive-statistical* (IS) to accommodate such cases, but the IS model struggles to account for low probability events. The alternatives analysed below are better equipped to handle statistical explanations.

mind, the DN model can easily accommodate algorithmic explanations as well. We simply explain the behaviour of some trained model f by combining input datapoints S with model parameters L, thereby rendering predicted outcome E. If the function f is deterministic – as function classes for most prominent ML algorithms are, at least at the token level (more on this in Chapters 4 and 5) – then the relationship between explanans and explanandum is even one of logical implication.

For reasons to be explained below, this account of algorithmic explanation will not quite do – but it is a start. Notable competitors to (and improvements upon) Hempel's DN model include the statistical relevance model (Salmon, 1971); the causal mechanical model (Dowe, 2000; Salmon, 1984); and the unificationist model (Kitcher, 1989). A thorough explication of these theories is beyond the scope of this literature review. For a good introduction to all three, see (Woodward, 2019). The following sections focus instead on counterfactual, interventionist, and pragmatic accounts, which I believe better capture the most important and interesting aspects of successful explanations.

§2.3.2 Counterfactuals and Interventionism

The counterfactual theory of explanation is most closely associated with Lewis (1973a, 1973b), who develops a rich metaphysics on the basis of his possible worlds semantics. Although Lewis famously endorses realism with respect to possible worlds, his account of explanation does not depend upon this ontological commitment. All that his theory requires is a similarity relation that distinguishes between worlds that are closer to and farther from the actual world. Truth conditions for counterfactual statements can then be defined as follows:

(1) 'If p were the case, then q would be the case' is true in the actual world if and only if (i) there are no possible p-worlds; or (ii) some p-world where q holds is closer to the actual world than is any p-world where q does not hold.

Ignoring the trivial case (i), we may employ this notion of counterfactuals to give truth conditions for causal dependence claims:

(2) Where c and e are two distinct actual events, e causally depends on c if and only if, if c were not to occur, then e would not occur.

Lewis argues that causal dependence is sufficient but not necessary for causation, as only the latter is a transitive relation (1973b). We can define causation more generally in terms of causal chains, which string together a finite sequence of successive causal dependencies:

(3) c is a cause of e if and only if there exists a causal chain leading from c to e.

This model can be generalised to handle issues of temporal asymmetry (Lewis, 1979) and probabilistic dependencies (Lewis, 1986). It has some difficulties accommodating instances of overdetermination and causal pre-emption – see (Menzies & Beebee, 2020) for an overview –

but later refinements incorporate contextual information that arguably defuses these objections (Lewis, 2000).

The counterfactual theory of explanation has been explicitly endorsed by at least one group of researchers working on iML (Wachter, Mittelstadt, & Russell, 2018). Their proposal is examined more closely in subsequent chapters. Briefly, the idea runs as follows. Say that some bank uses opaque ML model f to predict the creditworthiness of loan applicants. The model judges applicant a to be high risk, and the bank therefore denies her the loan. We would like an explanation why. One strategy would be to identify the statistical regularities that govern the model's outputs, at least in the region of a – this is (a localised version of) the DN approach – but these may be difficult to establish if f is complex and/or inaccessible. Another method would be to find the nearest neighbour on the opposite side of the decision boundary, i.e. the most proximal counterfactual applicant who receives the opposite prediction. Call this applicant a'. An explanation for why a was denied her loan can then be given by pointing out the differences between a and a', which should, by construction, be minimal. If those differences include sensitive attributes – if, for instance, the only difference between a and a' is race or gender – then this is evidence of algorithmic bias.

The interventionist theory of causation builds upon Lewis's counterfactual account, although arguably its roots are more directly traced to foundational work in experimental design (Fisher, 1935) and econometrics (Haavelmo, 1944), as well as more recent research in computer science (Pearl, 2000). The most vocal proponent of interventionism in contemporary philosophy is Woodward (2003; 2008; 2010; 2015), who articulates what he calls a "minimal" theory of explanation as follows (2003, p. 203). Let E stand for some explanandum assigning a particular value y to an output variable Y. Let L be some generalisation relating inputs X to outputs Y. Then explanans M is explanatory with respect to E if and only if the following conditions are met:

(i) M and E are true, or at least approximately so.

(ii) According to L, $Y = y$ under an intervention that sets X to x.

(iii) There is some intervention that changes the value of X from x to x', where $x \neq x'$, with L correctly describing the value y', where $y \neq y'$, that Y would assume under this intervention.

This theory poses no small number of complications that are beyond the scope of this literature review.[5] Nonetheless, it represents a valuable refinement of the counterfactual model with |important implications for iML.

[5] For book-length treatments of the topic, see (Halpern, 2016; Potochnik, 2017; Strevens, 2010; Woodward, 2003). For relevant articles, see, e.g., (Franklin-Hall, 2014; Kinney, 2018; Potochnik, 2015; Weslake, 2010; Woodward & Hitchcock, 2003).

Woodward emphasises that, whereas Lewis aspires to reduce causality to relations of counterfactual dependence, Woodward has no such reductionist ambitions. In fact, he is "skeptical that any reductive account will turn out to be adequate" (2003, p. 20). Just as Hempel sought to avoid any direct mention of causality by tying explanation to law-like regularities that could be defended on empiricist grounds, so Lewis attempts to explain events through counterfactuals that serve as primitives within his account. Both philosophers inherit a Humean scepticism with respect to causality, fearing that anything less would amount to question begging. Woodward objects that Lewis's notion of similarity metrics over worlds requires some prior concept of causality to render accurate judgments on real world cases regardless. Meanwhile, the circularity inherent in his own interventionism, Woodward argues, is more illuminating than vicious. A final distinction of note is that, whereas Lewis is focused exclusively on causal *tokens* (i.e., local explanations), Woodward expands his analysis to include causal *types* (i.e., global explanations). Briggs (2012) and Fine (2012) examine several instances where Lewis's and Woodward's theories diverge, and in all cases favour the interventionist model over the counterfactual account.

Whatever their differences, it is important to acknowledge how both theories improve upon the DN model. Specifically, they address sufficiency objections originally raised by Salmon (1971), who observes that DN-compliant explanations cannot distinguish between relevant and irrelevant regularities. To rephrase the point in slogan form: *correlation does not imply causation*. The mere fact that males who take birth control pills regularly fail to get pregnant does not by itself entail some law of nature. It merely suggests a certain hypothesis that must be tested by disentangling the two universals it conjoins. This is the logic behind randomised control trials (RCTs), the gold standard of causal inference, which attempt to resolve potential confounding effects by assigning treatment conditions uniformly throughout the target population.

For all interventionism's merits, Woodward's account fails to accommodate most cases of interest in iML. To see why, consider that every supervised learning model is by definition a mapping of inputs X to outputs Y. It may be applied to both actual feature vectors (i.e., true observations) or counterfactual feature vectors (i.e., perturbed datapoints). But though this setup meets all of Woodward's criteria, we do not judge some particular prediction to be explained just by pointing to the model parameters L and the output y when L is not itself intelligible. Such an explanation may be maximally accurate, but it is far too complex to be of any use. An opaque supervised learning algorithm cannot provide an explanation of itself. The upshot of this critique is that some further constraints on L are required before we regard its dicta as explanatory. Raw instrumentalism was found lacking in §2.2.3; it fares no better here. Without some understanding of the mechanisms that L instantiates – its operating logic – we

cannot predict outcomes on future points or envision results of individual perturbations. Explanations must run deeper.

§2.3.3 Epistemological Pragmatism

Lurking in the background of this critical exegesis has been a collection of related objections and proposals that could all be said to fly under the banner of *epistemological pragmatism*. With origins in nineteenth century American thought – particularly the works of Peirce (1999), James (1975), and Dewey (1999) – pragmatism has a rich and somewhat controversial history in analytic philosophy. For a good overview, see (Legg & Hookway, 2020). Central to all varieties of pragmatism is the primacy of agents and contexts over ideas and abstractions. Conceptual advances are only valuable insomuch as they are useful. A theory with no practical implications is little more than a formal exercise.

Twentieth century pragmatic accounts of scientific explanation are numerous and varied (see, e.g., Achinstein, 1983; Bromberger, 1966; Scriven, 1962), but perhaps no one crystallises their collective thrust so neatly as van Fraassen:

> The discussion of explanation went wrong at the very beginning when explanation was conceived of as a relation like description: a relation between a theory and a fact. Really, it is a three-term relation between theory, fact, and *context*. No wonder that no single relation between theory and fact ever managed to fit more than a few examples! Being an explanation is essentially relative for an explanation is an *answer*…it is evaluated vis-à-vis a question, which is a request for information. But exactly…what is requested differs from context to context. (1980, p. 156)

This approach marks a radical departure from all previously considered accounts. The DN, counterfactual, and interventionist models may differ in the particulars, but they all share the goal of enumerating some objective criteria deemed necessary and sufficient for successful explanation. The pragmatist, by contrast, rejects this undertaking altogether. Instead she starts from the simple, indisputable observation that explanations do not occur in a vacuum. Rather, they are the product of interactions between epistemic agents with certain beliefs and interests. Ignoring these contingencies does not result in deeper, more general models; it merely vitiates our true target, producing instead some formal theory with little to say about scientific practice.

The contention that good explanations should be flexible with respect to their levels of abstraction – ranging in target from models to systems, and in resolution from local to global, as required – is at least partially motivated by pragmatic considerations. What is simple for some agents may be complex for others, depending on their capabilities, experience, background knowledge, and so on. A theory of explanation must be able to accommodate such variation without artificially imposing a one-size-fits-all solution. Similarly, the objection that Woodward's interventionism is not explanatory beyond a certain threshold of complexity is

essentially a pragmatic one. Perhaps there is some abstract sense in which a neural network's behaviour is only *fully* explained by listing the values of every parameter associated with every node located in every layer of the model – but this is not much help to the defendant who just wants to know why the algorithm has predicted that he is likely to reoffend.

Some of the most important and relevant contemporary work on epistemological pragmatism is due to Floridi (2011a, 2013, 2019), who places *information* at the centre of all philosophical inquiry. Specifically, Floridi is interested in the semantic (rather than Shannon's [1948] syntactic) notion of information, which he defines as well-formed, meaningful, and veridical data (2011a). The ability to move clearly and smoothly between different levels of abstraction is not just a theoretical desideratum for Floridi, but a guiding methodological principle (2008a). There can be no theoretical analysis without first identifying the relevant observables and typed variables, for these both delimit the conceptual space and ground the semantics of all propositions, informative or otherwise. But a level of abstraction is not enough on its own. According to Floridi's (2011b) *correctness theory of truth* (CTT), semantic information can always be polarised into question/answer pairs – but only once we have specified a particular context, level of abstraction, and purpose (collectively labelled "CLP parameters"). These specifications ensure the epistemic relevance of information, which is a prerequisite for successful coordination between agents (Floridi, 2008b). An answer is *true* if and only if it *saturates* the question (verifying and validating it), thereby generating an adequate model of the target system. However, Floridi cautions that "Queries cannot acquire their specific meaning in isolation or independently of CLP parameters" (2011b, p. 155).

These elements all come together in Floridi's (2012) unique model of explanation, which he dubs the *network theory of account* (NTA). NTA builds upon the question/answer format envisioned by van Fraassen above and developed in detail as part of CTT. Since *information* is a strictly weaker concept than *knowledge* – only the latter can possibly admit of positive or negative introspection (Floridi, 2006), which means only the former is free of the otherwise insoluble Gettier problem (Floridi, 2004; Gettier, 1963) – we need some method of upgrading mere information into full-blown knowledge. Floridi's solution is to embed the information in a flow network with certain graph theoretic properties. A source *s* and target *t* are connected by a number of directed edges through which information flows. *t* queries *s* with well-formed questions about some explanandum of interest – being careful to specify the relevant CLP parameters – while *s* sends truthful answers back in reply. Floridi shows that under modest assumptions, such a network will tend to elevate *t*'s information into knowledge. Thus NTA avoids (but does not quite "resolve", since this is impossible) Gettier-type objections through reliabilist pragmatics, thereby providing necessary (but insufficient) conditions for successful explanation.

In a companion paper that explicitly acknowledges an intellectual debt to the American pragmatists, Floridi (2010) considers sceptical objections to his philosophy of information. He frames the problem in a possible worlds semantics, pointing out that Lewis is deliberately vague about how to define or quantify inter-world similarity. Floridi's strategy is to characterise each world by a unique Borel number, some (potentially infinite) string of 1's and 0's representing the answers to a fixed sequence of CLP-indexed yes-or-no questions. Given this setup, the Hamming distance – which simply counts the differences between two Boolean strings of equal length – offers a simple and effective way to measure similarity between worlds. Floridi then exploits the properties of this metric to prove that radical scepticism reduces to an innocuous redundancy, while moderate scepticism is actually beneficial, insomuch as it promotes methodological rigor. This latter argument he attributes to Peirce, who emphasises the social dimension of scientific inquiry, especially the communal commitment to advancing knowledge by attempting to falsify one another's theories.

The upshot of Floridi's analysis is that explanation is essentially a *process* – not a deductive argument or a structural causal model. We may describe the sequence of questions and answers using formal tools, but the explanation itself is no formalism. It is a messy, dynamic, social interaction in which at least two agents iteratively trade information within some particular context, at some particular level of abstraction, and for some particular purpose. This insight is beginning to gain traction within the iML community, as evidenced by a number of recent articles and conference papers explicitly endorsing pragmatic approaches to algorithmic explanation (Miller, 2019; Mittelstadt et al., 2019; Murdoch et al., 2019; Páez, 2019). These lessons are also central both to the critiques advanced in Chapter 4 and the solutions proposed in Chapters 5 and 7.

§2.3.4 Trust and Testing

I noted in Chapter 1 that there are epistemic motivations for the iML project – to validate the performance of algorithms, especially when outputs are unexpected, and to discover new mechanisms when statistical models outperform human experts. In the former case, our goal is to determine the extent to which the algorithm is *reliable* or *trustworthy*; in the latter case, to exploit that reliability to our epistemic advantage.

But how do we determine whether any method, computational or otherwise, is reliable? The question is a familiar one in analytic philosophy. One prominent answer comes from Goldman (1979), who led the vanguard what Williams calls "the reliabilist revolution" (2016) in anglophone epistemology. Goldman's theory is simple (and decidedly pragmatic): a process is reliable if it has a historical track record of cognitive success. Performance over time can be boiled down to what Goldman calls a "truth ratio", i.e. the rate of true judgments among all those attributable to the process in question. High truth ratios are evidence of reliable

methods. In a somewhat different context, Taddeo (2010a, 2010b) defines trustworthiness along similar lines. She describes a model in which rational agents (human or artificial) evaluate one another's historical performance on particular tasks to calculate success ratios. An agent is deemed trustworthy with respect to a given task if and only if its success ratio over time exceeds some threshold value. This is an intuitive, if somewhat idealised account that ignores statistical niceties such as base rates and effect sizes, to say nothing of the variable costs associated with different kinds of errors. (A false negative may be far more dangerous than a false positive in certain medical contexts, for example.) Yet Goldman and Taddeo are right to observe that reliability must be gradually earned and steadily maintained. In the context of ML, a new algorithm is always regarded with scepticism until it proves itself on unseen data. Even then, our confidence in the model is never more than a few mistakes away from being irreparably dashed.

Peirce was perhaps the first to describe this dynamic at play in the natural sciences. Systematic scepticism is elevated to an organising principle by both Merton (1942) and Popper (1934), two towering intellectual figures widely credited with founding the modern disciplines of sociology and philosophy of science, respectively. Merton enumerates four norms that collectively "comprise the ethos of modern science" (p. 270): universalism, communism, disinterestedness, and organised scepticism. Last but not least among these norms, organised scepticism "is both a methodological and an institutional mandate" (p. 277), he writes. The point is developed in considerable detail by Popper (1934), who argues that the demarcation criterion of science – what distinguishes it from all other modes of inquiry – is the empirical falsifiability of the theories it produces. According to his philosophy of falsificationism, science advances knowledge through an iterative procedure of conjectures and refutations with the formal structure of a *modus tollens* inference.

This view, which is closely related to the DN model of §2.3.1, can be explicated as follows. If L denotes a scientific theory, then we must be able to combine it with some initial conditions S to predict some empirical consequence(s) E:

(1) $(S \,\&\, L) \rightarrow E$.

Presuming that S includes all causally relevant information,[6] we may test L by checking whether it is in fact the case that E. If

(2a) $\sim E$,

then we can logically infer (via *modus tollens*) that

(3a) $\sim L$.

If, on the other hand,

[6] This is a nontrivial assumption. According to the Duhem-Quine thesis, Popper's falsificationism fails precisely because it is impossible to design a test that isolates the effects of L. We can always salvage any theory no matter how anomalous the observation(s) E, provided we make sufficient amendments to the conjunct(s) S, e.g. adding auxiliary hypotheses. See (Duhem, 1914; Quine, 1951).

(2b) E,

then we have failed to falsify L and can infer only that

(3b) L has been *corroborated*.

Popper goes to great pains to stress that corroboration is not to be confused with verification, for Hume has definitively shown that this is impossible. To take a famous example,[7] let L be the universal statement "All swans are white", and let $S = \{s_1, \dots, s_n\}$ be a set of n swans. Either at least one swan is not white ($\sim E$), in which case we may definitively reject L; or else they are all white (E), in which case we have not yet rejected L – but neither have we ruled out the possibility that swan s_{n+1} might be black. Thus, Popper reasons that universal statements are asymmetrically decidable – they can be falsified but never verified.

Numerous authors have pointed out major difficulties with Popper's account. See (Hansson, 2017) for a critical discussion. Perhaps most problematic is that strict falsificationism struggles to differentiate between theories that are more or less corroborated by the evidence. In later writings, Popper (1963) would go on to adopt Tarski's (1935) formulation of the correspondence theory of truth, which provides a method for ranking theories by verisimilitude. Even Popper's most ardent acolytes generally judge this work unfavourably (Thornton, 2019), and Popper himself would come to disavow the undertaking (Popper, 1972). Other attempts to address the issue include Carnap's inductive logic (1950, 1952) and various Bayesian epistemologies (for a good overview, see [Talbott, 2016]). However, the most convincing and sophisticated resolution, I contend, is found in the error-statistical philosophy of Mayo (1996; 2018). Her notion of *severe testing* brings clarity and rigour to Popper's falsificationism while avoiding many of the traps that inevitably ensnare probabilist logics.

Mayo – who, coincidentally, also cites Peirce as a major source of philosophical inspiration – offers both weak and strong versions of her severity principle. I will focus on the strong form, which states that "We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C" (2018, p. 14). The basic intuition behind this principle is that not all tests are created equal. To adapt Mayo's own example, suppose I weigh myself on digital and analogue scales prior to an extended vacation in Argentina. I also weigh a copy of Floridi's book *The Philosophy of Information*, which clocks in at exactly one pound. Upon my return from Argentina, where I have consumed prodigious volumes of beef, wine, and potatoes, I am disheartened (but not entirely surprised) to discover that both scales report a weight gain of approximately ten pounds on my part, while stubbornly insisting that Floridi's monograph remains exactly one

[7] Technically, this example should be formalised in first-order logic to quantify predicates over sets. I stick with propositional logic here for consistency with previous sections and ease of presentation. The example is sufficiently simple and familiar that I doubt the ambiguity will lead to any confusion.

pound. In this case, we can safely reject the hypothesis that I have lost weight in my travels. Moreover, we should and do have greater confidence in this conclusion under the scenario above than we would if the scales had been mutually inconsistent, or if Floridi's book had somehow gained a proportional amount of weight. The hypothesis that I gained weight has been not just corroborated, but *severely tested*. More generally, Mayo argues that our justification for believing in any non-logical proposition – be it a hypothesis, explanation, or mere observation statement – is a function not just of the proposition itself, but of how severely it has been tested.

The formal details of this proposal will be examined more closely in Chapter 4. Briefly, we evaluate the severity of a test by observing how likely it is to detect all and only true effects, i.e. by computing its expected rate of false positives and false negatives as a function of effect size. This represents an advance over Goldman's truth ratios, which do not differentiate between easy and hard judgments, and Taddeo's success ratios, which do not distinguish between simple and challenging tasks. Mayo adopts the Neyman-Pearson hypothesis testing framework to derive optimal decision procedures for a wide range of parametric examples, thereby reframing model selection as a statistical inference problem.

§2.4 Conclusion

I opened this chapter with reference to the Luddite Rebellion of 1811. This historical event embodies many of the themes that are central not just to iML but to technological advancement more generally. Though they are sometimes misrepresented as prototypical sociotechnical sceptics, the Luddites are perhaps more fairly remembered as a disadvantaged, disenfranchised group actively seeking redress. Destroying textile machinery was a (technological) means toward a (social) end – namely, greater wage security and worker's rights. Purely technological solutions – e.g., fully automating the textile industry (as sociotechnical dogmatists might have liked) or banning such automation outright (as sociotechnical sceptics might have preferred) – fundamentally miss the point. A pragmatic alternative more likely to satisfy the Luddites would be to redistribute the surplus value created by increased automation in an equitable fashion through progressive tax policy. I do not claim that such an undertaking is easy or straightforward – striking the ideal balance between free market dynamism and social welfare protections is arguably the defining struggle of modern capitalism – but there is little to gain by ignoring the trade-off altogether. Similarly, ML algorithms often force us to evaluate our commitment to competing imperatives such as fairness, transparency, and accuracy. The solution, then as now, is to roll up our sleeves and do the work – no matter how unsavoury the challenge.

The parable of the Luddites holds lessons for the philosophy of explanation as well. The end users of iML are *people* – messy, imperfect creatures with particular preferences,

beliefs, and abilities. Models of explanation that operate at logical, statistical, or physical levels under the monist assumptions of naïve realism are of no value to agents who cannot use or understand them. Efforts to assuage the Luddites through economic projections of GDP growth as a function of increased automation in nineteenth century England would likely be met with blank stares – or worse. Thankfully, twentieth century philosophy and statistics provide a range of tools for rigorously specifying levels of abstraction, encoding agentive preferences, and quantifying uncertainty. Though these methods are not yet widespread in iML, I will argue in subsequent chapters that they can and should be.

The Statistics of Interpretable Machine Learning

§3 Abstract

Statisticians, especially those who do applied work in the natural and social sciences, have long been interested in understanding model parameters and predictions. However, the distinct subfield of interpretable machine learning (iML), with its focus on general purpose explanations at varying degrees of resolution, is a much more recent development. In this chapter, I go beyond mere taxonomies and survey some of the most influential iML proposals at length. My goal in so doing is not to undertake the Sisyphean task of constructing a comprehensive review – numerous articles and at least one monograph have attempted this feat, only to be rendered obsolete within a few months – but rather to catalogue the assumptions and methods that characterise certain directions of research in iML, specifically those that I shall critique in Part II and build upon in Part III. This overview is an essential step to ground the ensuing analysis, introducing a range of technical concepts and notation that will be put to work in the following chapters. The examples from this chapter give a sense of the methodological breadth of iML techniques, which draw on literature in statistics, computer science, and game theory in the effort to make ML models and predictions more interpretable.

§3.1 Introduction

Explainability is a relatively young subfield in ML, yet already the area is booming with active research on multiple fronts. In just the last few years – roughly concurrent to the dramatic uptick in academic publications on the topic (see Fig. 3.1) – iML has gone mainstream. It has been the subject of lengthy articles in both *The New Yorker* (Mukherjee, 2017) and *The New York Times Magazine* (Kuang, 2017), as well as numerous TED talks (Hasani, 2019; Holzinger, 2019) and other public lectures that have collectively gathered hundreds of thousands of views online (Doshi-Velez, 2017; Hall, 2018; Lundberg, 2019). Meanwhile, major tech companies have begun to take notice. Google,[1] Microsoft,[2] and IBM[3] have all released open-source algorithmic explainability toolkits. Cloud computing services from Amazon,[4] Microsoft,[5] and Google[6] now include native implementations of various model interpretability methods.

[1] See https://pair-code.github.io/what-if-tool/.
[2] See https://github.com/interpretml/interpret.
[3] See http://aix360.mybluemix.net/.
[4] See https://aws.amazon.com/blogs/machine-learning/ml-explainability-with-amazon-sagemaker-debugger/.
[5] See https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability.
[6] See https://cloud.google.com/explainable-ai/.

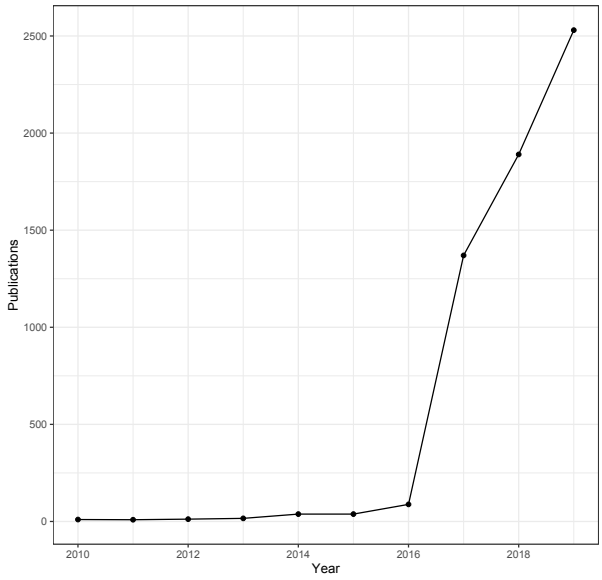

Figure 3.1. Surge in research interest. The plot depicts the number of academic publications with "interpretable machine learning" or "explainable artificial intelligence" in the title, abstract, or keywords published between 2010 and 2019. Source: Google Scholar.

In this chapter, I review some of the most prominent and promising proposals to have emerged from this burgeoning discourse. To go back to the typology outlined in Chapter 1 (§1.2), I will explore both intrinsic and post-hoc methods operating at global and local resolutions. However, I will generally focus on model-agnostic approaches to the exclusion of model-specific alternatives, as the latter are simply too large in number and narrow in scope to warrant comprehensive coverage in this literature review. For good survey articles, see (Adadi & Berrada, 2018; Arrieta et al., 2020; Guidotti et al., 2018; Murdoch et al., 2019); for book-length works, see (Molnar, 2020; Samek et al., 2019).

A number of commentators contend that the technical goals of iML are underspecified. Whereas formal criteria for algorithmic fairness abound (see §2.1.3), explainability is generally harder to quantify or optimise. Lipton identifies a vague and heterogeneous collection of concepts that jointly form "the mythos of model interpretability" (2018). He argues that without greater clarity on what exactly interpretability means and why it is important, efforts to build more explainable models will continue to be dragged down by implicit presuppositions that make it difficult or impossible to compare proposals. Doshi-Velez & Kim (2017) likewise emphasise the lack of consensus regarding the definition or operationalisation of algorithmic explainability. They provide a thorough taxonomy of iML, ultimately advocating a pragmatic focus on empirically demonstrating a model's ability to promote greater understanding among human subjects. In the three years since their influential preprint was posted to *arXiv*, the research agenda outlined by Doshi-Velez & Kim has attracted enormous interest and helped inspire a wide array of new statistical methodologies.

The remainder of this chapter is structured as follows. I consider local linear approximators in §3.2, focusing especially on the popular iML tools LIME and SHAP. A critical

discussion of rule lists follows in §3.3, examining both global and local recursive partitioning schemes. I review a number of case-based methods in §3.4, including algorithms that identify or construct prototypes and counterfactuals. I analyse feature importance measures in §3.5, including marginal and conditional metrics. I conclude in §3.6 with a preview of the role these concepts play in the proceeding chapters. In an effort to maintain consistency, I will adapt notation throughout to conform to my own formalisms.

§3.2 Local Linear Approximators

Many people believe, rightly or wrongly, that linear models are somehow inherently interpretable (Lipton, 2018). Several features of this function class probably contribute to this dogma, but I will focus on three assumptions in particular: that effects are *additive*, *monotonic*, and *constant*.[7] The additivity assumption means that predictions can be expressed as a weighted sum of input features. This is a convenient decomposition that allows users to quickly scan model parameters and potentially even grasp the relative importance of variables (presuming they have been properly standardised and are not too numerous). The monotonicity constraint means that the association between a predictor and an outcome is either positive, negative, or zero. No further variation is permitted. For instance, say we want to evaluate the impact of alcohol consumption on sociability via linear regression, which assigns a coefficient of 1.5 to the predictor variable. Then, according to this model, more alcohol *always* equates with greater sociability. The regression does not care whether you just arrived at a party or just woke up from a hangover. Moreover, because linear effects are constant, the increase in sociability is directly proportional to the increase in alcohol intake. For every unit of alcohol imbibed, sociability goes up exactly 1.5 units, on average.

Of course, these traits that make linear models so easy to interpret are precisely what can make them so inaccurate in practice. Many systems of interest are not additive, monotonic, or constant, and imposing these assumptions where they do not apply results not only in poor predictive performance but in model parameters with no clear interpretation. To continue with the example above, say we add a binary predictor to our model that indicates whether or not data subjects have a history of alcoholism. In this case, sociability is not a simple additive function of the inputs; instead, we need an interaction term that assigns one (presumably positive) coefficient to alcohol intake for non-alcoholics, and another (presumably negative) coefficient for alcoholics. Linear models can adapt to this setting when interactions or mixed effects are made explicit, but cannot detect such subtleties on their own the way many ML algorithms do. Moreover, the monotonicity constraint is unreasonable here. A bit of alcohol may tend to make people chattier, but there are quantities of alcohol that would

[7] In generalised linear models (GLMs), it should be noted that these properties hold only in a transformed space defined by the link function. For instance, they are true in logit space (not probability space) for logistic regression.

make even the world's most gregarious individual decidedly unsociable. In any event, the relationship is surely not constant. The effect of another drink on one's sociability is largely determined by how many drinks one has already had.

In full awareness of these considerations, many researchers in iML have promoted a solution I shall term *local linear approximation*. The idea is not to reconstruct the global behaviour of some complex regressor or classifier with a linear model – for this is likely impossible – but rather to use linear techniques merely to approximate the regression surface or decision boundary near a datapoint of interest. If you zoom in close enough to any point on a differentiable function, you will eventually find a tangent that can be expressed as a linear combination of input variables (see Fig. 3.2). By analysing the formula for this approximation, we may gain some intuition for the target function's behaviour in a particular region of the feature space.

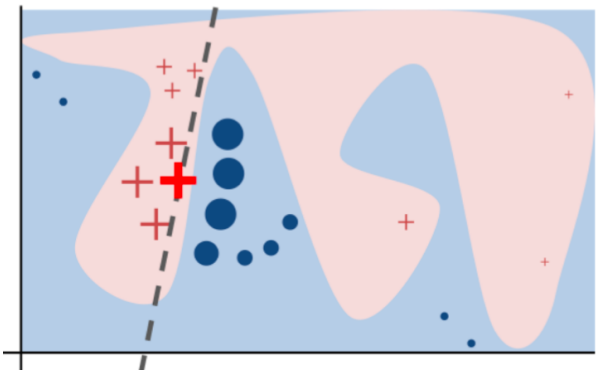

Figure 3.2. A complex decision boundary (the pink blob/blue background) separates red crosses from blue circles. This function cannot be well-approximated by a linear model. But the boundary near the large red cross is roughly linear, as indicated by the dashed line. From (Ribeiro et al., 2016, p. 1138).

This is the logic behind the locally interpretable model explanations (LIME) algorithm, an early and influential iML tool (Ribeiro et al., 2016). Specifically citing issues of algorithmic trust as a motivation for their work, the authors propose a novel explainability technique based on random sampling and regularised linear models. Given an input datapoint $\boldsymbol{x}$ and some target function f, LIME simulates a synthetic dataset by perturbing the coordinates of $\boldsymbol{x}$, thereby creating a collection of counterfactual observations $\boldsymbol{x}'$. By querying f at each point, LIME generates a training dataset of $(\boldsymbol{x}', y' = f(\boldsymbol{x}'))$ pairs. The goal now is to learn an explanation model $g: \boldsymbol{X}' \to Y'$. Points are weighted by their distance from the original target $\boldsymbol{x}$ using some appropriate similarity kernel k – the authors recommend the radial basis $\exp(-d(\boldsymbol{x}, \boldsymbol{x}')^2/\sigma^2)$, where d is a user-selected distance function and σ is the kernel bandwidth – and a sparse linear regression is fit to the data via weighted least squares (WLS). The lasso penalty that Ribeiro et al. apply to compute g is a sparsity-inducing regularisation technique that automatically assigns zero weight to uninformative predictors through a tuneable Lagrangian penalty λ on the L_1-norm of model coefficients (Tibshirani, 1996). Rather than selecting λ via cross-validation,

as one typically would in a lasso prediction problem (Friedman, Hastie, & Tibshirani, 2010), the parameter is selected to ensure at most m nonzero coefficients, where m is a user-selected hyperparameter. Ribeiro et al. demonstrate LIME's utility on text classification and image recognition problems, empirically validating their method through a human subject study in which participants were able to correctly identify which of two models was superior based on explanations extracted by LIME. The algorithm has been implemented in a popular Python library, available through the Python package index (PyPI).

A number of other local linear approximators debuted around the same time as LIME, each relying on different assumptions and optimised for different use cases. Examples include quantitative input influence (Datta, Sen, & Zick, 2016); layer-wise relevance propagation (Bach et al., 2015); DeepLIFT (Shrikumar et al., 2017); and Shapley sampling values (Štrumbelj & Kononenko, 2014). In an award-winning NeurIPS paper, Lundberg & Lee (2017) show that these methods are all formally equivalent in the limit, up to some variation in their choice of kernel. The authors advocate for a particular kernel inspired by Shapley values, a foundational concept in game theory originally derived to solve *the attribution problem*, which asks how to fairly distribute surplus across a coalition of players in cooperative games (Shapley, 1953). It can be shown that Shapley values are the unique solution to the attribution problem satisfying certain desirable properties (see below). In this reframing, players are replaced by input features and Shapley values measure their contribution to a given prediction. Directly computing classical Shapley values is NP-hard, however numerous approximations have been proposed (Sundararajan & Najmi, 2019). Lundberg & Lee are not the first to use this framework to explain model predictions, but their iML algorithm, SHAP, is especially efficient and user-friendly. Model-specific variants have been optimised for deep neural networks and tree-based ensembles (Lundberg et al., 2020), while a model-agnostic version is freely available through PyPI and distributed with all of the explainability toolkits mentioned above.

Lundberg & Lee (2017) formulate the explainability problem a bit differently than Ribeiro et al. (2016). Let f be the target function and g a corresponding explanation model. An input point $\boldsymbol{x} \in \mathbb{R}^p$ is associated with a simplified input $\boldsymbol{x}' \in \{0,1\}^p$ through a mapping function, $\boldsymbol{x} = h_x(\boldsymbol{x}')$. The goal is then to ensure that $g(\boldsymbol{z}') \approx f\big(h_x(\boldsymbol{z}')\big)$ whenever $\boldsymbol{z}' \approx \boldsymbol{x}'$, subject to certain constraints on the explanation model g. Specifically, Lundberg & Lee identify three desirable properties:

(1) **Local Accuracy**

$$f(\boldsymbol{x}) = g(\boldsymbol{x}') = \phi_0 + \sum_{j=1}^{p} \phi_j x_j'$$

The explanation model matches the original model when $\boldsymbol{x} = h_x(\boldsymbol{x}')$, where $\phi_0 = f\big(h_x(\mathbf{0})\big)$ represents the model output with all simplified inputs set to zero.

(2) **Missingness**

$$x_j' = 0 \;\Rightarrow\; \phi_j = 0$$

Features with simplified values of zero have no attributed impact.

(3) **Consistency**

Let $f_x(\boldsymbol{z}') = f\big(h_x(\boldsymbol{z}')\big)$ and $\boldsymbol{z}'\backslash j$ denote setting $z_j' = 0$. For any two models f and f', if

$$f_x'(\boldsymbol{z}') - f_x'(\boldsymbol{z}'\backslash j) \geq f_x(\boldsymbol{z}') - f_x(\boldsymbol{z}'\backslash j)$$

for all inputs $\boldsymbol{z}' \in \{0,1\}^p$, then $\phi_j(f', \boldsymbol{x}) \geq \phi_j(f, \boldsymbol{x})$.

Adapting Shapley's (1953) original theorem, Lundberg & Lee (2017) show that only one possible linear function g satisfies these properties, with coefficients given by:

$$\phi_j(f, \boldsymbol{x}) = \sum_{\boldsymbol{z}' \subseteq \boldsymbol{x}'} \frac{m!\,(p - m - 1)!}{p!} [f_x(\boldsymbol{z}') - f_x(\boldsymbol{z}'\backslash j)]$$

where m is the number of nonzero entries in $\boldsymbol{z}'$, and $\boldsymbol{z}' \subseteq \boldsymbol{x}'$ represents all $\boldsymbol{z}'$ vectors where the nonzero entries are a subset of the nonzero entries in $\boldsymbol{x}'$. This result implies a so-called "Shapley kernel", which can be substituted for Ribeiro et al.'s exponential kernel to recover Shapley values through WLS regression.[8] For details, see Theorem 2 in (Lundberg & Lee, 2017).

Local linear approximators offer a fast and principled method for generating algorithmic explanations. However, they are inherently limited in several respects. First, as noted above, linear regression methods rely on strong assumptions that are often violated in practice. The restriction to local settings ameliorates these concerns somewhat but does not remove them altogether. Both LIME and SHAP are bound to produce unstable estimates for highly nonlinear regions of the model space. The best linear approximation can vary wildly depending on the range of application and/or associated kernel weights. It is unclear if and how classical methods for quantifying uncertainty in linear parameters may apply in this setting. Neither iML algorithm offers any method for evaluating the quality of the underlying approximation or its probable scope of applicability.

A second issue with these methods is that they provide no option for users to specify a contrast class of interest. The default behaviour of both LIME and SHAP is to explain why and how an outcome deviates from the mean response for the entire dataset, real or simulated. In many contexts, this makes sense. For instance, if a patient receives a rare and unexpected

[8] This relies on a strong but not uncommon assumption in the local linear approximation literature, namely that features are independent. See Equation 11 in (Lundberg & Lee, 2017), as well the critical discussion in (Sundararajan & Najmi, 2019). This topic will be revisited in §3.5, as well as Chapters 4 and 7.

diagnosis, then she may want to know what differentiates her from the majority of patients. However, it seems strange to suggest, as these algorithms implicitly do, that "normal" predictions are somehow inexplicable. There is nothing confusing or improper about someone wondering, for instance, why they received an average credit score instead of a better-than-average one. Yet in their current form, neither LIME nor SHAP can accommodate such inquiries. To address these issues, we need more flexible iML tools.

§3.3 Rule Lists

A growing body of psychological research suggests that humans are especially adept at generating and interpreting explanations in the form of *rule lists* – i.e., sequences of if-then statements (Lage et al., 2018). This accords with the privileged position of material implication in propositional logic, where $\rightarrow$ is typically regarded as a primitive relation, along with conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$). These logical connectives form a functionally complete class, capable of expressing all possible Boolean operations.

In statistical contexts, rule lists are generally learned through some process of recursive partitioning. For instance, the pioneering classification and regression tree (CART) algorithm (Breiman et al., 1984) predicts outcomes by finding the optimal split c in a continuous predictor X such that error is minimised by segregating those samples with X-values greater than or equal to c from those with X-values less than c. A collection of such rules can be visually depicted as a tree, with branches corresponding to split points. Trees violate all of the aforementioned assumptions of linear models. They naturally detect interaction effects – the depth of a tree (i.e., number of recursive partitions) corresponds to the degree of interactions the model can capture – and easily adapt to nonmonotonic and even discontinuous functions. Computing optimal decision trees is generally NP-complete (Hyafil & Rivest, 1976), but CART uses greedy heuristics that typically work well in practice. Because individual trees tend to be unstable predictors, they are frequently combined through ensemble methods such as bagging (Breiman, 2001a), in which predictions are averaged across trees trained on random bootstrap samples, and boosting (Friedman, 2001), in which predictions are summed over a series of trees, each sequentially optimised to improve upon the last. The resulting algorithms are extremely fast and flexible, putting recursive partitioning at the core of some of the most popular and powerful techniques in all of supervised learning (Biau & Scornet, 2016; Chen & Guestrin, 2016).

While combining basis functions tends to improve predictions, it unfortunately makes it difficult if not impossible to extract individual rules for better model interpretation. However, some regularisation schemes have been developed to post-process complex learning forests for precisely this purpose. For instance, Friedman & Popescu (2008) propose the RuleFit algorithm, which mines a collection of Boolean variables by extracting splits from a

gradient boosted forest. These engineered features are then combined with the original predictors in a lasso regression, producing a sparse linear combination of splits and inputs. Nalenz & Villani (2018) develop a similar procedure using a Bayesian horseshoe prior (Carvalho, Polson, & Scott, 2010) instead of an L_1 penalty to encourage shrinkage. They also add splits extracted from a random forest with those learned via gradient boosting to promote greater diversity.

if hemiplegia **and** age $>$ 60 **then** *stroke risk* 58.9% (53.8%–63.8%)
else if cerebrovascular disorder **then** *stroke risk* 47.8% (44.8%–50.7%)
else if transient ischaemic attack **then** *stroke risk* 23.8% (19.5%–28.4%)
else if occlusion and stenosis of carotid artery without infarction **then** *stroke risk* 15.8% (12.2%–19.6%)
else if altered state of consciousness **and** age $>$ 60 **then** *stroke risk* 16.0% (12.2%–20.2%)
else if age $\leq$ 70 **then** *stroke risk* 4.6% (3.9%–5.4%)
else *stroke risk* 8.7% (7.9%–9.6%)

Figure 3.3. Decision list for determining 1-year stroke risk following diagnosis of atrial fibrillation from patient medical history. Risk is given by the posterior mean, with 95% credible interval in parentheses. From (Letham et al., 2015, p. 1361.)

Another strand of research in this area has focussed on *falling* rule lists, which create monotonically ordered decision trees such that the probability of the outcome $Y = 1$ strictly decreases as one moves down the list. These models were originally designed for medical contexts, where doctors must evaluate patients quickly and accurately. For instance, Letham et al. (2015) design a Bayesian rule list to predict stroke risk, resulting in a model that outperforms leading clinical diagnostic methods while being small enough to fit on an index card (see Fig. 3.3). Falling rule lists can be very challenging to compute – see the note above about NP-completeness – and subsequent work has largely focussed on efficient optimisation strategies. Specifically, researchers have developed fast branch-and-bound techniques to prune the search space and reduce training time (Chen & Rudin, 2018; Yang, Rudin, & Seltzer, 2017), culminating in several tree-learning methods that are provably optimal under some restrictions on the input data (Angelino et al., 2018; Hu, Rudin, & Seltzer, 2019). For instance, Fig. 3.4 depicts the output of one such algorithm on the ProPublica dataset from Chapter 1. This simple tree outperforms COMPAS in predicting two-year recidivism and is provably independent of race.

if $(age = 18 - 20)$ **and** $(sex = male)$ **then predict** yes
else if $(age = 21 - 23)$ **and** $(priors = 2 - 3)$ **then predict** yes
else if $(priors > 3)$ **then predict** yes
else predict no

Figure 3.4. Output of the certifiably optimal rule list (CORELS) algorithm. This rule list predicts two-year recidivism in the ProPublica dataset. From (Angelino et al., 2018, p. 2).

More customisable solutions are proposed by Lakkaraju et al. (2016, 2017), who implement a number of methods for computing interpretable decision sets simultaneously optimised for accuracy and sparsity. Their model understanding through subspace explanations (MUSE) algorithm allows users to specify a fixed number of features through which to explain the behaviour of an underlying target function, effectively modulating between global and local resolutions with user-specified granularity. The resulting objective function is non-normal, non-negative, non-monotone, submodular, and constrained by matroids – a class of budgeted coverage problems known to be NP-hard (Khuller, Moss, & Naor, 1999). However, approximate optimality can be guaranteed under mild assumptions, which Lakkaraju et al. exploit to compute efficient and interpretable decision sets that perform favourably against alternative rule list approaches in a number of user studies.

For all the advantages of these fast and occasionally optimal algorithms, rule lists remain prohibitively expensive to compute on data with more than a few dozen variables. But recall that our goal in iML is often not to learn a globally explainable model, but just to understand particular algorithmic prediction(s). Locally interpretable decision trees are not nearly as common as their global counterparts, but there have been some recent advances in this direction. Guidotti et al. (2018) introduce local rule-based explanations (LORE), which simulate a balanced dataset of cases using a genetic algorithm designed to sample heavily from points near the decision boundary. A decision tree g is then fit to the synthetic dataset, with special emphasis on both the input point of interest and the nearest counterfactual cases on the opposite side of the boundary. Explanations extracted from g then take a conjunctive form, providing short rule lists to explain both why $f(\boldsymbol{x}) = 1$ and why $f(\boldsymbol{x}) \neq 0$. Sokol & Flach (2020) introduce LIMEtree, a rule list variant of the aforementioned *locus classicus* of explainability methods, which allows users to interrogate particular predictions through a series of "What-if?" questions about possible perturbations of feature values. The method comes with local fidelity guarantees and is more adaptable than its linear forebear.

More recently, the authors of LIME have proposed a follow up method called *anchors* (Ribeiro, Singh, & Guestrin, 2018). Anchors are specifically designed to address the scoping problem raised in §3.2. Given an explanandum $f(\boldsymbol{x})$, the goal is to find a set of Boolean conditions A (the eponymous anchor) such that $A(\boldsymbol{x}) = 1$ and

$$\mathbb{E}_{\mathcal{D}_{\boldsymbol{x}}(\boldsymbol{z}|A)}\big[\mathbb{I}\big(f(\boldsymbol{x}) = f(\boldsymbol{z})\big)\big] \geq \tau,$$

where the expectation is taken with respect to a conditional perturbation distribution $\mathcal{D}_{\boldsymbol{x}}(\cdot\,|A)$, which represents a density centred at $\boldsymbol{x}$ where the conditions in A hold. $\mathbb{I}(\cdot)$ denotes the indicator function and τ a tuneable threshold parameter that Ribeiro et al. (2018) call *precision*. Once τ is fixed, the goal is to maximise *coverage*, formally defined as $\mathbb{E}_{\mathcal{D}_{\boldsymbol{x}}(\boldsymbol{z})}[A(\boldsymbol{z}) = 1]$, i.e. the proportion of datapoints to which the anchor applies. This generalises the notion of rule lists to include both global and local explanations, as the former can simply be expressed as anchors

with high (ideally unit) coverage. The authors reframe recursive partitioning as a reinforcement learning problem, combining graph search with a multi-armed bandit algorithm to compute anchors. This iML approach is original and rigorous. It is a rare and welcome departure from standard work in this literature, where authors hardly ever bother to quantify the uncertainty of explanations or evaluate expected error rates (see Chapter 4). That said, anchors face some major hurdles in practice. First, continuous predictors often need to be discretised to avoid low-coverage anchors. Second, results depend heavily on tuning parameters buried in the method's subroutines. Finally, anchors do not scale well with data dimensionality.

Setting aside computational concerns, rule lists face a very different set of statistical challenges than linear models. Whereas the latter start from the assumption that all effects must be monotonic and constant, decision trees struggle to detect smooth, linear functions. Recursive partitioning naturally produces jagged regression surfaces that can only be smoothed out by increasing complexity, usually by incorporating more basis functions (see Fig. 3.5). The resulting models are generally not differentiable, which means parameters cannot be learned using popular optimisation techniques like gradient descent or the Newton-Raphson algorithm. An exception to this rule is posed by so-called "soft trees" (Kontschieder et al., 2015), which treat splits probabilistically rather than categorically. By parametrising the probability of splitting a given direction as a logistic function of a linear combination of inputs – a common formulation in neural networks – model parameters can be learned through back-propagation. Some have even proposed this as a method for distilling deep networks into shallow trees for greater interpretability (Frosst & Hinton, 2017).

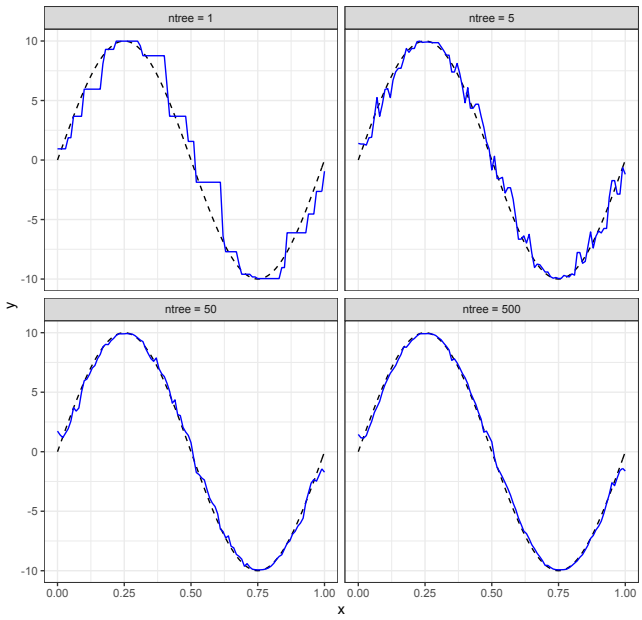

Figure 3.5. Predictions from a random forest regression converging on a sine function as the number of trees in the ensemble increases. From (Watson, 2019, p. 429).

§3.4 Case-Based Methods

Whatever the differences between linear models and rule lists, they share a certain formal similarity in the context of iML, in that both attempt to approximate some complex functional relationship with an alternative method considered more readily interpretable. The proposals considered in this subsection take an entirely different approach. Rather than building a new model g to explain a target model f, example-based methods opt for a strategy that might be summed up as "show, don't tell." They offer explanations in the form of particular cases intended either to exemplify a given class or highlight subtle, decisive differences between samples.

Some of the earliest work on exemplary methods was devoted to *prototypes*. A prototypical example of a given class might be thought of as a sort of Platonic ideal, a central theme upon which all other instances are merely variations. In statistical learning, the concept is perhaps most familiar in the context of clustering, where the classic k-means algorithm (Forgy, 1965) identifies k separate centroids, one for each cluster. A centroid is a kind of prototype where coordinates for each variable are set to the group-wide mean. This may create impossible datapoints, however, for example when some variables are categorical or integer-valued. An alternative method more appropriate in such settings is the k-medoids algorithm (Kaufman & Rousseeuw, 1990), which replaces centroids with medoids, where coordinates for each variable are set to the group-wide median. By construction, these coordinates must exist somewhere in the training data. The prototype selection (PS) method of Bien & Tibshirani (2011) is conceptually similar to these approaches. Unlike in unsupervised learning, PS uses labels to partition the data. The prototype for class $Y = 1$ can then be understood in geometric terms as whichever sample is closest to other $Y = 1$ cases and farthest from all $Y \neq 1$ cases on average. Thus the output of PS is neither a centroid (containing coordinates found nowhere in the data) nor a medoid (containing coordinates cobbled together from various samples), but rather some actual training case that best represents its label.

Figure 3.6. Learned prototypes and criticisms from the ImageNet dataset (two types of dog breeds). From (Kim et al., 2016, p. 8).

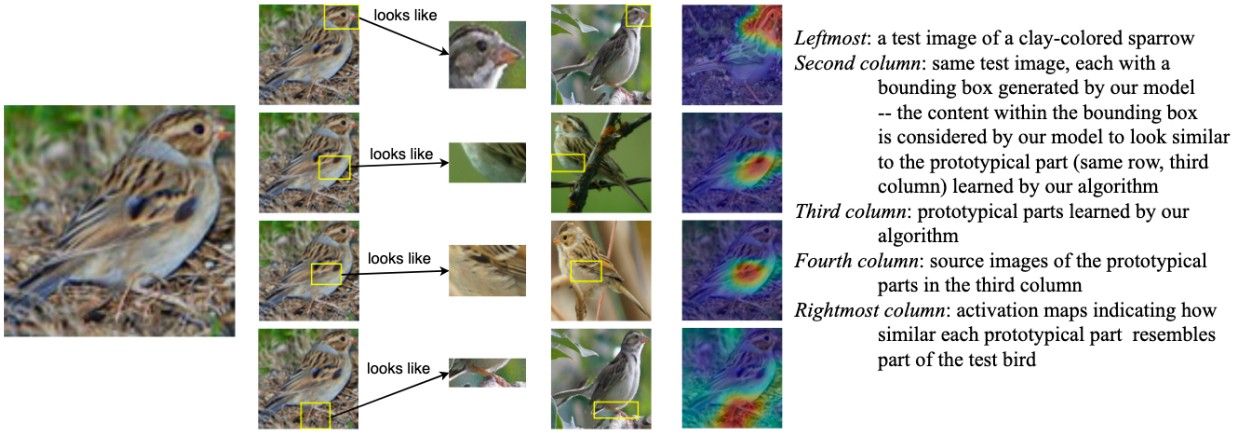

Figure 3.7. Example output from the ProtoPNet model on a clay-coloured sparrow. From (Chen et al., 2019, p. 2).

Another approach in this area is the Bayesian case model (BCM), a generative algorithm for identifying prototypes (Kim, Rudin, & Shah, 2014). BCM treats observations as the result of a mixture of k classes, where k is supplied by the user. The method works by learning the subspace of features most strongly associated with a particular label. The sample that maximises the posterior probability of class membership conditional on the cluster subspace is the prototype. This inference relies on a complex set of hierarchical priors and collapsed Gibbs sampling procedures. Kim, Khanna, & Koyejo (2016) modify the original BCM procedure by finding not just a prototype for each class, but a *criticism* as well – i.e., the datapoint least well represented by its prototype. Their MMD-critic algorithm embeds the data in a reproducing kernel Hilbert space (RKHS), where distances between distributions can be efficiently computed with the maximum mean discrepancy (MMD) statistic (Gretton et al., 2007). Prototypes and criticisms are found via greedy search in the RKHS, by maximising and minimising, respectively, the distance between their distributions and the rest of the data. (See Fig. 3.6.)

One especially promising work in this area is the ProtoPNet, which Chen et al. (2019) implement using state of the art image classification techniques based on deep convolutional neural networks. Their model architecture includes a so-called "prototype layer" in between the convolutional layers and the fully connected layer, tasked with learning prototypes of each class label based on high-level features (e.g., classifying bird species by the shape of beaks and wings) and computing distances between such prototypes and input image segments. The resulting model can be said to reason by analogy – the title of the paper is "*This* looks like *that*" – which, the authors persuasively argue, is more intelligible to humans than complex optimisation procedures over some high-dimensional parameter space. (See Fig. 3.7.)

More recent work in case-based reasoning has focused on *counterfactuals*, as opposed to prototypes. I briefly described the work of Wachter et al. (2018) in §2.2.2, where I noted its conceptual overlap with the philosophy of Lewis. To recapitulate, they suggest explaining the algorithmic decision $f(\boldsymbol{x}) = y$ by finding a set of nearest counterfactual neighbours – say, all

simulated datapoints $\boldsymbol{x}'$ within an ε-ball of $\boldsymbol{x}$ such that $f(\boldsymbol{x}') = y' \neq y$. Differences between these $\boldsymbol{x}'$ and $\boldsymbol{x}$ are explanatory to the extent that they indicate a minimal set of perturbations sufficient to change a prediction by some prespecified threshold. Building on seminal results in deep learning, Wachter et al. use generative adversarial networks (GANs) to optimise for $\boldsymbol{x}'$. GANs were originally developed in machine vision, where Goodfellow et al. (2014) first observed that minor perturbations of input pixels can result in new synthetic images indistinguishable from the original – at least to the human eye – yet capable of fooling otherwise high-performing machine vision algorithms into strange and consistent misclassifications. Using a similar approach, Wachter et al. propose the following objective function:

$$\lambda(f(\boldsymbol{x}') - y')^2 + d(\boldsymbol{x}, \boldsymbol{x}')$$

where d is a distance measure between points and λ controls the trade-off between squared error and distance. The authors recommend a standardised L_1 metric for d, as this tends to promote sparse solutions that minimise the number of axes along which $\boldsymbol{x}$ and $\boldsymbol{x}'$ are likely to differ. A simplified version of this algorithm that avoids optimisation altogether is implemented in Google's What-If Tool (Wexler et al., 2020), which selects counterfactual cases by finding the observed case(s) $\boldsymbol{x}'$ nearest to $\boldsymbol{x}$ such that $f(\boldsymbol{x}') = y'$.

The counterfactual method is attractive in several respects. Not only does it elegantly tie together disparate disciplinary influences, but it does so in a manner that is sophisticated and novel. Wachter et al.'s (2018) paper has shifted the iML discourse away from the approximation methods chronicled in §§3.2–3.3 and towards case-based reasoning with greater success than the prototype methods that preceded it (Artelt & Hammer, 2019). It has spawned an especially active and promising subfield known as "algorithmic recourse," which studies automatic methods for explaining unfavourable outcomes and recommending actions to alter them. See (Karimi et al., 2020) for a recent survey.

However, counterfactual explanations are not without their difficulties. First, the method is liable to produce a wide range of plausible explanations $\boldsymbol{x}'$ within some ε-ball of $\boldsymbol{x}$. The authors consider this a virtue, since different explanations can provide different details about a given outcome. But without some further method for evaluating the quality of these candidate $\boldsymbol{x}'$, this pluralism may become confusing or worse – it could give bad actors an easy way to avoid accountability should some explanations reflect more favourably than others upon algorithmic decisions. A second, arguably more important concern regards the unrestricted nature of adversarial attacks, which offer no guidance on how to limit the search space to genuine possibilities. The problem is especially acute in high dimensions, where data are often presumed to lie on a low-dimensional manifold from which the GAN is likely to deviate. This could potentially be mitigated through pre-processing steps that reduce the dimensionality of the input data, but this introduces whole new sources of potential confusion and error. Subsequent work has focused on restricting the search space to counterfactuals that are

"actionable" (Ustun, Spangher, & Liu, 2019) or "coherent" (Russell, 2019) using mixed integer programming; or else those that are "feasible" in some causal sense (Karimi, Schölkopf, & Valera, 2020; Mahajan, Tan, & Sharma, 2019), which is crucial to guarantee actionable recourse (Barocas, Selbst, & Raghavan, 2020). However, these methods only work at the expense of parametric constraints on f and/or strong assumptions about the data generating process.

§3.5 Variable Importance

A final class of iML techniques I will consider in this literature review are variable importance (VI) measures. A helpful way to think about these methods is as different sorts of *interventions*, in the technical sense of the term common in computer science and statistics (Imbens & Rubin, 2015; Pearl, 2000). Let $\boldsymbol{Z} = (\boldsymbol{X}, Y)$ denote a dataset of $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ pairs drawn i.i.d. from some fixed but unknown joint probability distribution, $\mathbb{P}(\boldsymbol{Z}) = \mathbb{P}(\boldsymbol{X}, Y)$. If our goal is to estimate the importance of some feature subset $\boldsymbol{X}^S \subset (X_1, \dots, X_p)$ on learner $f: \boldsymbol{X} \to Y$, then we can simply compare model performance before and after various perturbations of $\boldsymbol{X}^S$. Let $\tilde{\boldsymbol{z}}_i$ denote the i^{th} datapoint in $\boldsymbol{Z}$ following an intervention on subset $\boldsymbol{X}^S$ (e.g., permutation or deletion). Let $L(f, \boldsymbol{z}_i)$ denote the loss function for model f, evaluated at point $\boldsymbol{z}_i$. Then we may define a random variable

$$\Delta_i = L(f, \tilde{\boldsymbol{z}}_i) - L(f, \boldsymbol{z}_i)$$

as the sample-wise difference in loss between perturbed and original data. When $\boldsymbol{X}^S$ has no descendants in the causal graph, this measure of local VI is formally equivalent to an individual treatment effect (ITE). The global measure can be estimated by taking the mean of this variable across the complete dataset, which in this causal reframing could be considered an average treatment effect (ATE).

This high-level overview glosses over many important details discussed below, such as how particular interventions are designed, whether $\mathbb{P}(\boldsymbol{Z})$ factorises into a (known) structural model such that causal effects can be propagated throughout the associated graph, and whether f is held fixed or retrained following the perturbation of $\boldsymbol{X}^S$. But hopefully this brief exposition provides some intuition for what unifies these heterogeneous VI methods. Whatever their assumptions or strategies, they all represent different attempts to quantify just how much predictive information is encoded in some feature subset – typically, just a single variable X_j.

VI methods can be categorised by three dichotomies: local/global, model-specific/model-agnostic, and marginal/conditional. The first contrast is formally outlined above. The second is a familiar distinction in iML. As noted previously, I will restrict my focus here to model-agnostic methods. The third dichotomy is arguably the most fundamental. To

evaluate response variable Y's marginal dependence on predictor X_j, we test against the following hypothesis:

$$H_0^m: X_j \perp Y, \boldsymbol{X}_{-j},$$

where $\boldsymbol{X}_{-j}$ denotes a set of covariates. A measure of conditional dependence, on the other hand, tests against a different null hypothesis:

$$H_0^c: X_j \perp Y | \boldsymbol{X}_{-j}.$$

Observe that the former entails the latter, as conditional independence is just one possible form of marginal independence. Since H_0^c is more restrictive, we may find instances in which it holds but H_0^m does not. Specifically, this will be the case whenever X_j's marginal VI is high due to its association with $\boldsymbol{X}_{-j}$ rather than Y. This is why measures of marginal importance tend to favour correlated predictors. Often, however, our goal is to determine whether X_j adds any *new* information – in other words, whether Y is dependent on X_j even after conditioning on $\boldsymbol{X}_{-j}$. This becomes especially important when the assumption of feature independence is violated.

A number of popular marginal VI methods are what I will call permute and predict (PaP) procedures. The first and most famous of these is Breiman's (2001a) permutation importance, originally designed for random forests. The approach has since been generalised to other function classes, including most recently by Fisher, Rudin, & Dominici (2019), who introduce a number of "reliance" statistics that can be computed for any supervised learning algorithm. The basic idea with PaP procedures is to perturb some feature subset $\boldsymbol{X}^S$ by permuting its rows and observe the impact on performance for some fixed model f. Fisher et al. study the behaviour of this VI measure in individual models and diverse Rashomon sets, establishing uniform error bounds with the theory of U-statistics and drawing some unexpected connections between their measure and several causal estimands of interest.

The permutation approach has also led to a number of popular graphical tools that can help visualise complex associations in data. For instance, Friedman's (2001) partial dependence plots (PDPs) are widely used to evaluate the shape of relationships between predictors and outcomes. The partial dependence function is closely related to Breiman and Fisher et al.'s permutation measures, and can be formally defined as follows:

$$\mathrm{PD}(\boldsymbol{x}_i^S) = \mathbb{E}_{\boldsymbol{X}^R}\left[f(\boldsymbol{x}_i^S, \boldsymbol{X}^R)\right] = \int f(\boldsymbol{x}_i^S, \boldsymbol{X}^R)\, \mathrm{d}\mathbb{P}(\boldsymbol{X}^R)$$

In other words, we integrate predicted values for f over the marginal distribution of covariates $\boldsymbol{X}^R = \boldsymbol{X} \backslash \boldsymbol{X}^S$ while holding the point $\boldsymbol{x}_i^S$ constant. Zhao & Hastie (2019) observe that this is formally identical to Pearl's (2000) famous backdoor adjustment for estimating causal effects in graphical models, thereby enabling a causal interpretation of PD when the complementary

subset $\boldsymbol{X}^R$ satisfies the backdoor criterion.[9] Plotting empirical PD-estimates against $\boldsymbol{X}^S$ provides a visual summary of the (potentially causal) effect of the feature subset on model predictions $f(\boldsymbol{X}^S, \boldsymbol{X}^R)$. PDPs can also be adapted to visualise feature interactions and visually check for additive effects (Friedman, 2001; Friedman & Popescu, 2008). Goldstein et al. (2015) extend the method by decomposing partial dependence into n unique curves, each calculated by ranging over the empirical distribution of $\boldsymbol{X}^S$ while holding $\boldsymbol{x}_i^R$ constant. The mean of these curves at any given point is the corresponding PD-value. Assuming once again that the backdoor criterion holds, this helps visualise not just average but individual treatment effects. (See Fig. 3.8.)

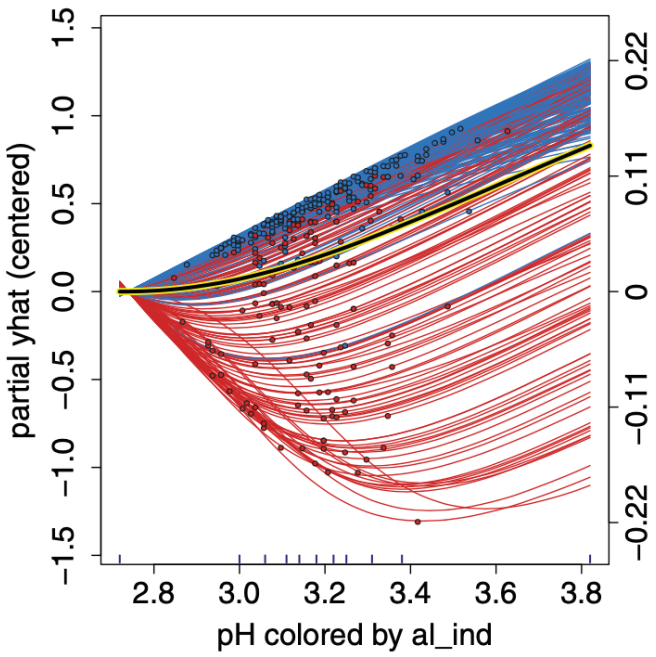

Figure 3.8. An individual conditional expectation (ICE) plot depicting the partial dependence of wine ratings on pH level, with colour indicating whether the alcohol content is high (blue) or low (red). The black curve depicts Friedman's (2001) partial dependence function. From (Goldstein et al., 2015, p. 58).

Despite the success and conceptual appeal of PaP approaches, they face several practical and theoretical obstacles. First, the computational expense of permutations can be enormous for large datasets. Sampling procedures can offset the cost of such operations, but also introduce new sources of error and compromise any theoretical guarantees that hold in the full sample setting. More troubling, numerous commentators have pointed out that these methods tend to be badly biased in favour of correlated predictors (Gregorutti, Michel, & Saint-Pierre, 2015; Nicodemus et al., 2010; Tolo̧şi & Lengauer, 2011). Hooker & Mentch (2019) explain the issue as one of extrapolation. Permuting features that are strongly associated with

[9] A set of variables Z satisfies the backdoor criterion relative to an ordered pair of variables (X_i, X_j) in a directed acyclic graph G if (i) no node in Z is a descendant of X_i; and (ii) Z blocks every path between X_i and X_j that contains an arrow into X_i. See (Pearl, 2000).

covariates results in improbable or even impossible samples very far from any in the algorithm's training data. For example, they note that a PaP procedure evaluating the importance of pregnancy status in a model f that also includes sex would force f to predict outcomes for pregnant males as often as pregnant females. Should f perform poorly on such datapoints – as we might expect – then pregnancy will receive high VI, even if it is independent of the response Y. Without further analysis, PaP methods cannot distinguish between variables that are truly predictive and those that merely appear so due to their association with covariates.

Broadly, there are two (model-agnostic) ways to evaluate VI that avoid the errors of PaP methods. One is to design a more targeted intervention on $\boldsymbol{X}^S$ that breaks its tie with the outcome Y while preserving the general covariance structure of the predictors. This requires some model of $\mathbb{P}(\boldsymbol{X}^S|\boldsymbol{X}^R)$. Options in this direction include structural causal models (Pearl, 2000; Peters, Janzing, & Schölkopf, 2017), which can rule out impossible coordinate combinations; semiparametric inference (van der Laan, 2006; van der Laan & Rose, 2011; van der Laan & Rose, 2018), which maximises likelihood for a single estimand while regressing out the effect of nuisance parameters; conditional permutation schemes (Berrett et al., 2019; Doran et al., 2014; Strobl et al., 2008), which constrain the set of allowable permutations in a data-adaptive manner; or knockoff approaches (Barber & Candès, 2015; Candès et al., 2018), which will be examined more closely in Chapter 6. A second way to estimate conditional VI is to simply relearn the model after permuting or deleting $\boldsymbol{X}^S$. That is, we train null and alternative models from the same function class $\mathcal{F}$ on perturbed and original data, respectively:

$$f_1 = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L\left(f, \tilde{\boldsymbol{Z}}\right) \quad f_0 = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} L(f, \boldsymbol{Z})$$

We then redefine the Δ variable accordingly:

$$\Delta_i = L(f_1, \tilde{\boldsymbol{z}}_i) - L(f_0, \boldsymbol{z}_i)$$

and use it to compute unbiased conditional estimates of global and local VI. In the parametric setting, the remove-and-relearn approach corresponds to the well-known likelihood ratio test in classical statistics (Lehmann & Romano, 2005). A more general leave one covariate out (LOCO) procedure is proposed by Lei et al. (2018) and studied further by Rinaldo, Wasserman, & G'Sell (2019). Mentch & Hooker (2016) advocate the permute-and-relearn approach, which they point out has the advantage of maintaining the same dimensionality for perturbed and original data. This helps ensure fair comparison of f_0 and f_1, especially when testing large feature subsets.

Of course, there are drawbacks to all of these methods. For example, it is not always clear how or if one can model $\mathbb{P}(\boldsymbol{X}^S|\boldsymbol{X}^R)$ while maintaining independence from Y without assuming a great deal of domain knowledge or incurring an impractical computational burden. These are major hurdles for the structural modelling framework and targeted maximum

likelihood estimation. Conditional permutation methods are limited to low-dimensional settings or particular model architectures, and generally require binning strategies to discretise continuous predictors. The original knockoff method relies on strong parametric assumptions. More general alternatives have been proposed (see, e.g., Bates et al., 2020; Berrett et al., 2019; Romano, Sesia, & Candès, 2019), but their applicability remains a matter of some dispute. Meanwhile, relearning f for each input variable – let alone the powerset of such variables – can be infeasible with even moderately large datasets and complex learning algorithms.

I propose a number of strategies for tackling these problems in Chapters 6 and 7, where I develop new model-agnostic VI measures for generating global and local explanations, respectively. Before doing so, however, I will first establish a conceptual critique of iML (Chapter 4) and advance a formal theory of explanation games (Chapter 5).

Part II: *Praxis*

Conceptual Challenges for Interpretable Machine Learning

§4 Abstract

As machine learning has gradually entered into ever more sectors of public and private life, there has been a growing demand for algorithmic explainability. How can we make the predictions of complex statistical models more intelligible to end users? A subdiscipline of computer science known as interpretable machine learning (iML) has emerged to address this urgent question. Numerous influential methods have been proposed, from local linear approximations to rule lists and counterfactuals. In this chapter, I highlight three conceptual challenges that are largely overlooked by authors in this area. I argue that the vast majority of iML algorithms are plagued by (1) ambiguity with respect to their true target; (2) a disregard for error rates and severe testing; and (3) an emphasis on product over process. Each point is developed at length, drawing on relevant debates in epistemology and philosophy of science. Examples and counterexamples from iML are considered, demonstrating how failure to acknowledge these problems can result in counterintuitive and potentially misleading explanations. Without greater care for the conceptual foundations of iML, future work in this area is doomed to repeat the same mistakes.

§4.1 Introduction

Machine learning (ML) is increasingly ubiquitous in modern society. Complex learning algorithms are widely deployed in private industries like finance (Heaton, Polson, & Witte, 2017) and insurance (Lin et al., 2017), as well as public services such as healthcare (Topol, 2019b) and education (Peters, 2018). Their prevalence is largely driven by results. ML models outperform humans not just at strategy games like chess (Silver et al., 2018) and Starcraft (Vinyals et al., 2019), but at important scientific tasks like chemical synthesis (Segler, Preuss, & Waller, 2018) and tumour diagnosis (McKinney et al., 2020).

High-performance algorithms are often *opaque*, in the sense that it is difficult or impossible for humans to understand the internal logic behind individual predictions. This raises fundamental issues of trust. How can we be sure a model is right when we have no idea why it predicts the values it does? Accuracy on previous cases may suggest reliability (Goldman, 1979), but epistemologists are well aware that a good track record on past cases is no guarantee of future success (Hume, 1739). Training datasets are often biased or unrepresentative in myriad ways that are not reflected by in-sample performance metrics. This can lead to discriminatory predictions with potentially disastrous consequences when algorithms are deployed in high-stakes settings like healthcare (Obermeyer et al., 2019) and criminal justice (Angwin et al., 2016). European regulators, sensitive to these concerns, have begun

introducing explainability guidelines into data protection law (GDPR, 2018), although the proper interpretation of the relevant texts remains a matter of some dispute (Selbst & Powles, 2017; Wachter et al., 2017).

While interpreting models is by no means a new concern in computer science and statistics, it is only in the last few years that a formal subfield has emerged to address the issues surrounding algorithmic opacity. I shall refer to this subdiscipline as interpretable machine learning (iML), also sometimes called explainable artificial intelligence (xAI), or more generally "explainability". iML comprises a diverse collection of technical approaches intended to render statistical predictions more intelligible to humans. I offered a brief typology of these methods in §1.2, and a more in-depth review of particularly influential methods in Chapter 3. For general surveys, see (Guidotti et al., 2018; Molnar, 2020; Murdoch et al., 2019). My focus in this chapter is primarily on model-agnostic post-hoc methods, which attempt to explain the outputs of some underlying target function without making any assumptions about its form. Such explanations may be global (spanning the entire feature space) or local (applying only to some subregion of the feature space). Both types are considered here.

The last few years have seen considerable advances in iML, several of which will be examined in detail below. Despite this progress, I contend that the field has yet to overcome or even properly acknowledge certain fundamental conceptual obstacles. In this chapter, I highlight three in particular:

(1) *Ambiguous fidelity*. Everyone agrees that algorithmic explanations must be faithful – but to what exactly? The target model or the data generating process? Failure to appreciate the difference has led to confusing and unproductive debates.

(2) *Error rate control*. The vast majority of iML methods do not even bother to quantify expected error rates. This makes it impossible to subject algorithmic explanations to severe tests, as is required of any scientific hypothesis.

(3) *Process vs. Product*. Current approaches overwhelmingly treat explanations as static deliverables, computed once and for all. In fact, successful explanations are more of a process than a product. They require dynamic, iterative refinements between multiple agents.

A number of other conceptual challenges of iML have already garnered much attention in the literature, especially those pertaining to subtle distinctions between explanations, interpretations, and understanding (Krishnan, 2019; Páez, 2019; Zednik, 2019); as well as the purported accuracy-explainability trade-off (Rudin, 2019; Zerilli et al., 2019). I have little to add to those debates here, which I believe have been well argued by numerous other authors. The challenges I highlight in this chapter, by contrast, are woefully under-examined despite their obvious methodological import. To make my case, I shall draw upon copious literature from epistemology and philosophy of science and demonstrate the relevance of these issues through

a number of real and hypothetical examples. Together, they point toward a singular conclusion – that despite undeniable technical advances, the conceptual foundations of iML remain shaky at best. Fortunately, there are glimmers of hope to be found in this burgeoning discourse. I consider exceptions to each trend that collectively suggest a promising horizon of possibility for iML research.

The remainder of this chapter is structured as follows. I review relevant background material in §4.2, framing the iML task as a causal inference problem. In §4.3, I distinguish between two oft-conflated notions of explanatory fidelity, revealing the apparent contradiction to be a simple confusion between complementary levels of abstraction. In §4.4, I draw on error-statistical considerations to argue that popular iML methods fail to meet minimal severity criteria, making it difficult to judge between competing explanations. I defend a dialogic account of explanation in §4.5, arguing that satisfactory solutions must include some degree of user interaction and feedback. I conclude in §4.6 with some reflections on the role of philosophy in iML.

§4.2 Background

In this section, I provide necessary background on iML methods, as well as some formal details on empirical risk minimisation and structural causal models. Building on Woodward's causal interventionism and Pearl's *do*-calculus, I frame the iML project as a certain sort of causal inquiry. This perspective helps elucidate the conceptual challenges that follow.

§4.2.1 All iML is Causal

Say some high-performance supervised learner f has been trained on copious volumes of biomedical data, and diagnoses Jack with rare disease y. His general practitioner, Dr. Jill, is as perplexed as he is by this unexpected diagnosis. Jack shows no outward symptoms of y and does not match the typical disease profile. Treatment for y is aggressive and potentially dangerous, so Jack wants to be certain before he proceeds. When Jack and Dr. Jill try to find out why f made this prediction, they receive a curt reply from the software company that licenses the technology, informing them that they should accept the diagnosis because f is very accurate. Most commentators would likely agree that this answer is unsatisfactory. But how exactly should we improve upon it? What is the proper form of explanation in this case?

I shall argue that what Jack and Dr. Jill seek here is a *causal* account of why f made the particular prediction it did. Following the interventionist tradition, I regard an explanation as causal insomuch as it identifies a set of variables whose values are sufficient to bring about the outcome in question, as well as an alternative set of values for those variables sufficient to alter the outcome in some prespecified way. Woodward (2003, p. 203) formalises these

criteria with three conditions. On his view, the model $\mathcal{M}$ is causal with respect to inputs X and outputs Y if and only if:

(i) The generalisations described by $\mathcal{M}$ are accurate, or at least approximately so, as are the observations $Y = y$ and $X = x$.

(ii) According to $\mathcal{M}$, $Y = y$ under an intervention that sets $X = x$.

(iii) There exists some possible intervention that sets $X = x'$ (where $x \neq x'$), with $\mathcal{M}$ correctly describing the value $Y = y'$ (where $y \neq y'$) that Y would assume under the intervention.

In Jack's case, we may be able to do this empirically by finding some other patient who is medically similar to Jack but receives a different diagnosis. Alternatively, we could query the model f directly using synthetic data in which we perturb Jack's input features until we achieve the desired outcome. If, for instance, we devise an input vector x' identical to Jack's input x except along one dimension – say, decreased heartrate – and the model does not diagnose this hypothetical datapoint with rare disease y, then we may justifiably conclude that heartrate is causally responsible for the original prediction. This kind of counterfactual explanation constitutes at least one viable explanans for the target explanandum.

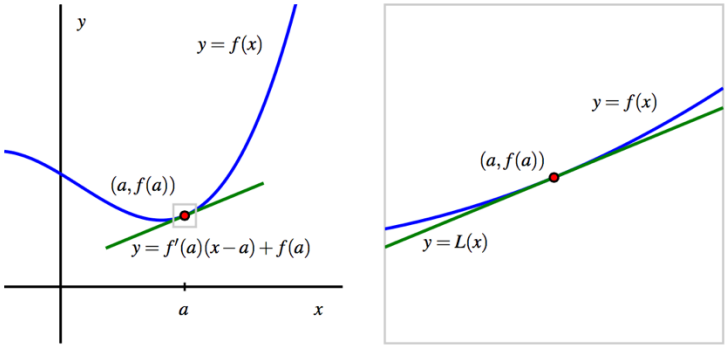

Figure 4.1. A nonlinear function $f(x)$ (blue curve) is approximated by a linear function (green curve) at the point a. Computing such tangents is the basic idea behind local linear approximators like LIME and SHAP.

Current iML approaches can be roughly grouped into three classes: feature attribution methods, case-based explanations, and rule lists. The latter category poses considerable computational challenges for large datasets, which may explain why the first two are generally more popular. Local linear approximators, a kind of feature attribution technique, are the most widely used approach in iML (Bhatt et al., 2020). Notable instance include local interpretable model-agnostic explanations, aka LIME (Ribeiro et al., 2016); and Shapley additive explanations, aka SHAP (Lundberg & Lee, 2017). Specifics vary, but the goal with these methods is essentially the same – to compute the linear combination of input features that best explains the decision boundary or regression surface near an input point of interest (see Fig. 4.1). Counterfactual explanations (Wachter et al., 2018), which account for predictions by generating similar datapoints with different predicted outcomes – a kind of synthetic matching

technique – are another common approach. Variants of LIME, SHAP, and counterfactual explanations have recently been implemented in open-source algorithmic explainability toolkits distributed by major tech firms such as Google,[1] Microsoft,[2] and IBM.[3] When I speak of "popular iML methods", I have these algorithms in mind.

No matter one's methodological approach, the central aim of iML is always, more or less explicitly, to answer questions of the form:

> *Q. Why did model f predict outcome $\hat{y}_i$ as opposed to alternative $y_i' \neq \hat{y}_i$ for input vector $\boldsymbol{x}_i$?*

A global explanation answers Q for each $i \in [n]$, while local explanations limit themselves to individual samples. Successful answers must be *causal* in Woodward's interventionist sense. This is perhaps most obviously true in the case of rule lists, which specify sufficient conditions (i.e., causal rules) for certain sorts of model predictions. An explanatory rule list for Jack's diagnosis may say something like, "If heartrate is decreased, then predict y'." The causal connection is similarly straightforward for feature attribution methods, which attempt to quantify the predictive impact of particular variables. In Jack's case, it may be that heartrate receives the largest variable importance score because it has the greatest causal effect on model outcomes. Interestingly, the creators of the counterfactual explanation algorithm explicitly motivate their work with reference to Lewis's theory of causation (1973). According to this view, we causally explain Jack's prediction by appealing to the nearest possible world in which he receives a different diagnosis.

If the causal underpinnings of iML are not always clear, perhaps this is because most authors in this area are steeped in a tradition of statistics and computer science that has historically prioritised prediction over explanation (Breiman, 2001b; Shmueli, 2010). I will briefly formalise the distinction between supervised learning and causal modelling to preempt any potential confusion and ground the following discussion in established theory.

§4.2.2 ERM and SCMs

A supervised learning algorithm is a method for predicting outcomes $\boldsymbol{Y} \in \mathbb{R}^k$ based on inputs $\boldsymbol{X} \in \mathbb{R}^p$ with minimal error.[4] This requires a training dataset of input/output pairs $\boldsymbol{z}_i = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where each sample $\boldsymbol{z}_i$ represents a draw from some fixed but unknown distribution $\mathbb{P}(\boldsymbol{Z})$. An algorithm is associated with a function space $\mathcal{F}$, and the goal is to find the model

[1] See https://pair-code.github.io/what-if-tool/.
[2] See https://github.com/interpretml/interpret.
[3] See http://aix360.mybluemix.net/.
[4] In the classification setting, we one-hot encode the k-class variable Y such that $Y \in \{0,1\}^k$ and $\forall i, \sum_{j=1}^k y_{ij} = 1$. While regressions typically assume a univariate target, more high-dimensional outputs are possible; this is known as multitask learning (Caruana, 1997). For simplicity's sake, I will generally assume that $k = 1$. Except where specified otherwise, all the analysis in this chapter applies equally to regression and classification problems, as well as cases where $k > 1$.

$f \in \mathcal{F}$ that minimises some predetermined loss function $L(f, \boldsymbol{Z})$, which quantifies the distance between observed outcomes $f(\boldsymbol{X}) = \hat{Y}$ and expected outcomes Y. Common examples include mean squared error for regression and cross-entropy for classification. The expected value of the loss is the risk, and empirical risk minimisation (ERM) is the learning strategy whereby we select whichever model attains the minimal loss within a given function class. ERM is provably consistent (i.e., guaranteed to converge upon the true risk) under two key assumptions (Vapnik & Chervonenkis, 1971): (1) samples are independently and identically distributed (i.i.d.); and (2) the function class $\mathcal{F}$ is of bounded complexity.[5]

The ERM approach provides the theoretical basis for all modern ML techniques, including support vector machines (Schölkopf & Smola, 2017), boosting (Schapire & Freund, 2012), and deep learning (Goodfellow et al., 2016). As noted in §4.1, these algorithms have proven incredibly effective at predicting outcomes for many complex tasks. However, critics argue that ERM ignores important structural dependencies between predictors, effectively elevating correlation over causation. The problem is especially acute when variables are confounded. To cite a famous example, researchers trained a neural network to help triage pneumonia patients at Mount Sinai hospital in New York (Caruana et al., 2015). The model was an excellent predictor, easily outperforming all competitors. Upon close inspection, however, the researchers were surprised to discover that the algorithm assigned low probability of death to pneumonia patients with a history of asthma, a well-known risk factor for those under acute pulmonary distress. The unexpected association was no mistake. Because asthmatics suffering from pneumonia are known to be high risk, doctors quickly send them to the intensive care unit (ICU) for monitoring. The extra attention they receive in the ICU lowers their overall probability of death. This confounding signal obscures a more complex causal picture that ERM is fundamentally incapable of capturing on its own.

Examples like this highlight the importance of interpretable explanations for high-stakes ML predictions like those commonly found in clinical medicine (Watson et al., 2019). They also demonstrate the dangers of relying on ERM when the i.i.d. assumption fails. If we were to deploy the pneumonia triage algorithm in a new hospital where doctors are not already predisposed to send asthma patients to the ICU – perhaps a hospital where doctors rely exclusively on a high-performance ML model to decide how to treat patients – then the aforementioned convergence guarantees no longer apply, and empirical risk may substantially underestimate the true generalisation error. In light of these considerations, a number of prominent authors have advocated for an explicitly causal approach to statistical learning (Imbens & Rubin, 2015; Pearl, 2000; Peters, Janzing, & Schölkopf, 2017; Spirtes, Glymour, &

[5] Exact proposals for bounding the complexity of $\mathcal{F}$ vary. In this chapter, I am more concerned with assumption (1) than (2), and so will have little to say about VC dimension, Rademacher complexity, or other learning theoretic measures. For details, see (Kearns & Vazirani, 1994; Schölkopf, 2003; Vapnik, 1998).

Scheines, 2000; van der Laan & Rose, 2011). The basic strategy can be elucidated through the formalism of structural causal models (SCMs). An SCM $\mathcal{M}$ is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, F, \mathbb{P}(\boldsymbol{u}) \rangle$, where:

- $\boldsymbol{U}$ is a set of exogenous variables, i.e. unobserved background conditions;
- $\boldsymbol{V}$ is a set of endogenous variables, i.e. observed features;
- F is a set of deterministic functions such that each $f_j \in F$ maps some subset of $\boldsymbol{U} \cup \boldsymbol{V}_{-j}$ to V_j, and such that $\bigcup F$ is a function from $\boldsymbol{U}$ to $\boldsymbol{V}$; and
- $\mathbb{P}(\boldsymbol{u})$ is a probability distribution over $\boldsymbol{U}$.

An SCM can be visually depicted as a directed graph, where nodes are variables and edges denote causal relationships (see Fig. 4.2). A fully specified $\mathcal{M}$ provides a map from the joint distribution of background conditions to a joint distribution of observables, $\mathcal{M}: \mathbb{P}(\boldsymbol{u}) \rightarrow \mathbb{P}(\boldsymbol{v})$.

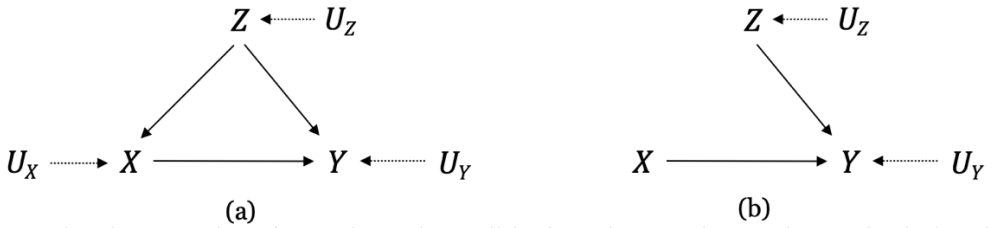

Figure 4.2. Simple examples of causal graphs. Solid edges denote observed causal relationships, dotted edges unobserved. (a) A model with confounding between variables *X* and *Y*. (b) The same model after intervening on *X*, thereby eliminating all incoming causal effects.

With SCMs, we can express the effects not just of conditioning on variables, but of *intervening* on them. In graphical terms, an intervention on a variable effectively deletes all incoming edges, resulting in a modified model $\mathcal{M}'$. Interventions are formally expressed by Pearl's (2000) *do*-operator. The interventional distribution $\mathbb{P}\big(Y \big| do(X = 1)\big)$ may deviate considerably from the observational distribution $\mathbb{P}(Y|X = 1)$ within a given model $\mathcal{M}$. For instance, if all and only men ($Z = 1$) take some drug ($X = 1$), then health outcomes Y could be the result of sex or treatment, since $\mathbb{P}(Y|X = 1) = \mathbb{P}(Y|Z = 1)$. However, if we randomly assign treatment to patients independent of their sex, then we may get a very different value for $\mathbb{P}\big(Y \big| do(X = 1)\big)$, especially if there is a confounding effect between sex and outcomes, for example if men are more likely than women to respond to treatment. Only by breaking the association between X and Z can we disentangle the relevant from the spurious effects. This is the motivating logic behind randomised control trials (RCTs), which are widely used by scientists and regulatory agencies to establish treatment efficacy.[6] The *do*-calculus provides a provably complete set of rules for reasoning about interventions (Huang & Valtorta, 2006), including criteria for deciding whether and how causal effects can be estimated from observational data.[7]

[6] The supremacy of RCTs for causal inference has not gone unchallenged. See (Deaton & Cartwright, 2018; Kaptchuk, 2001; Pearl, 2018; Worrall, 2007).
[7] It should be noted that the *do*-calculus is only complete with respect to atomic interventions. For recent work on identifiability results for stochastic and conditional interventions, see (Correa & Bareinboim, 2020).

Though the models we seek to explain with iML tools are typically ERM algorithms, the causal nature of this undertaking arguably demands an SCM approach. The mismatch between these two modelling strategies sets the stage for a number of conceptual problems, which are examined in greater detail below.

§4.3 Ambiguous Fidelity

One obvious desideratum for any iML tool is *accuracy*. We want explanations that are *true*, or at least *probably approximately correct*, to use Valiant's memorable phrase (1984). This accords with the first of Woodward's three criteria cited above. In this section, I argue that this uncontroversial goal is under-specified. Though the problem emerges for any iML approach, I will focus here on a longstanding dispute between proponents of marginal and conditional variable importance measures, two popular kinds of feature attribution methods. I show that the debate between these two camps is dissolved (rather than resolved) as soon as we recognise that each kind of measure is faithful to a different target. The question of which should be preferred for a given iML task cannot be answered without taking into account pragmatic information regarding the context, level of abstraction, and purpose of the underlying inquiry.

§4.3.1 Systems and Models

I have argued that iML's fundamental question Q poses a certain sort of causal problem. However, it is important to note how Q differs from more familiar problems in the natural and social sciences. Toward that end, I briefly review three well-known and interrelated challenges that complicate efforts to infer and quantify causal effects.

The problem of induction. Although commonly associated with Hume (1739, 1748) in the anglophone tradition, inductive scepticism goes back at least as far as the second century Pyrrhonist Sextus Empiricus (Bett, 2012; Floridi, 2002), and was independently developed by eleventh century Persian philosopher and Sufi mystic al-Ghazali (al-Ghazali, 2000; Griffel, 2020). The basic idea, familiar to philosophy undergraduates the world over, is that inference from particular observations to universal generalisations relies on the assumption (often implicit) of natural uniformity. For example, the leap from "All hitherto observed swans have been white" to "All swans are white" presumes that the regularity in question, corroborated in some bounded region of space and time, holds everywhere (and potentially always). Sceptics argue that such a premise cannot be justified by reason or experience. Deductive methods are insufficient because exceptions to the rule are logically possible; inductive reasoning, on the other hand, may not be invoked in support of induction without vicious circularity. This poses major challenges for any account of causality that seeks to go beyond mere correlations, since, according to the inductive sceptic, deeper structures are unobservable in principle. "One event

follows another," Hume writes, "but we never can observe any tie between them. They seem *conjoined*, but never *connected*" (1748, §7, Part II).

Possible confounders. Suppose you are not terribly concerned with the epistemological foundations of induction. Perhaps you view the problem as theoretical and overblown, clearly the work of philosophers who spend more time in armchairs than laboratories. You are a no-nonsense physician charged with testing the efficacy of a new drug on patients with some disease, and there can be no serious doubt that a causal relationship exists between treatment and response. Or can there? Reichenbach (1956) conjectures that any (sufficiently persistent) statistical dependency between two variables X and Y can only be explained by one of three circumstances: either (i) X causes Y; (ii) Y causes X; or (iii) some third variable Z causes both X and Y. In the latter case, we say that Z is a *confounder*, since it induces a spurious correlation between X and Y that tempts us into misclassifying an instance of (iii) as an instance of (i) or (ii). For example, demographic factors may spell disaster for your study if it turns out that treatment and control groups differ substantially along clinically relevant variables such as age or sex, as per the example above. This is why we rely on RCTs to mitigate the potentially deleterious effects of confounding variables. The problem is that we can never be certain we have controlled for all possible confounders, because we are limited by unavoidable constraints on our budget, instruments, and/or imagination. A version of this objection lies at the root of a radical scepticism first articulated by Duhem (1914) and later defended by Quine (1951), who argue that scientific theories are always underdetermined by evidence. Any observation can be made consistent with any theory, so long as we are willing to add sufficient auxiliary hypotheses (e.g., make exceptions or add latent confounders). If the association between two variables cannot be sufficiently well isolated, then it is impossible to estimate the causal effect of one on another.

Counterfactuals. RCTs may be the gold standard of causal inference, but there are fundamental limits to what they allow us to infer. This is because RCTs are designed to reveal *average* rather than *individual* treatment effects. Typically, we are not interested in how a large, random group of controls fares against a large, random group of treated samples, except insomuch as the evidence may help guide decisions with respect to particular cases. The trouble is that no single individual can simultaneously enter into both treatment arms – a person either does or does not undergo some intervention, and whichever path they take automatically forecloses the alternative. This is what Holland (1986) calls "the fundamental problem of causal inference": that individual treatment effects require some form of counterfactual reasoning. Lewis (1973) elevates this challenge into a unifying principle, reducing all causality to relations of counterfactual dependence. A similar idea motivates Rubin's (1974) potential outcomes framework, in which estimation is framed as a missing data problem. The goal is to impute unrealised alternatives based on available data and use these to infer causal

effects. The method is popular but not entirely without controversy. The metaphysical flavour of counterfactual events grates against the empiricist ethos. Among analytic philosophers, Quine (1960, 1980) is perhaps the most forceful in his opposition, arguing that all talk of so-called "possible worlds" is conceptually confused, not to mention ontologically profligate. Dawid (2000) makes a statistical case against counterfactuals, reasoning that they are misleading and unnecessary given careful Bayesian decision analysis.

Upon considering these challenges, it may appear that iML researchers are in luck. After all, when tracing causal effects from inputs to outputs in a supervised learning algorithm, *not one of these obstacles applies*. Assuming that the target model is:

(a) static (i.e., not retraining on the fly);

(b) deterministic (i.e., predictions do not involve random sampling); and

(c) accessible (i.e., researchers can query f at little or no cost),

then the task of causal inference should be remarkably straightforward. We can just dial each predictor up and down at will, one at a time or in conjunction, to observe the resulting behaviour. In this scenario, the future will always resemble the past, there are no possible confounders, and counterfactuals can be directly observed with the push of a button.

However, matters are not so simple – and not just because assumptions (a), (b), and (c) may not always hold. (For more on these points, see §5.4.) Recall the case of Jack. His unexpected diagnosis can be appreciated on (at least) two distinct levels of abstraction (LoAs). On the one hand, there is the model-LoA. At this level, when Jack asks Q, he is seeking information about the diagnostic algorithm itself. What about f – its training data, parameters, etc. – led to this particular prediction? On the other hand, there is the system-LoA. At this level, when Jack asks Q, he is seeking information about Jack *qua* biological organism. What set of physical circumstances account for the (presumed) fact that he has rare disease y despite showing no apparent symptoms?

There is an inherent ambiguity in iML's most obvious, uncontroversial goal. Of course we want algorithmic explanations that are *true*, or *accurate*, or *faithful* – but faithful to what? The model or the system? Do we care more about the diagnostic function that predicts Jack has rare disease y, or the biological facts that constitute truth conditions for the prediction? The two can quickly come apart, even when f attains perfect predictive accuracy. The issue, once again, is one of confounding. It may be that heartrate is a reliable proxy for some unobserved biological mechanism z that in fact drives y. Alternatively, heartrate may be strongly correlated with an observed covariate w (perhaps another proxy for z) such that any perturbation of one has an immediate effect on the other. We know that the synthetic datapoint x' achieves the desired outcome y', but there are legitimate concerns about how informative this is when x' is biologically impossible. The model f has no preconceived notions about how interventions on one predictor may impact others, but nature inevitably imposes certain non-

trivial constraints. These are just a few of the problems that emerge when we confuse explanatory levels of abstraction.

§4.3.2 Variable Importance Measures

The dichotomy between model- and system-LoAs echoes a debate between advocates of two different approaches to measuring variable importance (VI). As noted above, feature attribution methods are an active area of iML research, especially local linear approximators. The first major work in this area was arguably the quantitative input influence algorithm, aka QII (Datta, Sen, & Zick, 2016), although user-friendly Python implementations have made the aforementioned LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) algorithms dominant in recent years. These methods diverge in several significant respects – see §3.1 for an overview – however, each attempts to answer Q by means of a linear combination of input features optimised to hold around the point $\boldsymbol{x}_i$. Crucially, these methods all assume that predictors are mutually independent, i.e. for all $j, j' \in [p]$ such that $j \neq j', X_j \perp X_{j'}$. This enables something like the naïve approach described above, in which predictors may be dialled up and down at will without concern for the plausibility of the resulting inputs.

It is not always clear whether the authors fully appreciate just how strong the mutualy independence assumption really is, or just what its implications truly are. For example, Ribeiro et al. pass over the point in silence. Datta et al. explicitly defend the choice on causal grounds, which will be explored more thoroughly below. Lundberg & Lee appear almost apologetic, explaining that feature independence is not so much an assumption as an "approximation" (2017, p. 5). They go on to plead innocence by association, pointing out that a similar move is made by many others in the field. Subsequent work relaxed the assumption in the special case of tree-based models (Lundberg et al., 2020), further indicating that the creators of SHAP were never fully comfortable with the choice. A number of authors have criticised the original SHAP algorithm for failing to model covariate dependencies and proposed a number of "improvements" that incorporate conditional (Aas, Jullum, & Løland, 2019; Kumar et al., 2020) or causal (Frye, Feige, & Rowat, 2019) information. Meanwhile, Janzing et al. (2020) insist that the original SHAP algorithm is sound, and that purported improvements are conceptually misguided.

I shall argue that every one of these authors is right – or at least that none of them is entirely wrong. But that does not mean that the decision to incorporate or ignore dependencies between covariates should be made lightly. On the contrary, the choice has major implications for how results are interpreted. In statistical terms, we may formalise the difference as one between marginal and conditional association measures. The null hypothesis of a marginal feature attribution test is:

$$H_0^m: X_j \perp Y, \boldsymbol{X}_{-j}$$

where $\boldsymbol{X}_{-j}$ denotes a set of covariates. A conditional dependence measure, on the other hand, tests against a different null hypothesis:

$$H_0^c: X_j \perp Y | \boldsymbol{X}_{-j}$$

Observe that the former entails the latter, as conditional independence is just one possible form of marginal independence. Since H_0^c is more restrictive, we may find instances in which it holds but H_0^m does not. Specifically, this will be the case whenever X_j's marginal importance is high due to its association with $\boldsymbol{X}_{-j}$ rather than Y.

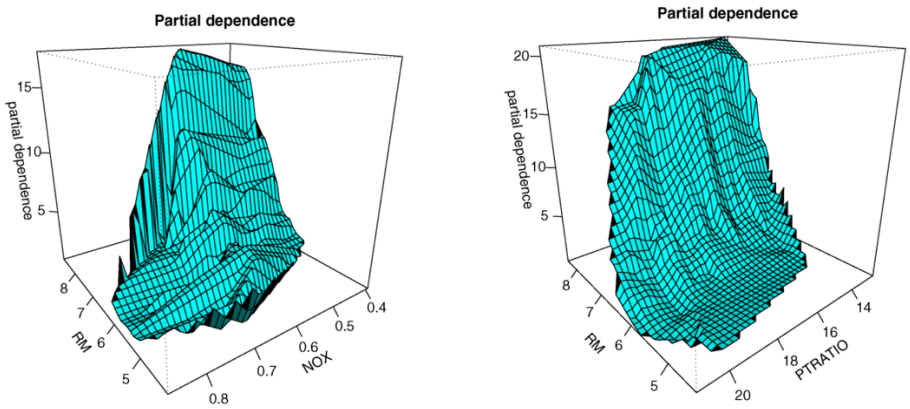

Figure 4.3. Three-dimensional partial dependence plots depicting the relationship between predictors in the benchmark Boston housing dataset in a tree-based ensemble model. From (Friedman & Popescu, 2008, p. 950).

The impulse to estimate feature importance in an entirely model-centric manner that ignores covariate dependencies altogether is evident in permute-and-predict (PaP) approaches, which take their inspiration from classical methods (Fisher, 1935). In supervised learning contexts, the most famous PaP technique is Breiman's (2001) permutation importance for random forests. He proposes to estimate the VI of X_j in a given forest f by comparing predictive performance on data before and after permuting X_j. Large post-permutation error inflations are interpreted as strong evidence that f relies on X_j to estimate outcomes Y. A more general "reliance" statistic is introduced by Fisher, Rudin, & Dominici (2019), who derive uniform bounds for a number of PaP tests, as well as analytic formulae for estimating reliance when models are additive functions in a reproducing kernel Hilbert space. The partial dependence function, originally proposed by Friedman (2001), is another popular PaP method. Partial dependence plots visualise the change in expected value of a function f as we marginalise over the empirical distribution of a feature subset while holding values for the complementary subset constant (see Fig. 4.3.).

Critics of PaP methods charge that marginal VI measures overstate the importance of uninformative variables when predictors are highly correlated. This has been the focus of considerable research in random forests, where a number of authors have proposed alternatives designed to overcome this perceived shortcoming of Breiman's permutation importance

(Altmann et al., 2010; Gregorutti, Michel, & Saint-Pierre, 2015; Mentch & Hooker, 2016; Nembrini, König, & Wright, 2018; Nicodemus et al., 2010; Strobl et al., 2008). Other, more general tests of conditional independence often rely on model refitting (Lehmann & Romano, 2005; Lei et al., 2018; Rinaldo, Wasserman, & G'Sell, 2019) or kernel methods (Doran et al., 2014; Fukumizu et al., 2008; Zhang et al., 2011), which can be computationally expensive for large datasets. In general, conditional independence testing is statistically *hard* in the precise sense that any procedure that controls the false positive rate at target level α cannot detect true positives for arbitrary alternative hypotheses with sensitivity greater than α (Shah & Peters, 2020). This result is somewhat surprising given the fact that permutation tests are exact and uniformly valid in the marginal case.

In a recent article with a blunt title – "Please stop permuting features: An explanation and alternatives" – Hooker & Mentch (2019) provide an intuitive explanation for how and why PaP methods can mislead. Permuting some X_j does not just break its dependence with the target Y, but also with the remaining features $\boldsymbol{X}_{-j}$. When covariance between predictors is high, the resulting inputs will tend to look unlike anything in f's training data. For example, Hooker & Mentch note that a PaP procedure evaluating the importance of pregnancy status in a model f that also includes sex would force f to predict outcomes for pregnant males as often as pregnant females. Should f perform poorly on such datapoints – as we might expect – then pregnancy will receive high VI, even if it is independent of the response Y. Since the efficacy of supervised learning relies crucially on the i.i.d. assumption, which states that training and test data are sampled from the same distribution, we should not be surprised when f errs in this new environment. When queried with exotic data, synthetically generated and far from its training manifold, the model has no choice but to *extrapolate*. Just because a model struggles to extrapolate well from real observations to imaginary hypotheticals does not mean that the permuted variable was predictive.

The validity of these objections notwithstanding, PaP methods occasionally boast attractive theoretical properties. For instance, Fisher et al. (2019, §8) demonstrate that under some common assumptions,[8] their model reliance measure may be decomposed into a sum of familiar terms from causal inference. Zhao & Hastie (2019) observe that Friedman's partial dependence function is formally identical to Pearl's (2000) backdoor adjustment when conditioning variables meet certain conditions.[9] In other words, the partial dependence of an outcome on a given feature subset in a purely predictive model may have a natural

[8] These assumptions include *conditional ignorability*, which states that potential outcomes are independent of treatment given covariates, and *positivity*, which states that propensity scores are bounded away from the extrema of the unit interval. Neither assumption is trivial, but both are fairly common in causal inference. See (Imbens & Rubin, 2015).

[9] A set of variables Z satisfies the backdoor criterion relative to an ordered pair of variables (X_i, X_j) in a directed acyclic graph G if (i) no node in Z is a descendant of X_i; and (ii) Z blocks every path between X_i and X_j that contains an arrow into X_i. See (Pearl, 2000).

interpretation as the causal effect of those variables on the outcome. A similar idea lies behind Datta et al.'s (2016) QII procedure and Janzing et al.'s (2020) defence of the original SHAP method. Both groups argue that marginalising over covariates is the right choice because the ERM algorithm itself does not explicitly model interdependencies. The resulting VI estimates are therefore causal at the model-LoA.

§4.3.3 The Correctness Theory of Truth

To review – advocates of conditional VI measures argue that their method alone recovers causal effects in the data generating process. Advocates of marginal VI measures respond that their method alone recovers causal effects in the target model. The solution to this impasse, as foreshadowed in §4.3.1, lies in the realisation that there is no impasse at all. Proponents of marginal and conditional association tests are asking different questions. They should not be surprised to receive different answers. One approach is preferable at the model-LoA, while the other performs better at the system-LoA. There is no statistical inconsistency here, merely underspecified pragmatics.

The essential role of pragmatics in ordinary language is highlighted by numerous twentieth century philosophers (Austin, 1961; Grice, 1989; Strawson, 1964). For a concise formalisation, I turn to Floridi's (2011a) correctness theory of truth (CTT), an elegant work of formal epistemology developed as part of his systematic philosophy of information (2011b). The full details Floridi's CTT, as well as those of this larger project, are beyond the scope of this chapter. For my purposes here, I shall focus merely on how the theory clarifies the essential semantic work done (often implicitly) by pragmatic auxiliaries. According to the CTT, semantic information can always be polarised into question/answer pairs – but these pairs may only be evaluated once we have specified a particular *context*, *level of abstraction*, and *purpose* (collectively labelled "CLP parameters"). The decomposition takes the form of a sum:[10]

$$i = [q + r]_{\mathrm{CLP}}$$

where i denotes the information in question, q the interrogative expression thereof, and r a Boolean yes/no reply to q. The CLP indexing is essential to ground q, and therefore define truth conditions for r. Floridi cautions that "Queries cannot acquire their specific meaning in isolation or independently of CLP parameters" (2011a, p. 155). This is most obviously the case when questions contain one or more indexicals – e.g., pronouns such as "I" and "she", or qualifiers like "here" and "presently" – however, the point applies much more broadly.

To borrow Floridi's own example, consider the proposition "The beer is in the fridge." This sentence conveys some semantic information, which constitutes the LHS of the above

[10] I take the liberty of reformulating somewhat for the sake of notational consistency. The content of the CTT remains unchanged.

equation, i. The interrogative form of this sentence is q = "Is there beer in the fridge?" Such a question does not and cannot occur in a vacuum. It must be uttered by and to embedded agents with certain beliefs and interests. Call these agents Inquirer and Responder. Perhaps Inquirer is preparing for a party (context). Or maybe the question is not about the immediate circumstance, but about whether, in general, Responder is in the habit of keeping beer in the fridge (level of abstraction). Inquirer may be asking because she wants a cold beer, or perhaps because she just bought some groceries and is worried they will not fit in the overcrowded fridge (purpose). Note that these pragmatic considerations all interact with one another, and may not even be well distinguished in some cases. However, with sufficient modification of the CLP parameters, we can always alter the meaning of q, and with it, truth conditions for r.

This lesson has immediate implications for iML. Our guiding question, as specified in §4.2.1, is Q. But the proper interpretation of this query varies as a function of the CLP parameters. For instance, if we have reason to believe that Jack's unexpected diagnosis is a result of algorithmic discrimination – e.g., that he was misdiagnosed with disease y due to a sensitive attribute such as race – then we probably want to focus on the model-LoA. Our goal here is simply to find out what the algorithm has learned, in full awareness that this may deviate arbitrarily from the ground truth. In this case, marginal VI measures such as those provided by LIME and SHAP are appropriate, and the causal inferences they license tell us how f relies on race to make predictions. Alternatively, if we have full confidence in f, we may want iML methods to help shed light on poorly understood mechanisms. Perhaps the algorithm has correctly identified some elusive biomarker for y, and Dr. Jill would like an explanation of Jack's diagnosis at the system-LoA. In this case, conditional VI measures are needed in order to find the right causal structure.

In general, the choice of whether to assume the mutual independence of predictive features cannot be determined by mathematical or computational considerations alone. Both methods of quantifying feature attributions have their place; neither dominates the other as an all-purpose tool for model interpretability. Marginal tests are ideally suited to tasks such as auditing or validating a supervised learning model, where we are generally more concerned with internal properties of the algorithm than with the data generating process. Conditional tests are better suited to tasks of discovery, where the model is not an object of inherent value so much as an instrument through which we may learn about an underlying system that is too complex, expensive, and/or risky to probe directly. In this setting, we do not have the luxury of ignoring feature covariance and treating all points on some large grid as approximately equiprobable. We must take great pains to understand the joint distribution of the data, create an SCM that approximates the behaviour of the system, and use conditional tests to evaluate feature importance in a principled manner.

Some authors speak of conditional VI as an "improvement" over marginal measures (Aas et al. 2019; Frye et al., 2019), perhaps because the former is generally more complex or closer to nature. But the concept of "improvement" is misplaced here. The relationship between marginal and conditional measures is not that of a line progressing from less to more refined or informative statistics. A more apt geometric analogy would be a pair of nested spheres, representing the hierarchy of abstraction; or perhaps even two perpendicular lines that intersect at a point. They may agree at or near their region of intersection, but the approaches are orthogonal. Our decision to use one VI method or another essentially depends upon the CLP parameters of our underlying inquiry. *Q* does not simply ask itself. The question is posed by one or more agents with particular interests, embedded in some context. Failure to appreciate pragmatic variation creates unnecessary and unhelpful ambiguity in iML.

§4.4 Error Rates and Severe Testing

In §4.3, I argued that pragmatic considerations can and must inform iML analyses. In this section, I argue that, regardless of CLP parameters, we cannot rely on algorithmic explanations that do not pass severe tests. The utter lack of severity in the vast majority iML methods represents a missed opportunity to establish some much-needed rigour in this young and fast-evolving field of research.

§4.4.1 Severity Criteria

In a series of influential articles and monographs, Mayo (1996; 2006; 2010; 2018) advances a statistically sophisticated philosophy of science in which the problem of induction is reduced to the practice of *severe testing*. The basis for this reduction is her severity principle, which, in its strong form, states that "We have evidence for a claim *C* just to the extent it survives a stringent scrutiny. If *C* passes a test that was highly capable of finding flaws or discrepancies from *C*, and yet none or few are found, then the passing result, *x*, is evidence for *C*" (2018, p. 14). This principle shifts the focus of scientific discourse from physical theories to testing procedures. On Mayo's view, the justification for believing a given hypothesis is a function not of the hypothesis itself or the data it purportedly explains, so much as the tests it has passed. When tests are sufficiently sensitive (i.e., likely to detect true effects) and specific (i.e., likely to reject false effects), then we say they are *severe*.

Mayo works in the falsificationist tradition of Popper (1934). However, she aims to move beyond his negative result – that science can only advance knowledge by *disproving* theories – to a positive conclusion – that severe tests provide (fallible, statistical) evidence in favour of particular hypotheses. Unlike Bayesian epistemologists, who typically interpret

probabilities as degrees of belief computed by combining subjective priors[11] with evidential likelihoods, Mayo places her approach in the frequentist tradition, emphasising the importance of hypothesis testing with fixed error rates. Her work is grounded in a descriptive fact – ignored or lamented by both Popper and Bayesians, albeit for different reasons – that null hypothesis significance testing has been the dominant method of statistical inference across the natural and social sciences for the better part of a century. This is no oversight. The procedures originally conceived by Fisher (1925) and later extended by Neyman & Pearson (1933) provide a firm groundwork for rational and progressive theory testing, even if the founders themselves did not always see eye to eye on what exactly those methods were intended to show.

ML is not inherently associated with either Bayesian or frequentist interpretations of probability. Some may choose to view this agnosticism as a sign of strength – the algorithms work no matter how you feel about *p*-values or prior distributions – others as a dangerous portent of ML's theoretical vacuity. I will not have much to say here regarding the (occasionally bitter) foundational debates between competing schools of probability theory.[12] It will suffice to observe that most applied statisticians are unmoved by the dogmatism of either camp, and willing to use whatever combination of methods is best suited to a given problem. Partisans of both traditions largely agree on particular inferences, especially when Bayesians use uninformative priors and/or when datasets are sufficiently large. Numerous convergence theorems have shown that priors wash out in the limit, as we might hope and expect (Earman, 1992).[13] Moreover, methodological syntheses are possible. Empirical Bayes inference (Efron, 2010) and PAC-Bayes learning (Guedj, 2019) are just two examples of popular methods that borrow heavily from both traditions. In any case, though Mayo's allegiance undoubtedly skews frequentist, her error-statistical philosophy has been reinterpreted along Bayesian lines (Gelman & Shalizi, 2013). In what follows, I will generally stick to her frequentist exposition more out of convenience than conviction.

There are subtle distinctions between the severity criteria articulated in Mayo's various writings on the topic, as well as some occasionally confusing discussion of a so-called "severity function" that need not concern us here. To make matters concrete, I will stick to the spirit if not the letter of Mayo's approach and explicate her theory entirely in terms of Neyman-Pearson (NP) testing. Specifically, I shall focus on arguably the simplest and most common sort of hypotheses in statistical inference, namely those positing some value or range of values for a

[11] Some Bayesians, it should be noted, self-identify as "objectivists"; see, e.g., (Berger, 2006; Jaynes, 2003). I will not concern myself too much with these distinctions. See (Talbott, 2016) for an overview.
[12] For a good introduction, see (Romeijn, 2017). For a more comprehensive compendium, see (Bandyopadhyay & Forster, 2011).
[13] Somewhat disconcertingly, it can also be shown that, for any body of evidence, we may construct (nonextreme) priors for a given hypothesis *H* such that corresponding posteriors differ by an arbitrarily large amount. See (Kyburg, 1992).

single parameter.[14] Let Θ denote the parameter space, and let T be a test that decides between $H_0: \theta \in \Theta_0$ and $H_1: \theta \in \Theta_1$, where Θ_0 and Θ_1 are some partition of Θ. We observe sample data x and compute sufficient statistic $d(x)$, which measures the disagreement between x and H_0. Test T rejects H_0 when $d(x)$ meets or exceeds the critical value c_α. We say that H_0 passes an (α, β)-severe test T with data x if and only if:

(S1) $d(x) < c_\alpha$; and

(S2) with probability at least $1 - \beta$, if H_1 were true, then we would observe some sufficient statistic $d(x')$ such that $d(x') \geq c_\alpha$.

Readers well-versed in frequentist inference will recognise some familiar concepts in these criteria. The critical value is indexed by the type I error rate α, such that, under H_0, the rejection region of statistics greater than or equal to c_α integrates to α. The type II error rate is given by β, such that, under H_1, the rejection region of statistics less than c_α integrates to β (see Fig. 4.4). The complement of this value, $1 - \beta$, denotes the power of the test. A test with small α is said to be *specific*, since it only accepts hypotheses that are likely to be true; a test with small β is said to be *sensitive*, since it is able to detect even slight deviations from the null. A maximally severe test is one that finds all ($\beta = 0$) and only ($\alpha = 0$) true effects. Such stringency is generally not possible in real-world experiments, where there is an inevitable trade-off between sensitivity and specificity. According to Mayo, science advances knowledge not just by falsifying theories, as Popper would have it, but by subjecting hypotheses to increasingly severe tests. Hypotheses earn their warrant by passing such tests, thereby providing positive justification for successful theories.

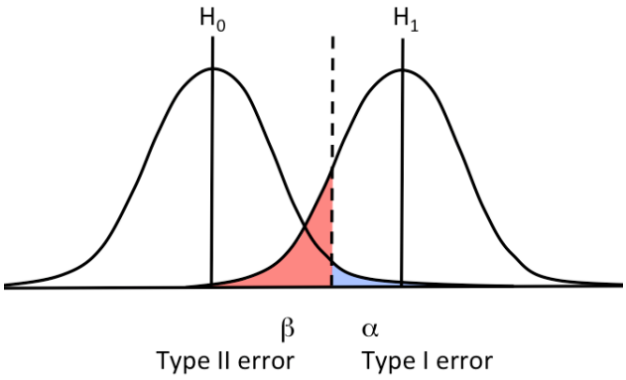

Figure 4.4. Null and alternative distributions for a given hypothesis test. The critical value is denoted by the dashed line. Type I error is represented by the blue integral; type II error is depicted by the red integral.

My brief account here glosses over a number of important subtleties that matter a great deal in practice, such as how exactly one goes about defining hypotheses, gathering data, and computing probability distributions. This arguably constitutes the bulk of puzzle-solving

[14] Extensions to compositive hypotheses and/or multiple testing scenarios are conceptually straightforward, but technically tedious and beyond the scope of this chapter.

activity that Kuhn (1970) regards as central to "normal science". There is no simple recipe for any of these crucial steps, however a handful of valuable heuristics are known to work well in a variety of settings. Of course, failure to adequately consider these sorts of questions could doom any experiment and inevitably opens the door to sceptical objections such as the aforementioned underdetermination of theory by evidence. The best antidote is generally a combination of clearly stated statistical assumptions, epistemological theory, and rigorous misspecification tests; see (Mayo & Spanos, 2004; Spanos, 2010) for an overview. No matter the details of particular testing methods, the point I want to stress is that minimising expected errors of the first and second kind is an obvious desideratum for any inference procedure, not to mention a faithful description of most modern scientific practice.[15]

Frequentist overtones notwithstanding, this account of severity is in fact very general. Unlike Mayo, I am agnostic with respect to how conditional probabilities ought to be computed or interpreted. Whether we use sampling distributions or posteriors makes no substantive difference. Some Bayesians may take issue with the dichotomous thinking implicitly endorsed here. Why stipulate that test outcomes can only be of the "accept" or "reject" variety? Surely there are shades of grey, degrees of uncertainty, etc. I have no objection to more inclusive approaches, such as plotting relationships between test statistics, error rates, and sample sizes, or visually inspecting conditional distributions. But note that the (α, β) parameters themselves serve to qualify test outcomes by specifying thresholds at which the relevant hypothesis is accepted or rejected. A similar procedure occurs in Bayesian analysis, where decisions typically turn on Bayes factors or credible intervals. Blind insistence on particular error rate thresholds, such as $\alpha = 0.05$, is obviously problematic, especially when conventions arising from one discipline or experimental design are mindlessly transported into another with different statistical properties (Ioannidis, 2005; Wasserstein & Lazar, 2016; Ziliak & McCloskey, 2008). Yet the mistake here lies in how people use or interpret severity criteria, not with the criteria themselves (Greenland, 2019). Matters would be no better if we were to replace an uncritical dogmatism about *p*-values with an uncritical dogmatism about Bayes factors, completely independent of any concerns regarding the provenance of prior distributions. A severe test is just one that should detect errors if they are present. As Floridi's CTT foreshadowed in §4.3, the fact that tolerable type I and type II error rates vary according to context, level of abstraction, and purpose is only to be expected.

§4.4.2 Severity in iML

Given the prevalence of ML in high-stakes public and private sector applications (to say nothing of scientific research), one might expect authors in this area to take error rates very

[15] Other sorts of statistical errors have also been formalised, such as those pertaining to sign and magnitude (Gelman & Carlin, 2014). For the sake of brevity, I will limit myself to type I and type II errors.

seriously. In fact, there is a shocking dearth of methods for estimating the sensitivity and specificity of algorithmic explanations. The most popular open-source software solutions make no effort to test the causal effects they infer, evaluate the uncertainty of their outputs, or bound their region of relevance. Some notable counterexamples exist (more on these below), but they are conspicuously, scandalously few. Given that algorithmic explanations are essentially causal claims, and that causal claims are typically the realm of science, we may justifiably wonder whether Mayo's severity criteria can be fruitfully applied in this setting. I argue that they can and should. In this subsection, I highlight two ways that algorithmic explanations mislead when severity criteria are not taken into account.

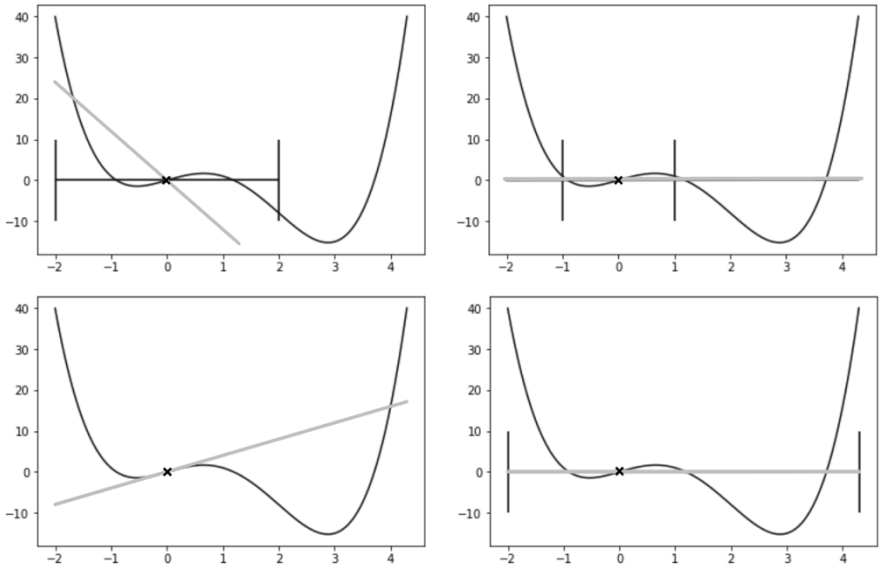

Figure 4.5. Unstable linear approximations. The grey line in each subfigure shows a local approximation of the same function centred at the same location. The varying range is indicated by the black bars, leading to vastly different linear explanations. From (Wachter et al., 2018, p. 885).

§4.4.2.1 How Local is "Local"?

Local explanations are constructed to apply only in some fixed region of the feature space. Yet iML methods do not generally provide information about the bounds of a given explanation or goodness of fit within the target region, facts that may be crucial for someone facing a consequential decision on the basis of an algorithmic explanation. For illustration, I will focus here on linear approximators, but the point applies more broadly.

If you zoom in far enough to any point on a continuous function, you will eventually find a linear tangent. This is the intuition behind methods like LIME and SHAP. However, when the regression surface or decision boundary around the target point is extremely nonlinear, the linear region tends to be very small and the estimated coefficients highly unstable. In this case, model weights are acutely sensitive to regional bounds. In a simple two-dimensional example, Wachter et al. (2018) visually demonstrate how a linear explanation for the same model prediction may assign positive, negative, or zero weight to a feature depending on

the scope of the linear window (see Fig. 4.5). This is a simple consequence of model misspecification. Recall that the output of local linear approximators is just a weighted sum of inputs. This explanation inherits all the virtues and vices of linear functions, which are often preferred for their relative ease of interpretation but bemoaned for their inflexible model assumptions. The less these assumptions hold near a given point, the less reliable our linear approximation will be.

The most obvious statistical solution here, should we insist on sticking with linear approximators, would be to augment iML outputs with some information regarding the scope and fit of the approximation. It is common, for instance, in linear regression to compute the significance and standard error of model coefficients. This would satisfy (S1). Power analysis typically requires parametric assumptions or data simulations, which could be used to satisfy (S2). Unfortunately, these strategies are not readily available to algorithms like LIME and SHAP, which use unconventional sampling techniques, kernel weights, and regularisation penalties that preclude easy analytic solutions for calculating expected error rates. Resampling methods such as bootstrapping (Davison & Hinkley, 1997) could help evaluate parameter uncertainty; however, this would substantially reduce the computational efficiency of these algorithms, which is arguably one of their greatest selling points. The problem could become especially acute as the number of explananda increases.

While reporting standard errors would certainly be an improvement over current practice, it would by no means resolve the fundamental problem of model misspecification. To evaluate the utility of a given linear approximation, we need a better sense of the target function's topology near our input point of interest. Formally, we are focused on a d-dimensional hypersphere around $\boldsymbol{x}_i$ with radius ε.[16] For each feature X_j, we need to know how the corresponding weight ϕ_j and standard error σ_j vary with ε. This three-dimensional surface will likely be more informative than the linear approximation itself. Extreme sensitivity to ε on the part of VI scale and location parameters indicates a highly nonlinear neighbourhood around $\boldsymbol{x}_i$, which means that any local linear approximation should be interpreted with caution – or, better yet, abandoned altogether. Statistical tests offer a principled way to evaluate these relationships, but informal methods may serve just as well. Old-fashioned scatterplots can be enormously helpful in exploring these sorts of multivariate associations. Of course, this can quickly become impractical as the number of input features grows.

§4.4.2.2 Correlated Predictors

Another challenging scenario for iML tools is when predictors are strongly correlated. For in-

[16] Assume here, for simplicity, that all predictors are continuous and scaled to unit variance, and that our distance metric is L_2. Of course, these assumptions are (almost always) violated in practice, but such complications do nothing to mitigate the problem. On the contrary, they only make matters worse.

stance, as noted above, it will be difficult if not impossible to decide whether sex or treatment best explains drug trial outcomes when the two are strongly confounded. This issue can be especially nefarious in the setting of algorithmic fairness. When a sensitive attribute is associated with a permissible variable – e.g., if race is well predicted by zip code (Datta et al., 2017) – then the latter can serve as a proxy for the former. This allows bad actors to get away with discrimination, so long as they can fool an auditor into believing they were using the permissible variable rather than the sensitive one. The concern is not merely speculative. Lakkaraju & Bastani (2020) demonstrate how such deceptive practices are possible even under perfect explanatory fidelity, and generate misleading explanations on a range of real-world examples. Pruthi et al. (2020) use similar methods to manipulate weights in a way that makes a discriminatory natural language processing model appear fair in a user study. Slack et al. (2020) design an adversarial procedure for obscuring biases from LIME and SHAP, and use it to create a racist classifier from the COMPAS dataset that passes fairness audits according to both iML methods.

Severe testing cannot, on its own, prevent bad actors from engaging in discriminatory behaviour. However, it can make it harder for them to get away with it by elucidating the uncertainty associated with algorithmic explanations under confounding. Just as standard errors for regression coefficients are inflated by collinear predictors, the severity of particular explanations will tend to decrease with strongly correlated features. Reporting the error rates of given outputs at local or global scales will provide some much-needed context for users and regulators alike. When predictors are strongly correlated, then it is very difficult to assert with high confidence that one variable and not another is causally responsible for the observed outcome without introducing some structural assumptions. Such assumptions may be justifiable, but they will need to be articulated and defended. Even better, they can in many cases be severely tested themselves.

Algorithmic fairness is a complex and contested topic. Dozens of statistical fairness criteria have been proposed – see (Barocas et al., 2019) for a good overview – while impossibility theorems have shown that most of the popular definitions are mutually incompatible except in trivial cases (Kleinberg, Mullainathan, & Raghavan, 2017). No matter which criteria one adopts for a given application, almost all may be expressed in terms of marginal or conditional independence relations, which means that classical NP tests can be used for auditing purposes. Despite Shah & Peters's aforementioned hardness result, a large number of conditional independence tests have been developed over the years, many with impressive performance on real-world datasets.[17] Severity therefore has a central role to play in holding people and institutions accountable for their algorithmically mediated decisions.

[17] For a good review of such methods, see (Heinze-Deml, Peters, & Meinshausen, 2018).

§4.4.3 Severity and Trust

Many authors motivate the iML project with appeals to *trust*. "Why should I trust you?" reads the title of Ribeiro et al.'s (2016) paper introducing LIME. Successful algorithmic explanations "engender appropriate user trust," (Lundberg & Lee, 2017, p. 1) write the creators of SHAP on the first page of their award-winning NeurIPS paper. In their *Harvard Journal of Law and Technology* article introducing counterfactual explanations, Wachter et al. argue that "Building trust is essential to increase societal acceptance of algorithmic decision-making" (2018, p. 843). So long as complex black box algorithms remain opaque and impenetrable, users will harbour suspicions about their reliability in particular cases. That is why we seek transparent explanations that can assuage concerns about unfair or unreasonable model predictions.

Yet do methods like LIME and SHAP really settle matters, or merely push the problem one rung up the ladder? After all, why should we trust the outputs of iML algorithms? Presumably the original function f at least has the advantage of performing well on some test dataset. According to reliabilist philosophers, this may be sufficient to justify belief in its predictions (Goldman, 1979). Can we say the same of algorithms like LIME or SHAP? Their outputs are readily intelligible, and that is clearly a start. But does that necessarily mean that their explanations should all be given equal weight? Or are some more reliable than others? How can we be sure that they have not produced unstable estimates or selected the wrong features? Are there principled methods for critically evaluating individual explanations, much like we can critically evaluate individual predictions?

I argue that severe testing holds the key to securing the trustworthiness of algorithmic explanations. Recall that the response to Q is always a certain sort of causal claim, and causal claims can in principle be tested. That, for instance, is how we come to trust scientific theories – by mercilessly subjecting them to numerous tests with quantifiable error rates. It is not always immediately obvious how one ought to go about testing algorithmic explanations, especially those that do not boil down to particular parameter estimates. However, some iML authors have begun to try. In a follow up to their LIME article, Ribeiro et al. (2018) introduce a novel iML algorithm called "anchors". Anchors are set of Boolean conditions that hold at the target point, selected to ensure some minimal level of precision (i.e., probability of securing the desired outcome) and optimised for coverage (i.e., applicable region of the feature space). These parameters effectively bound resulting error rates. Other methods for testing local explanations include the localised knockoff procedure (Gimenez & Zou, 2019), as well as leave-one-covariate-out (LOCO) inference techniques (Lei et al., 2018; Rinaldo et al., 2019), both of which quantify and control errors of the first kind. The CXPlain algorithm (Schwab & Karlen, 2019), presented at the 2019 NeurIPS conference, evaluates the uncertainty of its explanations via bootstrap resampling techniques.

These methods are not without their difficulties. However, they represent a notable advance over the previous state of the art, in that they explicitly try to quantify and optimise the quality of individual explanations with testable claims. Unfortunately, the majority of iML authors have yet to take notice. Until severe testing is built into iML, the field will fail to meet the standards of scientific rigour required for widespread adoption and user trust.

§4.5 Process vs. Product

One way to classify iML algorithms is by their output class. Saliency methods, which are popular for image classification tasks, produce visual explanations highlighting the pixels (or superpixels) that are most relevant in generating particular predictions. VI methods, by contrast, produce statistical outputs measuring importance at global or local resolutions. Counterfactual and case-based explanations generate examples intended to elucidate model predictions. In each case, the output is a *product* – that is, a static deliverable that is computed once and for all. In this section, I argue that a more helpful way to think of explanations is as a *process* – an iterative exchange between (at least) two agents engaged in a certain sort of causal inquiry. Such explanations are not just more mimetic of how explanations unfold in real life, but are also more likely to ensure understanding on the part of the inquiring agent.

§4.5.1 Dialogical Explanations

There is a tendency in analytic philosophy to think of explanations as *arguments* or *models* with certain characteristics. Famous twentieth examples include the deductive-nomological model (Carl Hempel, 1965), the statistical relevance model (Salmon, 1971), the causal mechanical model (Dowe, 2000; Salmon, 1984), and the unificationist model (Kitcher, 1989).[18] However, there is a more ancient tradition that conceives of explanations in a very different way – as fundamentally *interactive* and *dialogical*. The roots of this form go back to Ancient Greece, although adherents may be found among the scholastics (e.g., Anselm, 1080) and early moderns (e.g., Berkeley, 1713). Even within the ranks of the most staid contemporary logicians, there are those who find it helpful to frame formal proofs as dialogues or games (Hintikka, 1999; Hodges & Väänänen, 2019; Keiff, 2011). I believe there are good reasons to prefer explanations of this sort as well.

I highlighted the role of pragmatic information – specifically, CLP parameters – in §4.3. However, there is more to pragmatics than assiduously indexing the context, level of abstraction, and purpose of particular inquiries. By defining the overarching goal of iML as answering *Q*, I have already framed the undertaking as essentially interrogative, with the implicit suggestion that at least two agents are engaged in some form of inquiry regarding the

[18] For more on these and other related proposals, see (Woodward, 2019).

predictions of a target algorithm. Yet even if we allow CLP parameters to vary, the static form of explanation remains irreparably impoverished. The idea that tacking on some extra information will always suffice to explain predictions is restrictive and naïve. It ignores the possibility that answers to *Q* may leave an agent confused or open up whole new avenues of inquiry. Interactive explanations allow the inquiring agent to get a more complete picture of the explanans and its place in a wider body of knowledge. This is essential as soon as we acknowledge that agents will request algorithmic explanations with different motivations, expectations, and beliefs. Rather than creating a one-size-fits-all solution, the dialogical approach lets the inquiring agent guide the discussion to best satisfy her needs and curiosity.

Pragmatists have long argued against monolithic theories of explanation. A number of notable twentieth century philosophers have proposed alternative accounts (Achinstein, 1983; Bromberger, 1966; Scriven, 1962), but perhaps no one crystallises their collective critique so neatly as van Fraassen:

> The discussion of explanation went wrong at the very beginning when explanation was conceived of as a relation like description: a relation between a theory and a fact. Really, it is a three-term relation between theory, fact, and *context*. No wonder that no single relation between theory and fact ever managed to fit more than a few examples! Being an explanation is essentially relative for an explanation is an *answer*…it is evaluated vis-à-vis a question, which is a request for information. But exactly…what is requested differs from context to context. (1980, p. 156)

If van Fraassen is right, then there can be no objective criteria that constitute necessary and sufficient conditions for successful explanation, no single set of parameters to optimise. The pragmatist starts from the simple, indisputable observation that explanations do not occur in a vacuum. Rather, they are the product of interactions between epistemic agents with certain beliefs and interests. For example, Dr. Jill may be satisfied with an explanation for Jack's unexpected diagnosis in terms of transcriptomic signatures and cellular phenomena. Jack, by contrast, may seek a higher level explanation in terms of more familiar biological functions. This reflects a difference not just in background knowledge, but in goals. Dr. Jill's aim is to understand disease mechanisms in order to better detect warning signs in future patients; Jack's aim is to treat his own condition, preferably through non-invasive behavioural adjustments. Of course, these goals are not mutually exclusive, but they suggest different explanatory emphases. In general, successful explanations must take into account both the epistemic state and guiding interests of whoever is asking the questions.

Among contemporary commentators, Walton has most extensively developed the dialogic model of explanation. In a series of articles (2004; 2006; 2011), he puts forward a framework for explanatory dialectics in which one agent imparts understanding to another through a sequence of well-structured exchanges. Though his focus is primarily on the *closing* stage of such dialogues – how can we be sure that an explanation has successfully concluded?

– I am especially interested in the intermediate *explanation* stage (Walton, 2011, §7), during which agents jointly explore a target system's behaviour and anomalies. Building on Moulin et al.'s (2002) tripartite distinction between trace explanations, strategic explanations, and deep explanations in AI systems, Walton argues that dialogic models are especially well suited to the latter category, in which agents help each other fill gaps in one another's knowledge. Current iML strategies at best provide trace explanations, which track the sequential reasoning steps of a target model, or strategic explanations, which outline more abstract problem-solving approaches. Yet all three modes are essential to help agents like Jack and Dr. Jill understand predictions like his unexpected diagnosis. On Walton's model, the two may each chart their own course through the algorithm's reasoning, addressing whatever strikes either them as unclear or anomalous through a series of speech acts that gradually bring them closer to understanding the original prediction. Such a personalised process is simply impossible with popular iML algorithms.

§4.5.2 Advantages for iML

There are at least two clear advantages to interactive explanations for iML. First, such approaches are inherently customisable, since they must respond to each agent's idiosyncratic questions. This accommodates the inevitable variability among users, who will approach unexpected predictions with a range of different assumptions and background beliefs. Second, interactive explanations promote greater user trust. Whereas a static iML algorithm simply spits out a set of parameters with no obvious account of how they were derived or how they are meant to fit in with other known facts about the system, an interactive method can address ambiguous or unexpected aspects of a model's reasoning one step at a time. This ensures users that the model is working as it should, or, alternatively, helps to isolate the error that led to an anomalous prediction.

As anyone who has spent time with young children can attest, an initial explanans often merely sets the stage for further questions about constituent terms or related phenomena.[19] This recursive pattern may continue more or less indefinitely, subject to constraints on the child's interest and the adult's patience. A similar pattern unfolds in scientific inquiry, with researchers assuming the role of curious children and nature the informed (though stubbornly coy) adult. Indeed, causal discovery has been described as a game between Scientist and Nature in which the latter attempts to hide its secrets for as long as possible while the former seeks to uncover them at minimal cost (Eberhardt, 2010). A preliminary question about some particular observation (e.g., "Why do finch beaks vary so widely across the Galápagos

[19] From a recent conversation with my nephew: "Why is the sky blue?" "Because of how the earth's atmosphere scatters sunlight." "What's the atmosphere?" "A layer of gases surrounding the planet." "Why don't the gases fly away?" "Because of gravity." "What's gravity?" "Go to bed."

Islands?") can quickly lead to profound questions about fundamental mechanisms (e.g., "How do species evolve over time?"). It is tempting to regard the final product of such an inquiry – say, Darwin's *On the Origin of Species* – as the explanation we were seeking all along. But this, I contend, would be a vast oversimplification. The journey counts every bit as much as the destination. We learn best not through the passive transmission of knowledge, but rather by actively formulating questions, gathering data, designing experiments, and generally engaging with the material. Because agents will tend to take different paths towards a discovery, converging on it from various angles, it would be a mistake for iML algorithms to ignore humans' natural epistemic heterogeneity. Giving all users the same answer, with no allowance for follow up questions, turns algorithmic explanations into oracular pronouncements, overstating our confidence in these potentially unstable outputs and precluding the most fruitful aspects of the natural explanation processes.

The inflexibility of an iML method that delivers static explanations, as the vast majority of algorithms in use today do, works against the goal of promoting greater user trust. The problem is especially acute when those explanations vary wildly in response to minor perturbations of hyperparameters or even due to random sampling, as we saw in §4.4.2. To avoid the (justified) perception that such methods merely replace one black box (for predictions) with another (for explanations), we need algorithms that can address various aspects of the learning pipeline, answering a range of questions about model behaviour under real and hypothetical interventions. Just as a knowledgeable scientist should be able to answer a student's questions about a target system,[20] a successful iML algorithm should promote learning and trust among users.

§4.5.3 Interactive iML Approaches

There is an acknowledged dearth of interactive methods in iML, despite some recent calls for more research in this area (Miller, 2019; Mittelstadt, Russel, & Wachter, 2019; Murdoch et al., 2019). A small group of intrepid computer scientists is actively working to fill the lacuna. Lakkaraju et al. (2019), specifically motivated by clinical applications for ML, develop a customisable decision set algorithm that allows users to specify features of interest. The explanations provided by their Model Understanding through Subspace Explanations (MUSE) algorithm are compact and provably optimal within a bounded subspace defined by the user. Akula et al. (2019) propose a natural language interaction method – literally instantiating a dialogic model – through which users may query a target algorithm about particular predictions. The approach is not very scalable, however, as it requires hand-crafted ontologies as well as unique and-or graphs for each application. In a pair of recent articles, Sokol & Flach

[20] Hopefully with greater detail and patience than I was able to provide my nephew.

(2020a, 2020b) develop and implement techniques for interactively interrogating black box algorithms. Their LIMEtree method is especially promising, providing local fidelity guarantees. However, at the time of writing, source code for this approach has not been made publicly available, which makes it difficult to benchmark against alternatives.

The last few years have also seen tepid first steps into interactive methodologies for the closely related field of algorithmic fairness. Jung et al. (2019), acknowledging the inherent difficulties in defining a context-independent metric for measuring the similarity of individuals, propose a flexible learning procedure in which human judges evaluate pairs of data points on a case by case basis. The resulting similarity scores are plugged directly into their algorithm, even though measures almost certainly deviate from the classic criteria for a metric.[21] The Jung et al. approach is specifically designed to accommodate such unorthodox kernels, which means it may be deployed in any setting where human judges can claim legitimate expertise. Meanwhile Canetti et al. (2019) develop post-processing methods that allow users to make targeted revisions to potentially unfair classifiers. Their approach effectively circumvents the aforementioned impossibility results, which purport to show that intuitive measures of statistical fairness cannot simultaneously hold except under extreme and improbable circumstances.

These examples of interactive algorithms are perhaps most notable for their scarcity. None of these methods has yet to gain much popularity among practitioners. However, it should also be noted that no algorithm mentioned in this section has yet to reach its second birthday. While the importance of the problem is widely acknowledged, the jury is still out on proposed solutions.

§4.6 Conclusion

Feyerabend (1975) famously argues that the ideal structure of scientific discovery is neither a logical sequence of conjectures and refutations (Popper, 1934) nor an orderly cycle of rising and falling paradigms (Kuhn, 1970), but rather a *marketplace* – a teeming bazaar in which theories multiply, combine, and clash in a protean struggle for supremacy. If Feyerabend's epistemological anarchism is to be believed, then iML may be in a sort of golden era. Research is expanding at a remarkable rate, with few checks on the proliferation of proposals.

However, Feyerabend's pluralism is too inclusive. The last three sections have chronicled major shortcomings of popular iML software. Practice has outpaced theory in this realm, and the result is a dizzying number of tools that suffer from similar oversights. Conceptual foundations are necessary in this new and urgent area of research. By articulating these critiques, my goal is not to inaugurate some new paradigm in which all iML research must be

[21] A distance function is called a metric if it follows three axioms: identity of indiscernibles, symmetry, and the triangle inequality.

conducted henceforth. A degree of pluralism is welcome and fruitful in young, dynamic subdisciplines such as this. Instead, my aim is merely to set up some pragmatic guardrails, to alert stakeholders to potential failures, and to identify promising new directions that are already being pursued by pioneering computer scientists.

I am sensitive to charges of pessimism. It is far easier to point out what is wrong with existing approaches than it is to advance positive counterproposals. However, this negative move in the dialectic is a critical first step toward that end. The technical work of developing practical algorithms for computing local and global explanations begins with an act of conceptual desk clearing. Wittgenstein's comments from the *Philosophical Investigations* are particularly apposite:

> It is the business of philosophy, not to resolve a contradiction by means of a mathematical or logico-mathematical discovery, but to make it possible for us to get a clear view of the state of mathematics that troubles us: the state of affairs *before* the contradiction is resolved.…One might also give the name "philosophy" to what is possible *before* all new discoveries and inventions. (1953, §§125-126)

I have argued that epistemology and philosophy of science are uniquely positioned to diagnose what ails iML, thereby setting the stage for new discoveries in this area. Building on centuries' worth of lessons from the analysis of scientific and statistical inquiry, philosophy has a key role to play in disambiguating interrelated concepts, drawing instructive analogies, and suggesting standards and strategies that are likely to promote greater algorithmic explainability.

For a relatively young research program, iML has come a long way in a short time. Numerous sophisticated proposals have been developed and implemented in just the last few years, including a number of popular off-the-shelf open-source tools. The rapid adoption of such software is understandable given the recent widespread deployment of supervised learning algorithms in high-risk applications. Public and private stakeholders all share an interest in making ML models more intelligible and trustworthy. The creators of iML software credibly argue that their solutions can ensure greater fairness, accountability, and transparency in artificial intelligence.

I have argued that despite the urgency of iML's mission, the conceptual foundations of the field are underdeveloped. I have highlighted three especially pressing, largely unacknowledged problems – ambiguous fidelity, lack of severe testing, and an emphasis on product over process – that undermine the vast majority of explainability software in use today. Without greater attention to these concerns, algorithmic explanations run the risk of being unclear, unstable, and unhelpful. Research has meticulously demonstrated failure conditions for a number of popular iML tools. The bad news is, it does not take much to break these methods. Some confounding between predictors is typically sufficient. They are prone to adversarial attacks, as well as simple misspecification. These worries are especially urgent as algorithms expand into ever more sensitive and high-stakes areas of public and private life. Data

regulation policy notwithstanding, the so-called "right to explanation" will remain not just unrealised but functionally impossible without technical procedures for overcoming these obstacles.

Fortunately, there is room for optimism. I have identified counterexamples to each of these problems from the iML literature that point the way toward more satisfactory solutions. Just because today's most popular methods do not always meet the highest standards is no cause for despair. On the contrary, a process of iterative refinement is only to be expected for a research program still in its infancy. The explainability discourse is teeming with novel methods and promising research on a number of fronts.

There are of course practical challenges to enacting the changes proposed herein. The resulting algorithms may be relatively slow and unfamiliar, requiring more user input than some people would like. Considerations of proper design, typically the domain of human computer interaction, will be paramount. But if the stakes are sufficiently high that we need an algorithmic explanation in the first place – perhaps even a legally mandated one – then it is important that we get that explanation right. Shortcuts and heuristics do us no favours here. A healthy mix of Feyerabendian pluralism and Kuhnian collective focus will go a long way toward advancing the state of the art for iML. That is an outcome that data scientists, policymakers, and end users alike can all get behind.

The Explanation Game:
A Formal Framework for Interpretable Machine Learning

§5 Abstract

I propose a formal framework for interpretable machine learning. Combining elements from statistical learning, causal interventionism, and decision theory, I design an idealised *explanation game* in which players collaborate to find the best explanation(s) for a given algorithmic prediction. Through an iterative procedure of questions and answers, the players establish a three-dimensional Pareto frontier that describes the optimal trade-offs between explanatory *accuracy*, *simplicity*, and *relevance*. Multiple rounds are played at different levels of abstraction, allowing the players to explore overlapping causal patterns of variable granularity and scope. I characterise the conditions under which such a game is almost surely guaranteed to converge on a (conditionally) optimal explanation surface in polynomial time, and highlight obstacles that will tend to prevent the players from advancing beyond certain explanatory thresholds. The game serves a descriptive and a normative function, establishing a conceptual space in which to analyse and compare existing proposals, as well as design new and improved solutions.

§5.1 Introduction

Machine learning (ML) algorithms have made enormous progress on a wide range of tasks in just the last few years. Some notable recent examples include mastering perfect information games like chess and Go (Silver et al., 2018), diagnosing skin cancer (Esteva et al., 2017), and proposing new organic molecules (Segler et al., 2018). These technical achievements have coincided with the increasing ubiquity of ML, which is now widely used across the public and private sectors for everything from film recommendations (Bell & Koren, 2007) and sports analytics (Bunker & Thabtah, 2019) to genomics (Zou et al., 2019) and predictive policing (Perry et al., 2013). ML algorithms are expected to continue improving as hardware becomes increasingly efficient and datasets grow ever larger, providing engineers with all the ingredients they need to create more sophisticated models for signal detection and processing.

Recent advances in ML have raised a number of pressing questions regarding the epistemic status of algorithmic outputs. One of the most hotly debated topics in this emerging discourse is the role of explainability. Because many of the top performing models, such as deep neural networks, are essentially black boxes – dazzlingly complex systems optimised for predictive accuracy, not user intelligibility – some fear that this technology may be inappropriate for sensitive, high-stakes applications. The call for more explainable algorithms has been especially urgent in areas like clinical medicine (Watson et al., 2019) and military operations (Gunning, 2017), where user trust is essential and errors could be catastrophic. This has

led to a number of international policy frameworks that recommend explainability as a requirement for any ML system (Floridi & Cowls, 2019).

Explainability is fast becoming a top priority in statistical research, where it is often abbreviated as xAI (explainable Artificial Intelligence) or iML (interpretable Machine Learning). I adopt the latter initialism here to emphasise my focus on supervised learning algorithms (formally defined in §5.3.1) as opposed to other, more generic artificial intelligence applications.

Several commentators have argued that the central aim of iML is underspecified (Doshi-Velez & Kim, 2017; Lipton, 2018). They raise concerns about the irreducible subjectivity of explanatory success, a concept that they argue is poorly defined and difficult or impossible to measure. In this chapter, I tackle this problem head on. I provide a formal framework for conceptualising the goals and constraints of iML systems by designing an idealised *explanation game*. This model clarifies the trade-offs inherent in any iML solution, and characterises the conditions under which epistemic agents are almost surely guaranteed to converge on an optimal set of explanations in polynomial time. The game serves a descriptive and a normative function, establishing a conceptual space in which to analyse and compare existing proposals, as well as design new and improved solutions.

The remainder of this paper is structured as follows. In §5.2, I identify three distinct goals of iML. In §5.3, I review relevant background material with a focus on formalisms that will be deployed in the following sections. I clarify the scope of my proposal in §5.4. In §5.5, I articulate the rules of the explanation game and outline the procedure in pseudocode. A discussion follows in §5.6. I consider five objections in §5.7, before concluding in §5.8.

§5.2 Why Explain Algorithms?

As I noted in Chapter 1, there are three primary goals that guide those working in iML: to *audit*, to *validate*, and to *discover*. These objectives help motivate and focus the discussion, providing an intuitive typology for the sorts of explanations we are likely to seek and value in this context. Counterarguments to the project of iML are delayed until §5.7.

§5.2.1 Justice as (Algorithmic) Fairness

Perhaps the most popular reason to explain algorithms is their large and growing social impact. ML has been used to help evaluate loan applications (Munkhdalai et al., 2019) and student admissions (Waters & Miikkulainen, 2014), predict criminal recidivism (Dressel & Farid, 2018), and identify military targets (Nasrabadi, 2014), to name just a few controversial examples. Failure to properly screen training datasets for biased inputs threatens to automate injustices already present in society (Mittelstadt et al., 2016). For instance, studies have indicated that algorithmic profiling consistently shows online advertisements for higher paying

jobs to men over women (Datta et al., 2015); that facial recognition software is often trained on predominantly white subjects, making them inaccurate classifiers for black and brown faces (Buolamwini & Gebru, 2018); and that predatory lenders use financial data to disproportionately target poor communities (Eubanks, 2018). Critics point to these failures and argue that there is a dearth of fairness, accountability, and transparency in ML – collectively acronymised as FAT ML, an annual conference on the subject that began meeting in 2014.

Proponents of FAT ML were only somewhat mollified by the European Union's 2018 General Data Protection Regulation (GDPR), which includes language suggesting a so-called "right to explanation" for citizens subject to automated decisions. Whether or not the GDPR in fact guarantees such a right – some commentators insist that it does (Goodman & Flaxman, 2017; Selbst & Powles, 2017), while others challenge this reading (Edwards & Veale, 2017; Wachter et al., 2017) – there is no question that policymakers are beginning to seriously consider the social impact of ML, and perhaps even take preliminary steps towards regulating the industries that rely on such technologies (HLEGAI, 2019; OECD, 2019). Any attempt to do so, however, will require the technical ability to audit algorithms in order to rigorously test whether they discriminate on the basis of protected attributes such as race and gender (Barocas & Selbst, 2016).

§5.2.2 The Context of (Algorithmic) Justification

Shifting from ethical to epistemological concerns, many iML researchers emphasise that their tools can help debug algorithms that do not perform properly. The classic problem in this context is *overfitting*, which occurs when a model predicts well on training data but fails on test data. This happened, for example, with a recent image classifier designed to distinguish between farm animals (Lapuschkin et al., 2016). The model attained 100% accuracy on in-sample evaluations but mislabelled all the horses in a subsequent test set. Close examination revealed that the training data included a small watermark on all and only the horse images. The algorithm had learned to associate the label "horse" not with equine features, as one might have hoped, but merely with this uninformative trademark.

The phenomenon of overfitting, well known and widely feared in the ML community, will perhaps be familiar to epistemologists as a sort of algorithmic Gettier case (Gettier, 1963). If a high-performing image classifier assigns the label "horse" to a photograph of a horse, then we have a justified true belief that this picture depicts a horse. But when that determination is made on the basis of a watermark, something is not quite right. Our path to the fact is somehow crooked, coincidental. The model is right *for the wrong reasons*. Any true judgments made on this basis are merely cases of epistemic luck, as when we correctly tell the time by looking at a clock that stopped exactly 24 hours before.

Attempts to circumvent problems like this typically involve some effort to ensure that agents and propositions stand in the proper relation, i.e. that some reliable method connects knower and knowledge. Process reliabilism was famously championed by Goldman (1979), who arguably led the vanguard of what Williams calls "the reliabilist revolution" (2016) in anglophone epistemology. Floridi (2004) demonstrates the logical unsolvability of the Gettier problem (in non-statistical contexts), while his network theory of account (2012) effectively establishes a pragmatic, reliabilist workaround.

Advances in iML represent a statistical answer to the reliabilist challenge, enabling sceptics to analyse the internal behaviour of a model when deliberating on particular predictions. This is the goal, for instance, of all local linear approximation techniques, including popular iML algorithms like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), which assign weights to input variables so users can verify that the model has not improperly focused on uninformative features like the aforementioned watermark.

§5.2.3 The Context of (Algorithmic) Discovery

I consider one final motivation for iML: *discovery*. This subject has so far received relatively little attention in the literature. However, I argue that it could in fact turn out to be one of the most important achievements of the entire algorithmic explainability project, and therefore deserves special attention.

Suppose I design an algorithm to predict subtypes of some poorly understood disease using biomolecular data. The model is remarkably accurate. It unambiguously classifies patients into distinct groups with markedly different prognostic trajectories. Its predictions are robust and reliable, providing clinicians with actionable advice on treatment options and suggesting new avenues for future research. In this case, we want iML methods not to audit for fairness or test for overfitting, but to reveal underlying mechanisms. The algorithm has clearly learned to identify and exploit some subtle signal that has so far defied human detection. If we want to learn more about the target system, then iML techniques applied to a well-specified model offer a relatively cheap and effective way to identify key features and generate new hypotheses.

The case is not purely hypothetical. A wave of research in the early 2000s established a connection between transcriptomic signatures and clinical outcomes for breast cancer patients (e.g., Sørlie et al., 2001; van 't Veer et al., 2002; van de Vijver et al., 2002). The studies employed a number of sophisticated statistical techniques, including unsupervised clustering and survival analysis. Researchers found, among other things, a strong association between BRCA1 mutations and basal-like breast cancer, an especially aggressive form of the disease. Genomic analysis remains one of the most active and promising areas of research in the natural sciences, and whole new subfields of ML have emerged to tackle the unique challenges

presented by these high-dimensional datasets (Bühlmann et al., 2016; Hastie et al., 2015). Successful iML strategies will be crucial to realising the promise of high-throughput sciences.

§5.3 Formal Background

In this section, I introduce concepts and notation that will be used throughout the remainder of the chapter. Specifically, I review the basic formalisms of supervised learning, causal interventionism, and decision theory.

§5.3.1 Supervised Learning

The goal in supervised learning is to estimate a function that maps a set of predictor variables to some outcome(s) of interest. To discuss learning algorithms with any formal clarity, we must make reference to values, variables, vectors, and matrices. I denote scalar values using lowercase italicised letters, e.g. x. Variables, by contrast, are identified by uppercase italicised letters, e.g. X. Matrices, which consist of rows of observations and columns of variables, are denoted by uppercase boldfaced letters, e.g. $\mathbf{X}$. I sometimes index values and variables using matrix notation, such that the i^{th} element of variable X is x_i and the j^{th} variable of the matrix $\mathbf{X}$ is X_j. The scalar x_{ij} refers to the i^{th} element of the j^{th} variable in $\mathbf{X}$. When referring to a row-vector, such as the coordinates that identify the i^{th} observation in $\mathbf{X}$, I use lowercase, boldfaced, and italicised notation, e.g. $\boldsymbol{x}_i$.

Each observation in a training dataset consists of a pair $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i$ denotes a point in d-dimensional space, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})$, and y_i represents the corresponding outcome. I assume that samples are drawn i.i.d. from some fixed but unknown joint probability distribution $\mathbb{P}(\mathbf{Z}) = \mathbb{P}(\mathbf{X}, Y)$. Using n observations, an algorithm maps a dataset to a function, $a: \mathbf{Z} \rightarrow f$; the function in turn maps features to outcomes, $f: \mathbf{X} \rightarrow Y$. I consider both cases where Y is categorical (in which case f is a classifier) and where Y is continuous (in which case f is a regressor). I make no additional assumptions about the structure or properties of f.

Model f is judged by its ability to *generalise*, i.e. to accurately predict outcomes on test data sampled from $\mathbb{P}(\mathbf{Z})$ but not included in the training dataset. For a given test sample $\boldsymbol{x}_i$, we compute the predicted outcome $f(\boldsymbol{x}_i) = \hat{y}_i$ and observe the true outcome y_i. The hat symbol denotes that the value has been estimated. A model's performance is measured by a loss function L, which quantifies the distance between Y and $\hat{Y}$ over a set of test cases. The expected value of this loss function with respect to $\mathbb{P}(\mathbf{Z})$ for a given model f is called the *risk*:

$$R(f, \mathbf{Z}) = \mathbb{E}_{\mathbf{Z}}[L(f, \mathbf{Z})] \quad (5.1)$$

We estimate this population parameter with the empirical risk over a set of n samples:

$$R_{\text{emp}}(f, \mathbf{Z}) = \frac{1}{n} \sum_i L(f, \boldsymbol{z}_i) \quad (5.2)$$

A learning algorithm is said to be *consistent* if empirical risk converges to true risk as $n \to \infty$. A fundamental result of statistical learning theory states that an algorithm is consistent if and only if the space of functions it can learn is of finite VC dimension (Vapnik & Chervonenkis, 1971). This latter parameter is a capacity measure defined as the cardinality of the largest set of points the algorithm can shatter.[1] The finite VC dimension criterion will be important to define convergence conditions for the explanation game in §5.3.

Some philosophers have argued that statistical learning provides a rigorous foundation for all inductive reasoning (Corfield et al., 2009; Harman & Kulkarni, 2007). Although I am sympathetic to this position, none of the proceeding analysis depends upon this thesis.

§5.3.2 Causal Interventionism

Philosophers often distinguish between causal explanations (for natural events) and personal reasons (for human decisions). It is also common – though extremely misleading – to speak of algorithmic "decisions". Thus, we may be tempted to seek *reasons* rather than *causes* for algorithmic predictions, on the grounds that they are more decision-like than event-like. I argue that this is mistaken in several respects. First, the talk of algorithmic "decisions" is an anthropomorphic trope granting statistical models a degree of autonomy that dangerously downplays the true role of human agency in sociotechnical systems (Watson, 2019). Second, we may want to explain not just the top label selected by a classifier – the so-called "decision" – but also the complete probability distribution over possible labels. In a regression context, we may want to explain a prediction interval in addition to a mere point estimate. Finally, there are good pragmatic reasons to take a causal approach to this problem. As I argue in §5.4, it is relatively easy and highly informative to simulate the effect of causal interventions on supervised learning models, provided sufficient access.

Our approach therefore builds on the causal interventionist framework originally formalised by Pearl (2000) and Spirtes et al. (2000), and later given more philosophical treatment by Woodward (2003; 2008; 2010; 2015). A minimal explication of the theory runs as follows. X is a cause of Y within a given structural model $\mathcal{M}$ if and only if some hypothetical intervention on X (and no other variable) would result in a change in Y or the probability distribution of Y. This account is minimal in the sense that it places no constraints on $\mathcal{M}$ and imposes no causal efficacy thresholds on X or Y. The notion of an intervention is kept maximally broad to allow for any possible change in X, provided it does not alter the values of other variables in $\mathcal{M}$ except those that are causal descendants of X.

[1] The class of sets C shatters the set A if and only if for each $a \subset A$, there exists some $c \in C$ such that $a = c \cap A$. For more on VC theory, see (Vapnik, 1995; 1998). Popper's "degree of falsifiability" arguably anticipates the VC dimension. For a discussion, see (Corfield et al., 2009).

Under certain common assumptions,[2] Pearl's *do*-calculus provides a complete set of formal tools for reasoning about causal interventions (Huang & Valtorta, 2006). A key element of Pearl's notation system is the *do* operator, which allows us to denote, for example, the probability of Y, conditional on an intervention that sets variable X to value x, with the concise formula $\mathbb{P}\big(Y \big| do(X = x)\big)$. A structural causal model $\mathcal{M}$ is a tuple $\langle \mathbf{U}, \mathbf{V}, F, \mathbb{P}(\mathbf{u}) \rangle$ consisting of exogenous variables $\mathbf{U}$, endogenous variables $\mathbf{V}$, a set of functions F that map each V_j's causal parents to its observed values, and a probability distribution $\mathbb{P}$ over exogenous states. $\mathcal{M}$ may be visually depicted as a graph with nodes corresponding to variables and directed edges denoting causal relations between endogenous features (see Fig. 5.1). I restrict my attention here to directed acyclic graphs (DAGs), which are the focus of most work in causal interventionism.

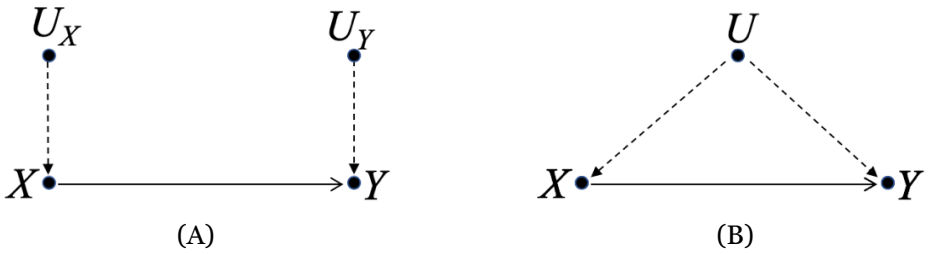

Figure 5.1. Two examples of simple causal models. (A) A Markovian graph. Two exogenous variables, U_X and U_Y, have unobserved causal effects on two endogenous variables, X and Y, respectively. (B) A semi-Markovian graph. A single exogenous variable, U, has unobserved confounding effects on two endogenous variables, X and Y.

If the model $\mathcal{M}$ contains no exogenous confounders, then $\mathcal{M}$ is said to be *Markovian*. In this case, factorisation of a graph's joint distribution is straightforward and causal effects can be computed directly from the data. However, when one or more unobserved variables has a confounding effect on two or more observed variables, as in Fig. 5.1B, then we say that $\mathcal{M}$ is *semi-Markovian*, and more elaborate methods are needed to estimate causal effects. Specifically, some sort of adjustment must be made by conditioning on an appropriate set of covariates. While several overlapping formulations have been proposed for such adjustments (Galles & Pearl, 1995; Pearl, 1995; Robins, 1997), I follow Tian & Pearl (2002), who provide a provably sound and complete set of causal identifiability conditions for semi-Markovian models (Huang & Valtorta, 2008; Shpitser & Pearl, 2008).

Their criteria are as follows. The causal effect of the endogenous variable V_j on all observed covariates $\mathbf{V}_{-j}$ is identifiable if and only if there is no consecutive sequence of confounding edges between V_j and V_j's immediate successors in the graph. Weaker conditions are

[2] The completeness of the *do*-calculus relies on the causal Markov and faithfulness conditions, which together state (roughly) that statistical independence implies graphical independence and vice versa. Neither assumption has gone unchallenged. I refer interested readers to (Hausman & Woodward, 2004) and (Cartwright, 2002) for a debate on the former; see (Cartwright, 2007) and (Weinberger, 2018) for a discussion of the latter.

sufficient when we focus on a proper subset $\mathbf{S} \subset \mathbf{V}$. In this case, $\mathbb{P}\left(\mathbf{S}\middle|do\left(V_j = v_{ij}\right)\right)$ is identifiable so long as there is no consecutive sequence of confounding edges between V_j and V_j's children in the subgraph composed of the ancestors of **S**.

I take it that the goal in most iML applications is to provide a causal explanation for one or more algorithmic outputs. Identifiability is therefore a central concern, and another key component to defining convergence conditions in §5.3. Fortunately, as I argue in §5.4.1, many cases of interest in this setting involve Markovian graphs, and therefore need no covariate adjustments. Semi-Markovian alternatives are considered in §5.5.2.2, although guarantees cannot generally be provided in such instances without additional assumptions.

If successful, a causal explanation for some algorithmic prediction(s) will accurately answer a range of what Woodward calls "what-if-things-had-been-different questions" (henceforth *w*-questions). For instance, we may want to know what feature(s) about an individual caused her loan application to be denied. What if she had been wealthier? Or older? Would a hypothetical applicant identical to the original except in terms of along the axis of wealth or age have had more luck? Several authors in the iML literature explicitly endorse such a counterfactual strategy (Kusner et al., 2017; Wachter et al., 2018). I shall revisit these methods in §5.4.6.

	c_1: Rain	c_2: No rain
a_1: Umbrella	1	−1
a_2: No umbrella	−2	0

Table 5.1. Utility matrix for Jones when deciding whether or not to pack his umbrella.

§5.3.3 Decision Theory

Decision theory provides formal tools for reasoning about choices under uncertainty. These will prove useful when attempting to quantify explanatory relevance in §5.5.2.3. I assume the typical setup, in which an individual considers finite set of actions A and a finite set of outcomes C. According to expected utility theory,[3] an agent's rational preferences may be expressed as a utility function u that maps the Cartesian product of A and C to the real numbers, $u: A \times C \rightarrow \mathbb{R}$. For instance, Jones may be unsure whether to pack his umbrella today. He could do so (a_1), but it would add considerable bulk and weight to his bag; or he could leave it at home (a_2) and risk getting wet. The resulting utility matrix is depicted in Table 5.1. The rational choice for Jones depends not just on his utility function u but also on his beliefs about

[3] The von Neumann-Morgenstern representation theorem guarantees the uniqueness (up to affine transformation) of the rational utility function u, provided an agent's preferences adhere to the following four axioms: completeness, transitivity, independence of irrelevant alternatives, and continuity. For the original derivation, see (von Neumann & Morgenstern, 1944).

whether or not it will rain. These are formally expressed by a (subjective) probability distribution over C, $\mathbb{P}(C)$. We compute each action's expected utility by taking a weighted average over outcomes:

$$\mathbb{E}_C[u(a_i, C)|E] = \sum_j \mathbb{P}(c_j|E)u(a_i, c_j) \quad (5.3)$$

where the set of evidence E is either empty (in which case Eq. 5.3 denotes a prior expectation) or contains some relevant evidence (in which case Eq. 5.3 represents a posterior expectation). Posterior probabilities are calculated in accordance with Bayes's theorem:

$$\mathbb{P}(c_i|E) = \frac{\mathbb{P}(E|c_i)\mathbb{P}(c_i)}{\mathbb{P}(E)} \quad (5.4)$$

which follows directly from the Kolmogorov axioms for the probability calculus (Kolmogorov, 1950). By solving Eq. 5.3 for each element in A, we identify at least one utility-maximising action:

$$a^* = \underset{a_i \in A}{\operatorname{argmax}}\, \mathbb{E}_C[u(a_i, C)|E] \quad (5.5)$$

An ideal epistemic agent always selects (one of) the optimal action(s) a^* from a set of alternatives.

It is important to note how a rational agent's beliefs interact with his utilities to guide decisions. If Jones is maximally uncertain about whether or not it will rain, then he assigns equal probability to both outcomes, resulting in expected utilities of

$$\mathbb{E}_C[u(a_1, C)] = 0.5(1) + 0.5(-1) = 0$$

and

$$\mathbb{E}_C[u(a_2, C)] = 0.5(-2) + 0.5(0) = -1,$$

respectively. In this case, Jones should pack his umbrella. But say he gains some new information E that changes his beliefs. Perhaps he sees a weather report that puts the chance of rain at just 10%. Then he will have the following expected utilities:

$$\mathbb{E}_C[u(a_1, C)|E] = 0.1(1) + 0.9(-1) = -0.8$$
$$\mathbb{E}_C[u(a_2, C)|E] = 0.1(-2) + 0.9(0) = -0.2$$

In this case, leaving the umbrella at home is the optimal choice for Jones.

Of course, humans can be notoriously irrational. Experiments in psychology and behavioural economics have shown time and again that people rely on heuristics and cognitive biases instead of consistently applying the axioms of decision theory or probability calculus (Kahneman, 2011). Thus, the concepts and principles outlined here are primarily normative. They prescribe an optimal course of behaviour, a sort of Kantian regulative ideal when utilities and probabilities are precise, and posterior distributions are properly calculated. For the practical purposes of iML, these values may be estimated via a hybrid system in which software aids an inquisitive individual with bounded rationality. I shall revisit these issues in §5.7.1.

§5.4 Scope

Supervised learning algorithms provide some unique affordances that differentiate iML from more general explanation tasks. This is because the target in iML is not the natural or social phenomenon the algorithm is designed to predict, but rather *the algorithm itself*. In other words, we are interested not in the underlying joint distribution $P(\mathbf{Z}) = P(\mathbf{X}, Y)$, but in the estimated joint distribution $P(\mathbf{Z}_f) = P(\mathbf{X}, \hat{Y})$. The distinction is crucial.

Strevens (2013) differentiates between three modes of understanding: *that*, *why*, and *with*.[4] Understanding *that* some proposition p is true is simply to be aware that p. Understanding *why* p is true requires some causal explanation for p. Strevens's third kind of understanding, however, applies only to theories or models. Understanding *with* a model amounts to knowing how to apply it in order to predict or explain real or potential phenomena. For instance, a physicist who uses Newtonian mechanics to explain the motion of billiard balls thereby demonstrates her ability to understand *with* the theory. Since this model is strictly speaking false, it would be incorrect to say that her explanation provides a true understanding of *why* the billiard balls move as they do. (Of course, she could be forgiven for sparing her poolhall companions the relativistic details of metric tensors and spacetime curvature in this case.) Yet our physicist has clearly understood something – namely the Newtonian theory itself – even if the classical account she offers is inaccurate or incomplete. Similarly, the goal in iML is to help epistemic agents understand *with* the target model f, independent of whatever realities f was intended to capture. The situation is slightly more complicated in the case of discovery. The strategy here is to use understanding *with* as an indirect path to understanding *why*, on the assumption that if model f performs well then it has probably learned some valuable information about the target system.

Despite the considerable complexity of some statistical models, as a class they tend to be *complete*, *precise*, and *forthcoming*. These three properties simplify the effort to explain any complex system.

§5.4.1 Complete

Model f is complete with respect to the input features $\mathbf{X}$ in the sense that exogenous variables have no influence whatsoever on predicted outcomes $\hat{Y}$. Whereas nature is full of unobserved confounders that may complicate or undermine even a well-designed study, fitted models are self-contained systems impervious to external variation. They therefore instantiate Markovian, rather than semi-Markovian graphs. This is true even if dependencies between predictors

[4] In what follows, I take it more or less for granted that explanations promote understanding and that understanding requires explanations. Both claims have been disputed. For a discussion, see (de Regt et al., 2009; Grimm, 2006; Khalifa, 2012). I revisit the relationship between these concepts in §5.7.3.

are not explicitly modelled, in which case we may depict f as a simple DAG with directed edges from each feature $X_1, \ldots, X_d$ to $\hat{Y}$.

In what follows, I presume that the agents in question know which variables were used to train f. This may not always be the case in practice, and without such knowledge it becomes considerably more difficult to explain algorithmic predictions. Whatever the epistemic status of the inquiring agent(s), however, the underlying model itself remains complete.

Issues arise when endogenous variables serve as proxies for exogenous variables. For instance, a model may not explicitly include a protected attribute such as race, but instead use a seemingly innocuous covariate like zip code, which is often a strong predictor of race (Datta et al., 2017). In this case, an intervention that changes a subject's race will have no impact on model f's predictions unless we take the additional step of embedding f in a larger causal structure $\mathcal{M}$ that includes a directed edge from race to zip code. I consider possible strategies for resolving problems of this nature in §5.5.2.2.

§5.4.2 Precise

Model f is precise in the sense that it always returns the same output for any particular set of inputs. Whereas a given experimental procedure may result in different outcomes over repeated trials due to irreducible noise, a fitted model has no such internal variability. Some simulation-based approaches, such as the Markov chain Monte Carlo methods widely used in Bayesian data analysis (Gelman et al., 2014), pose a notable exception to this rule. These models make predictions by random sampling, a stochastic process whose final output is a posterior distribution, not a point estimate. However, if the model has converged, then these predictions are still precise in the limit. As the number of draws from the posterior grows, statistics of interest (e.g., the posterior mode or mean) stabilise to their final values. The Monte Carlo variance of a given parameter can be bounded as a function of the sample size using well-known concentration inequalities (Boucheron et al., 2013).

Woodward (2003; 2010) emphasises the role of "stability" in causal generalisations, a concept that bears perhaps some superficial resemblance to what I here call precision. The difference is that stability in Woodward's sense can only be applied to a proper subset of the edges (usually just a single edge) in a causal graph. The generalisation that "variable X causes variable Y" is *stable* to the extent that it persists across a wide range of background conditions, i.e. alternative states of the model $\mathcal{M}$ and/or input distributions for X (Arjovsky et al., 2019; Bühlmann, 2020; Peters, Bühlmann, & Meinshausen, 2016). Precision in this case requires completeness, because it applies only to the causal relationship between the set of all predictors $\mathbf{X}$ and the outcome Y, which is strictly deterministic at the token level.

§5.4.3 Forthcoming

Model f is forthcoming in the sense that it will always provide an output for any well-formed input. Moreover, it is typically quite fast and cheap to query an algorithm in this way. Whereas experiments in the natural or social sciences can often be time-consuming, inconclusive, expensive, or even dangerous, it is relatively simple to answer w-questions in supervised learning contexts. In principle, an analyst could even recreate the complete joint distribution $\mathbb{P}(\mathbf{X}, \hat{Y})$ simply by saturating the feature space with w-questions. Of course, this strategy is computationally infeasible with continuous predictors and/or a design matrix of even moderate dimensionality.

Supervised learning algorithms may be less than forthcoming when shielded by intellectual property (IP) laws, which can also prevent researchers from accessing a model's complete list of predictors. In lieu of an open access programming interface, some iML researchers resort to reverse engineering algorithms from training datasets with known predicted values. This was the case, for instance, with a famous ProPublica investigation into the COMPAS algorithm, a proprietary model used by courts in several US states to predict the risk of criminal recidivism (Angwin et al., 2016; Larson et al., 2016). Subsequent studies using the same dataset reached different conclusions regarding the algorithm's reliance on race (Fisher et al., 2019; Rudin et al., 2018), highlighting the inherent uncertainty of model reconstruction when the target algorithm is not forthcoming. In what follows, I focus on the ideal case in which agents face no IP restrictions.

§5.5 The Explanation Game

In this section, I introduce a formal framework for iML. The proposal takes the form of a game in which an inquisitor (call her Alice) seeks an explanation for an algorithmic prediction $f(\boldsymbol{x}_i) = \hat{y}_i$. Note that our target (at this stage) is a *local* or *token* explanation, rather than a *global* or *type* explanation. In other words, Alice wants to know why this particular input resulted in that particular output, as opposed to the more general task of recreating the entire decision boundary or regression surface of f.

Unfortunately for Alice, f is a black box. But she is not alone. She is helped by a devoted accomplice (call him Bob), who does everything in his power to aid Alice in understanding $\hat{y}_i$. Bob's goal is to get Alice to a point where she can correctly predict f's outputs, at least in the neighbourhood of $\boldsymbol{x}_i$ and within some tolerable margin of error. In other words, he wants her to be able to give true answers to relevant w-questions about how f would respond to hypothetical datapoints near $\boldsymbol{x}_i$.

I make several nontrivial assumptions about Alice and Bob, some of which were foreshadowed above. Specifically:

- Alice is a rational agent. Her preferences over alternatives are complete and transitive, she integrates new evidence through Bayesian updating, and she does her best to maximise expected utility subject to constraints on her cognitive/computational resources.
- Bob is Alice's accomplice. He has data on the features $\mathbf{V} = (X_1, \ldots, X_d, \hat{Y})$ that are endogenous to f, as well a (possibly empty) set of exogenous variables $\mathbf{U} = (X_{d+1}, \ldots, X_{d+m})$ that are of potential interest to Alice. He may query f with any well-formed input at little or no cost.

One could easily envision more complex explanation games in which some or all of these assumptions are relaxed. Future work will examine such alternatives.

§5.5.1 Three Desiderata

According to Woodward (2003, p. 203), the following three criteria are individually necessary and jointly sufficient to explain some outcome of interest $Y = y$ that obtains when $X = x$ within a given structural model $\mathcal{M}$:

(iv) The generalisations described by $\mathcal{M}$ are accurate, or at least approximately so, as are the observations $Y = y$ and $X = x$.

(v) According to $\mathcal{M}$, $Y = y$ under an intervention that sets X to x.

(vi) There exists some possible intervention that sets X to x' (where $x \neq x'$), with $\mathcal{M}$ correctly describing the value y' (where $y \neq y'$) that Y would assume under the intervention.

This theory poses no small number of complications that are beyond the scope of this paper.[5] I adopt the framework as a useful baseline for analysis, as it is sufficiently flexible to allow for extensions in a number of directions.

§5.5.1.1 Accuracy

Woodward's account places a well-justified premium on explanatory accuracy. Any explanation that fails to meet criteria (i)-(iii) is not deserving of the name. However, this theory does not tell the whole story. To see why, consider a deep convolutional neural network f trained to classify images. The model correctly predicts that $\boldsymbol{x}_i$ depicts a cat. Alice would like to know why. Bob attempts to explain the prediction by writing out the complete formula for f. The neural network contains some hundred layers, each composed of 1 million parameters that together describe a complex nonlinear mapping from pixels to labels. Bob checks against Woodward's criteria and observes that his model $\mathcal{M}$ is accurate, as are the input and output values; that $\mathcal{M}$ correctly predicts the output given the input; and that interventions on the

[5] For book length treatments, see (Halpern, 2016; Strevens, 2010; Woodward, 2003). For relevant articles, see, e.g., (Franklin-Hall, 2014; Kinney, 2018; Potochnik, 2015; Weslake, 2010; Woodward & Hitchcock, 2003).

original photograph replacing the cat with a dog do in fact change the predicted label from "cat" to "dog".

Problem solved? Not quite. Bob's causal graph $\mathcal{M}$ is every bit as opaque as the underlying model f. In fact, the two are identical. So while this explanation may be maximally accurate, it is far too complex to be of any use to Alice. The result is not unlike the map of Borges's famous short story (1946), in which imperial cartographers aspire to such exactitude that they draw their territory on a 1:1 scale. Black box explanations of this sort create a kind of Chinese room (Searle, 1980), in which the inquiring agent is expected to manually perform the algorithm's computations in order to trace the path from input to output. Just as the protagonist of Searle's thought experiment has no understanding of the Chinese characters he successfully manipulates, so Alice gains no explanatory knowledge about f by instantiating the model herself. Unless she is comfortable computing high-dimensional tensor products on the fly, Alice cannot use $\mathcal{M}$ to build a mental model of the target system f or its behaviour near $\boldsymbol{x}_i$. She cannot answer relevant w-questions without consulting the program, which will merely provide her with new labels that are as unexplained as the original.

§5.5.1.2 Simplicity

Accuracy is a necessary but insufficient condition for successful explanation, especially when the underlying system is too complex for the inquiring agent to fully comprehend. In these cases, we tend to value *simplicity* as an inherent virtue of candidate explanations. The point is hardly novel. Simplicity has been cited as a primary goal of scientific theories by practically everyone who has considered the question (cf. Baker, 2016). The point is not lost on iML researchers, who typically impose sparsity constraints on possible solutions to ensure a manageable number of nonzero parameters (e.g., Angelino et al., 2018; Ribeiro et al., 2016; Wachter et al., 2018).

It is not always clear just what explanatory simplicity amounts to in algorithmic contexts. One plausible candidate, advocated by Popper (1959), is based on the number of free parameters. In statistical learning theory, this proposal has largely been superseded by capacity measures like the aforementioned VC dimension or Rademacher complexity. These parameters help to establish a syntactic notion of simplicity, which has proven especially fruitful in statistics. Yet such definitions obscure the semantic aspect of simplicity, which is probably of greater interest to epistemic agents like Alice. The kind of simplicity required for her to understand why $f(\boldsymbol{x}_i) = \hat{y}_i$ depends not just upon the functional relationships between the units of explanation, but more importantly upon the explanatory level of abstraction (Floridi, 2008a) – i.e., the choice of units themselves.

Rather than adjudicate between the various competing notions of simplicity that abound in the literature, I opt for a purely relational approach upon which simplicity is just

equated with *intelligibility for Alice*. I am unconvinced that there is any sense to be made of an absolute, mind-independent notion of simplicity. Yet even if there is, it would be of little use to Alice to insist that explanation g_1 is simpler than g_2 on our preferred definition of the term, despite the empirical evidence that she understands the implications of the latter better than the former. What is simple for some agents may be complex for others, depending on background knowledge and contextual factors. In §5.5.2, I operationalise this observation by measuring simplicity in explicitly agentive terms.

§5.5.1.3 Relevance

Some may judge accuracy and simplicity to be sufficient for successful explanation, and in many cases they probably are. But there are important exceptions to this generalisation. Consider, for example, the following case. A (bad) bank issues loans according to just two criteria: applicants must be either white or wealthy. This bank operates in a jurisdiction in which race alone is a protected attribute. A poor black woman named Alice is denied a loan and requests an explanation. The bank informs her that her application was denied due to her finances. This explanation is accurate and simple. However, it is also disingenuous – for it would be just as accurate and simple to say that her loan was denied because of her race, a result that would be of far greater relevance both to Alice and state regulators. Given Alice's interests, the latter explanation is superior to the former, yet the bank's explanation has effectively eclipsed it.

This is a fundamental observation: among the class of accurate and simple explanations, some will be more or less relevant to the inquiring agent (Floridi, 2008b). Alice has entered into this game for a reason. Something hangs in the balance. Perhaps she is a loan applicant deciding whether to sue a bank, or a doctor deciding whether to trust an unexpected diagnosis. A successful explanation will not only need to be accurate and simple; it must also inform her decision about how best to proceed. Otherwise, we have a case of *counterfactual eclipse*, in which an agent's interests are overshadowed by a narrow focus on irrelevant facts that do nothing to advance her understanding or help modify future behaviours.

The problem of *counterfactual eclipse* is a serious issue in any context where customers or patients, for example, may wish to receive (or perhaps exercise their right to) an explanation. However, I am unaware of any proposal in the iML literature that explicitly protects against this possibility.

§5.5.2 Rules of the Game

Having motivated an emphasis on accuracy, simplicity, and relevance, I now articulate formal constraints that impose these desiderata on explanations in iML. A schematic overview of the explanation game is provided in pseudocode.

Algorithm 5.1: The Explanation Game

Inputs:

Environment: supervised learner f, endogenous variables $\mathbf{V}$, data $D \sim \mathbb{P}(\mathcal{M})$ possibly including exogenous covariates $\mathbf{U}$

Alice: explanandum $f(\boldsymbol{x}_i) = \hat{y}_i$, contrastive outcome $f(\boldsymbol{x}_i) \neq \tilde{y}_i$, level of abstraction LoA, choice set A, causal hypotheses C, utility function u, prior distribution over causal hypotheses $\mathbb{P}(C)$, function space $\mathcal{H}$, loss function $L_{\mathcal{H}}$

Bob: set of B unique function spaces $\mathcal{G}_b$, loss function $L_{\mathcal{G}}$, kernel $k_{\mathcal{G}}$. If exogenous variables are relevant, then an additional function space $\mathcal{G}'$, loss function $L_{\mathcal{G}'}$, kernel $k_{\mathcal{G}'}$

for each round:

(1) Bob creates a map $\psi: \mathcal{Z}_f \rightarrow \mathcal{Z}_g$ from the original f-space to an explanatory g-space designed to (a) shift the input distribution to Alice's desired LoA and (b) help provide evidence for or against at least one hypothesis in C. Whereas $\mathbf{Z}_f = (\mathbf{X}, \hat{Y})$, $\mathbf{Z}_g = (\mathbf{X}', Y')$.

if $\mathbf{X}'$ includes variables $\mathbf{U}$ that are exogenous to f:

(2) Bob trains the model $g': \mathbf{V} \rightarrow \mathbf{U}$, optionally fit using kernel $k_{\mathcal{G}'}$, to minimize loss $L_{\mathcal{G}'}$ over function space $\mathcal{G}'$.

(3) Bob creates a training dataset by sampling points $\boldsymbol{v}_s$ from a distribution centred at $\boldsymbol{v}_i$ and repeatedly querying g' with w-questions of the form $\mathbb{E}_{\mathcal{M}}[\mathbf{U}|do(\mathbf{V} = \boldsymbol{v}_s)] = ?$ The resulting data are mapped to g-space via ψ.

end if

for each function space $\mathcal{G}_b$:

(4) Bob creates a training dataset by sampling points $\boldsymbol{x}_s$ from a distribution centred at $\boldsymbol{x}_i$ and repeatedly querying f with w-questions of the form $\mathbb{E}_{\mathbf{Z}_f}[Y|do(\mathbf{X} = \boldsymbol{x}_s)] = ?$ The resulting data are mapped to g-space via ψ.

(5) Bob trains a model $g: \mathbf{X}' \rightarrow Y'$, optionally fit using kernel $k_{\mathcal{G}}$, to minimise loss $L_{\mathcal{G}}$ over function space $\mathcal{G}_b$. Empirical risk is calculated in f-space via the inverse mapping ψ^{-1}, optionally weighted by $k_{\mathcal{G}}$.

(6) Alice creates a training dataset by repeatedly querying g with w-questions of the form $\mathbb{E}_{\mathbf{Z}_g}[Y'|do(X'_j = x'_{ij})] = ?$ Bob reports both the predicted outcome and the empirical risk.

(7) Alice trains a model $h: \mathbf{X}' \rightarrow Y'$ to minimise loss $L_{\mathcal{H}}$ over function space $\mathcal{H}$. Empirical risk is optionally weighted by $k_{\mathcal{G}}$ and estimated in g-space.

(8) The information Alice learns from and about g and h constitutes a body of evidence E, which she uses to update her beliefs regarding C.

(9) Alice calculates the posterior expected utility of each action in A, producing at least one optimal choice a^*.

Outputs: $R_{\text{emp}}(g, \mathbf{Z}_f), R_{\text{emp}}(h, \mathbf{Z}_g), \mathbb{E}_C[u(a^*, C)|E]$

end for

end for

This game has a lot of moving parts, but at its core the process is quite straightforward. Essentially, Bob does his best to proffer an accurate explanation in terms that Alice can understand. She learns by asking w-questions until she feels confident enough to answer such questions herself. The result is scored by three measures: accuracy (error of Bob's model), simplicity (error of Alice's model), and relevance (expected utility for Alice). Note that all explanations are indexed by their corresponding map ψ and explanatory function space $\mathcal{G}_b$. I suppress the dependency for notational convenience. All inputs and steps are discussed in greater detail below.

§5.5.2.1 Inputs

Alice must specify a contrastive outcome $f(\boldsymbol{x}_i) \neq \tilde{y}_i \in Y$. This counterfactual alternative may represent Alice's initial expectation or desired response. Consider, for example, a case in which f is trained to distinguish between handwritten digits, a classic benchmark problem in ML commonly referred to as MNIST, after the most famous database of such images.[6] Say f misclassifies $\boldsymbol{x}_i$ as a "7", when in fact y_i = "1". Alice wants to know not just why the model predicted "7", but also why it did *not* predict "1". Specifying an alternative $\tilde{y}_i$ is important, as it focuses Bob's attention on relevant regions of the feature space. An explanation such as "Because $\boldsymbol{x}_i$ has no closed loops" may explain why f did not predict "8" or "9", but that is of little use to Alice, as it eclipses the relevant explanation. The importance of contrastive explanation is highlighted by several philosophers (Blaauw, 2013; Hitchcock, 1999; Potochnik, 2015; van Fraassen, 1980), and has recently begun to receive attention in the iML literature as well (Miller, 2019; Mittelstadt et al., 2019).

Alice must specify some desired level of abstraction (LoA). The LoA specifies a set of typed variables and observables that are used to describe a system. Inspired by the Formal Methods literature in computer science (Boca et al., 2010), the levelist approach has been extended to conceptualise a wide array of problems in the philosophy of information (Floridi, 2008a; 2011; 2017). Alice's desired LoA will help Bob establish the preferred units of explanation, a crucial step toward ensuring intelligibility for Alice. In the MNIST example, Alice is unlikely to seek explanations at the pixel-LoA, but may be satisfied with a higher LoA that deals in curves and edges.

Pragmatism demands that Alice have some notion why she is playing this game. Her choices A, preferences u, and beliefs $\mathbb{P}(C)$ will guide Bob in his effort to supply a satisfactory explanation and constrain the set of possible solutions. The MNIST example is a case of iML for validation, in which Alice's choice set may include the option to deploy or not deploy the model f. Her degrees of belief with respect to various causal hypotheses are determined by her expertise in the data and model. Perhaps it is well known that algorithms struggle to differentiate between "7" and "1" when the former appears without a horizontal line through the digit. The cost of such a mistake is factored into her utility function.

Bob, for his part, enters into the game with three key components: (i) a set of $B \geq 1$ candidate algorithms for explanation; (ii) a loss function with which to train these algorithms; and (iii) a corresponding kernel. Popular options for (i) include sparse linear models and rule lists. The loss function is left unspecified, but common choices include mean squared error for regression and cross-entropy for classification. The kernel tunes the locality of the explana-

[6] The Modified National Institute of Standards and Technology database contains 60,000 training images and 10,000 test images, each 28 × 28 pixel grayscale photos of digits hand-written either by American high school students or United States Census Bureau employees. See http://yann.lecun.com/exdb/mnist/.

tion, weighting observations by their distance from the original input $\boldsymbol{x}_i$, as measured by some appropriate metric. Whether the kernel is used to train the model g or simply evaluate g's empirical risk is left up to Bob. Abandoning the kernel altogether results in a global explanation, with no particular emphasis on the neighbourhood of $\boldsymbol{x}_i$.

Bob may need an additional algorithm, loss function, and kernel to estimate the relationship between endogenous and exogenous features. If so, there is no obvious requirement that such a model be intelligible to Alice or Bob, so long as it achieves minimal predictive error.

§5.5.2.2 Mapping the Space

Perhaps the most consequential step in the entire game is Bob's mapping $\psi: \mathcal{Z}_f \to \mathcal{Z}_g$. In an effort to provide a successful explanation for Alice, Bob projects the input distribution $\mathbb{P}(\mathbf{Z}_f) = \mathbb{P}(\mathbf{X}, \hat{Y})$ into a new space $\mathbb{P}(\mathbf{Z}_g) = \mathbb{P}(\mathbf{X}', Y')$. The change in the response variable is set by Alice's contrastive outcome of interest. In the MNIST example, Bob maps the original 10-class variable $\hat{Y}$ onto a binary variable Y' indicating whether or not inputs are classified as "1". The contents of $\mathbf{X}'$ may be iteratively established by considering Alice's desired LoA and hypothesis set C. This will often amount to a reduction of the feature space. For instance, Bob may coarsen a set of genes into a smaller collection of biological pathways (Sanguinetti & Huynh-Thu, 2018), or transform pixels into super-pixels (Stutz, Hermans, & Leibe, 2018).

Alternatively, Bob may need to expand the input features to include exogenous variables hypothesised to be relevant to the outcome. In this case, he will require external data D sampled from the expanded feature space $\mathbb{P}(\mathcal{M})$, which can be used to train one or more auxiliary models to predict values for the extra covariate(s) in unobserved regions of g-space. For instance, when an algorithm is suspected of encoding protected attributes like race via unprotected attributes like zip code, Bob will need to estimate the dependency using a new function g' that predicts the former based on the latter (along with any other relevant endogenous variables). Note that in this undertaking, Bob is essentially back to square one. The target $\mathcal{M}$ is presumably not complete, precise, or forthcoming, and his task therefore reduces to the more general problem of modelling some complex natural or social system with limited information. This inevitably introduces new sources of error that will have a potentially deleterious impact on downstream results. Depending on the structural properties of the underlying causal graph, effects of interventions in g-space may not be uniquely identifiable.

In any event, the goal at this stage is to make the input features sufficiently intelligible to Alice that they can accommodate her likely w-questions and inform her beliefs about causal hypotheses C. General purpose methods for causal feature learning have been proposed (Chalupka et al., 2017), however, critics have persuasively argued that such procedures cannot be implemented in a context-independent manner (Kinney, 2018). Some areas of research, such as bioinformatics and computer vision, have well-established conventions on how to

coarsen high-dimensional feature spaces. Other domains may prove more challenging. Accessibility to external data on exogenous variables of interest will likewise vary from case to case. Even when such datasets are readily available, there is no guarantee that the functional relationships sought can be estimated with high accuracy or precision. As in any other explanatory context, Alice and Bob must do the best they can with their available resources and knowledge.

§5.5.2.3 Building Models, Scoring Explanations

Once ψ is fixed, the next steps in the explanation game are effectively supervised learning problems. This puts at Alice and Bob's disposal a wide range of well-studied algorithms and imports the corresponding statistical guarantees.

Bob creates a training dataset of $\mathbf{Z}_g = (\mathbf{X}', Y')$ and fits a model g from the explanatory function space $\mathcal{G}_b$. Alice explores g-space by asking a number of w-questions that posit relevant interventions. For instance, she may want to know if the presence of a horizontal line through the middle of a numeral determines whether f predicts a "7". If so, then this will be a hypothesis in C and we should find a corresponding variable in $\mathbf{X}'$. Because the target model f and/or Bob's explanation g may involve implicit or explicit structural equations, we use the *do*-calculus to formalise such interventions.

Bob and Alice can select whatever combination of loss function and algorithm makes the most sense for their given explanation task. g's error is measured by $R_{\text{emp}}(g, \mathbf{Z}_f)$; g's complexity is measured by $R_{\text{emp}}(h, \mathbf{Z}_g)$. We say that g is ε_1-accurate if $R_{\text{emp}}(g, \mathbf{Z}_f) \leq \varepsilon_1$ and ε_2-simple if $R_{\text{emp}}(h, \mathbf{Z}_g) \leq \varepsilon_2$. The content and performance of g and h constitute a body of evidence E, which Alice uses to update her beliefs about causal hypotheses C. Relevance is measured by the posterior expected utility of the utility-maximising action, $\mathbb{E}_C[u(a^*, C)|E]$. (For consistency with the previous desiderata, we in fact measure *irrelevance* by multiplying the relevance by -1.) Bob's explanation is ε_3-relevant to Alice if $-\mathbb{E}_C[u(a^*, C)|E] \leq \varepsilon_3$.

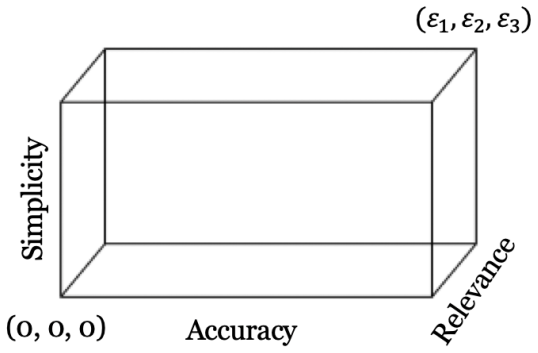

Figure 5.2. The space of satisfactory explanations is delimited by upper bounds on the error (ε_1), complexity (ε_2), and irrelevance (ε_3) of explanations Alice is willing to accept.

Explanations generated by this game may now be located in three-dimensional space, with axes corresponding to accuracy, simplicity, and relevance. An explanation is deemed *satisfactory* if it does not exceed preselected values of $\varepsilon_1, \varepsilon_2$, and ε_3. These parameters can be interpreted as budgetary constraints on Alice and Bob. How much inaccuracy, complexity, and irrelevance can they afford? I assign equal weight to all three criteria here, but relative costs could easily be quantified through a differential weighting scheme. Together, these points define the extremum of a cuboid, whose opposite diagonal is the origin (see Fig. 5.2). Any point falling within this cuboid is $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$-satisfactory.

§5.5.3 Consistency and Convergence

The formal tools of statistical learning, causal interventionism, and decision theory provide all the ingredients needed to state the necessary and sufficient conditions for convergence to a conditionally optimal explanation surface in polynomial time.

I define optimality in terms of a Pareto frontier. One explanation Pareto-dominates another if and only if it is strictly better along at least one axis and no worse along any other axis. If Alice and Bob are unable to improve upon the accuracy, simplicity, or relevance of an explanation without incurring some loss along another dimension, then they have found a Pareto-dominant explanation. A collection of such explanations constitutes a Pareto frontier, a surface of explanations from which Alice may choose whichever best aids her understanding and serves her interests. Note that this is a relatively weak notion of optimality. Explanations may be optimal in this sense without even being satisfactory, since the entire Pareto frontier may lie beyond the satisfactory cuboid defined by $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$. In this case, Alice and Bob have two options: (a) accept that no explanation will satisfy the criteria and adjust thresholds accordingly; or (b) start a new round with one or several different input parameters. Option (b) will generate entirely new explanation surfaces for the players to explore.

Without more information about the target function f or specific facts about Alice's knowledge and interests, conditional Pareto dominance is the strongest form of optimality one can reasonably expect. Convergence on a Pareto frontier is almost surely guaranteed on three conditions:

- *Condition 1*. The function spaces $\mathcal{G}_b$ and $\mathcal{H}$ are of finite VC dimension.
- *Condition 2*. Answers to all w-questions are uniquely identifiable.
- *Condition 3*. Alice is a rational agent and consistent Bayesian updater.

Condition (1) entails the statistical consistency of Bob's model g and Alice's model h, which ensures that accuracy and simplicity are reliably measured as sample size grows. Condition (2) entails that simulated datasets are faithful to their underlying data generating processes, thereby ensuring that g and h converge on the right targets. Condition (3) entails the existence of at least one utility-maximising action $a^* \in A$ with well-defined posterior expectation. If her

probabilities are well-calibrated, then Alice will tend to pick the "right" action, or at least an action with no superior alternative in A. With these conditions in place, each round of the game will result in an explanation that cannot be improved upon without altering the input parameters.

If all subroutines of the game's inner loops execute in polynomial time, then the round will execute in polynomial time as well. The only potentially NP-hard problem is finding an adequate map ψ, which cannot be efficiently computed without some restrictions on the solution set. A naïve approach would be to consider all possible subsets of the original feature space, but even in the Markovian setting this would result in an unmanageable 2^d maps, where d represents the dimensionality of the input matrix $\mathbf{X}$. Efficient mapping requires some principled method for restricting this space to just those of potential interest for Alice. The best way to do so for any given problem is irreducibly context-dependent.

§5.6 Discussion

Current iML proposals do not instantiate the explanation game in any literal sense. However, our framework can be applied to evaluate the merits and shortcomings of existing methods. It also provides a platform through which to conceptualise the constraints and requirements of any possible iML proposal, illuminating the contours of the solution space.

The most popular iML methods in use today are local linear approximators like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). The former explains predictions by randomly sampling around the point of interest. Observations are weighted by their distance from the target point and a regularised linear model is fit by weighted least squares. The latter builds on foundational work in cooperative game theory, using training data to efficiently compute pointwise approximations of each input feature's Shapley value.[7] The final result in both cases is a (possibly sparse) set of coefficients indicating the positive or negative association between input features and the response, at least near $\boldsymbol{x}_i$ and conditional on the covariates.

Using LIME or SHAP basically amounts to restricting the function space of Bob's explanation model g to the class of regularised linear models. Each method has its own default kernel k, as well as recommended mapping functions ψ for particular data types. For instance, LIME coarsens image data into super-pixels, while SHAP uses saliency maps to visualise the portions of an input image that were most important in determining its classification. While the authors of the two methods seem to suggest that a single run of either algorithm is sufficient for explanatory purposes, local linear approximations will tend to be unstable for datapoints near especially nonlinear portions of the decision boundary or regression surface.

[7] Shapley values were originally designed to fairly distribute surplus across a coalition of players in cooperative games (Shapley, 1953). They are the unique solution to the attribution problem that satisfies certain desirable properties, including local accuracy, missingness, and consistency. Directly computing Shapley values is NP-hard, however numerous approximations have been proposed. See (Sundararajan & Najmi, 2019) for an overview.

Thus, multiple runs with perturbed data may be necessary to establish the precision of estimated feature weights. This corresponds to multiple rounds of the explanation game, thereby giving Alice a more complete picture of the model space.

One major problem with LIME and SHAP is that neither method allows users to specify a contrast class. The default behaviour of both algorithms is to explain why an outcome is $\hat{y}_i$ as opposed to $\bar{y}$ – that is, the mean response for the entire dataset (real or simulated). In many contexts, this makes sense. For instance, if Alice receives a rare and unexpected diagnosis, then she may want to know what differentiates her from the majority of patients. However, it seems strange to suggest, as these algorithms implicitly do, that "normal" predictions are inexplicable. There is nothing confusing or improper about Alice wondering, for example, why she received an average credit score instead of a better-than-average one. Yet in their current form, neither LIME nor SHAP can accommodate such inquiries.

More flexible alternatives exist. Rule lists, which predict outcomes through a series of if-then statements, can model nonlinear effects that LIME and SHAP are incapable of detecting in principle. Several iML solutions are built on recursive partitioning (Guidotti et al., 2018; Ribeiro et al., 2018; Yang et al., 2017) – the statistical procedure that produces rule lists – and a growing number of psychological studies suggests that users find such explanations especially intelligible (Lage et al., 2018). If Alice is one of the many people who shares this preference for rule lists, then Bob should take this into account when selecting $\mathcal{G}_b$.

Counterfactual explanations are endorsed by Wachter et al. (2018), who propose a novel iML solution based on generative adversarial networks (GANs). Building on pioneering research in deep learning (Goodfellow et al., 2014), the authors demonstrate how GANs can be used to find the minimal perturbation of input features sufficient to alter the output in some prespecified manner. These models are less restrictive than linear regressions or rule lists, as they not only allow users to identify a contrast class but can in principle adapt to any differentiable function. Wachter et al. emphasise the importance of simplicity by imposing a sparsity constraint on explanatory outputs intended to automatically remove uninformative features.

	c_1: `wealth`	c_2: `race`
a_1: `sue`	−1	5
a_2: ¬`sue`	0	0

Table 5.2. Utility matrix for Alice in the (bad) bank scenario.

Rule lists and GANs have some clear advantages over linear approximators like LIME and SHAP. However, no method in use today explicitly accounts for user interests, an omission that may lead to undesirable outcomes. In short, they do not pass the eclipsing test. Recall the case of the (bad) bank in §5.5.1.3. Suppose that Alice's choice set contains just two options, A = {`sue`, ¬`sue`}, and she considers two causal hypotheses as potential explanations for her

denied loan, C = {`wealth`, `race`}. Alice's utility matrix is given in Table 5.2. Alice assigns a uniform prior over C to begin with, such that $\mathbb{P}(c_1) = \mathbb{P}(c_2) = 0.5$. She receives two explanations from Bob: g_1, according to which Alice's application was denied due to her wealth; and g_2, according to which Alice's application was denied due to her race. Using misclassification rate as our loss function and assuming a uniform distribution over both `wealth` $\in$ {`rich`, `poor`} and `race` $\in$ {`white`, `black`}, we find that both explanations are equally accurate:

$$R_{\text{emp}}(g_1, \mathbf{Z}_f) = R_{\text{emp}}(g_2, \mathbf{Z}_f) = 0.25$$

and equally simple:

$$R_{\text{emp}}(h, \mathbf{Z}_{g_1}) = R_{\text{emp}}(h, \mathbf{Z}_{g_2}) = 0.$$

However, they induce decidedly different posteriors over C:

$$\mathbb{P}(c_1|g_1) = \mathbb{P}(c_2|g_2) = 0.9$$
$$\mathbb{P}(c_1|g_2) = \mathbb{P}(c_2|g_1) = 0.1$$

The posterior expected utility of a_1 under g_1 is therefore

$$0.9(-1) + 0.1(5) = -0.4,$$

whereas under g_2 the expectation is

$$0.1(-1) + 0.9(5) = 4.4.$$

(The expected utility of a_2 is 0 under both explanations.) Since the utility-maximising action under g_2 is strictly preferable to the utility-maximising action under g_1, we regard g_2 as the superior explanation. In fact, the latter Pareto-dominates the former, since the two are equivalent in terms of accuracy and simplicity but g_1 is strictly less relevant for Alice than g_2. This determination can only be made by explicitly encoding Alice's preferences, which are currently ignored by all major iML proposals.

Methods that fail to pass the eclipsing test pose problems for all three iML goals outlined in §5.2. Irrelevant explanations can undermine tests of validity or quests of discovery by failing to recognise the epistemological purpose that motivated the question in the first place. When those explanations are accurate and simple, Alice can easily be fooled into thinking she has learned some valuable information. In fact, Bob has merely overfit the data. Matters are even worse when we audit algorithms. In this case, eclipsing explanations may offer loopholes to bad actors wishing to avoid controversy over questionable decisions. For instance, a myopic focus on accuracy and simplicity would allow (bad) banks to get away with racist loan policies so long as black applicants are found wanting along some other axis of variation.

§5.7 Objections

In this section, I consider five objections of increasing generality. The first three are levelled against the proposed game, the latter two against the entire iML project.

§5.7.1 Too Highly Idealised

One obvious objection to my proposal is that it demands a great deal of Alice. She must provide a contrastive outcome $\tilde{y}_i$, a level of abstraction LoA, a choice set A, some causal hypotheses C, a corresponding prior distribution $\mathbb{P}(C)$, and a utility function u. On top of all that, I also expect her to be a consistent Bayesian updater and expected utility maximiser. If Alice were so well-equipped and fiercely rational, then perhaps cracking black box algorithms would pose no great challenge to her.

My response is twofold. First, I remind the sceptical reader that idealisations are a popular and fruitful tool in conceptual analysis. There are no frictionless planes or infinite populations, but such assumptions have contributed to successful theories in physics and genetics. Potochnik (2017) makes a compelling case that idealisations are essential to scientific practice, enabling humans to represent and manipulate systems of incomprehensible complexity. Decision theory is no exception. The assumption that epistemic agents always make rational choices – though strictly speaking false – has advanced our understanding of individual and social behaviour in economics, psychology, and computer science.

Second, this setup is not nearly as unrealistic as it may at first appear. It is perfectly reasonable to assume that an agent would seek an algorithmic explanation with at least a counterfactual outcome and choice set to hand, as well as some (tentative) causal hypotheses. For instance, Alice may enter into the game expressly because she suspects her loan application was denied due to her race, and is unsure whether to seek redress. Utilities can be derived through a simple ranking of all action-outcome pairs. If new hypotheses emerge over the course of the game, they can easily be explored in subsequent rounds. Alice may have less confidence in ideal values for LoA and $\mathbb{P}(C)$, but there is no reason to demand certainty about these from the start. Indeed, it is advisable to try out a range of values for each, much like how analysts often experiment with different priors to ascertain the impact on posteriors in Bayesian inference (Gelman et al., 2014). Alice and Bob can iteratively refine their inputs as the rounds pass and track the evolution of the resulting Pareto frontiers to gauge the uncertainty associated with various parameters. Something like this process is how a great deal of research is in fact conducted.

Perhaps most importantly, I stress that Alice and Bob are generalised agents that can and often will be implemented by hybrid systems involving numerous humans and machines working in concert. There is no reason to artificially restrict the cognitive resources of either to that of any specific individual. The problems iML is designed to tackle are beyond the remit of any single person, especially one operating without the assistance of statistical software. When we broaden the cognitive scope of Alice and Bob, the idealisations demanded of them become decidedly more plausible. The only relevant upper bounds on their inferential capacities are computational complexity thresholds. The explanation game is an exercise in

sociotechnical epistemology, where knowledge emerges from the continuous interaction of individuals, groups, and technology (Watson & Floridi, 2018). The essential point is whether the explanation game described herein is possible and fruitful, not whether a specific Alice and a specific Bob can actually play it according to their idiosyncratic abilities.

§5.7.2 Infinite Regress

A common challenge to any account of explanation is the threat of infinite regress. Assuming that explanations must be finite, how can we be sure that some explanatory method concludes at the proper terminus? In this instance, how can we guarantee that the explanation game does not degenerate into an infinite recursive loop? Note that this is not a concern for any fixed Alice and Bob – each round ends once models g and h are scored, and Alice's expected utilities are updated – but the objection appears more menacing over shifting agents and games. For instance, we may worry that Alice and Bob together constitute a new supervised learning algorithm f_2 that maps inputs $\boldsymbol{x}_i$ to outputs $h(\boldsymbol{x}_i')$ through the intermediate model g. The resulting function may now be queried by a new agent Alice_2 who seeks the assistance of Bob_2 in accounting for some prediction $f_2(\boldsymbol{x}_i)$. This process could repeat indefinitely.

The error in this reasoning is to ignore the vital role of pragmatics. By construction, each game ends at the proper terminus *for that particular Alice*. There is nothing fallacious about allowing other agents to inquire into the products of such games as if they were new algorithms. The result will simply be t steps removed from its original source, where t is the number of Alice-and-Bob teams separating the initial f from the latest inquirer. The effect is not so unlike a game of telephone, where a message gradually degrades as players introduce new errors at each iteration. Similarly, each new Alice-and-Bob pair will do their best to approximate the work of the previous team. The end result may look quite unlike the original f for some large value of t, but that is only to be expected. So long as conditions (1)-(3) are met for any given Alice and Bob, then they are almost surely guaranteed to converge on a conditionally optimal explanation surface in polynomial time.

§5.7.3 Pragmatism + Pluralism = Relativist Anarchy?

The explanation game relies heavily on pragmatic considerations. I explicitly advocate for subjective notions of simplicity and relevance, allowing Bob to construct numerous explanations at various levels of abstraction. This combination of subjectivism and pluralism grates against the realist tradition in epistemology and philosophy of science, according to which there is exactly one true explanans for any given explanandum. Is there not a danger here of slipping into some disreputable brand of outright relativism? If criteria for explanatory success are so irreducibly subjective, is there simply no fact of the matter as to which of two competing explanations is superior? Is this not tantamount to saying that anything goes?

The short answer is no. The objection assumes that for any given fact or event there exists some uniquely satisfactory, mind- and context-independent explanation, presumably in terms of fundamental physical units and laws. Call this view explanatory monism. It amounts to a metaphysical doctrine whose merits or shortcomings are frankly beside the point. For even if the "true" explanation were always available, it would not in general be of much use. The goal of the explanation game is to promote greater *understanding for Alice*. This may come in many forms. For instance, the predictions of image classifiers are often explained by heatmaps highlighting the pixels that most contribute to the given output. The fact that complex mathematical formulae could in this case provide a maximally deep and stable explanation is irrelevant (see §5.5.1.1). Pragmatic goals require pragmatic strategies. Because iML is fundamentally about getting humans to understand the behaviour of machines, there is a growing call for personalised solutions (Páez, 2019). I take this pragmatic turn seriously and propose formal methods to implement it.

I emphatically reject the charge that the explanation game is so permissive that "anything goes". Far from it, I define objective measures of subjective notions that have long defied crisp formalisation. Once values for all variables are specified, it is a straightforward matter to score and compare competing explanations. For any set of input parameters, there exists a unique ordering of explanations in terms of their relative accuracy, simplicity, and relevance. Explanations at different levels of abstraction may be incommensurable, but together they can help Alice form a more complete picture of the target system and its behaviour near the datapoint of interest. This combination of pragmatism and explanatory ecumenism is flexible and rational. It embraces relationalism, not relativism (Floridi, 2017). One of the chief contributions of this chapter is to demonstrate that the desiderata of iML can be formulated with precision and rigour without sacrificing the subjective and contextual aspects that make each explanation game unique.

§5.7.4 No Trade-Off

Some have challenged the widespread assumption that there is an inherent trade-off between accuracy and interpretability in ML. Rudin (2019) argues forcefully against this view, which she suggests is grounded in anecdotal evidence at best, and corporate secrecy at worst. She notes that science has long shown a preference for more parsimonious models, not out of mere aesthetic whimsy, but because of well-founded principles regarding the inherent simplicity of nature (Baker, 2016). Recent results in formal learning theory confirm that an Occam's Razor approach to hypothesis testing is the optimal strategy for convergence to the truth under minimal topological constraints (Kelly et al., 2016).

Breiman (2001) famously introduced the idea of a *Rashomon set*[8] – a collection of models that estimate the same functional relationship using different algorithms and/or hyperparameters, yet all perform reasonably well (say, within 5% of the top performing model). Rudin's argument – expanded in considerable technical detail in a follow up paper (Semenova, Rudin, & Parr, 2019) – is premised on the assumption that sufficiently large Rashomon sets should include at least one interpretable model. If so, then it would seem there is no point in explaining black box algorithms, at least in high-stakes applications such as healthcare and criminal justice. If we must use ML for these purposes, then we should simply train a (globally) interpretable model in the first place, rather than reverse-engineer imperfect post-hoc explanations.

There are two problems with this objection. First, there is no logical or statistical guarantee that interpretable models will outperform black box competitors or even be in the Rashomon set of high-performing models for any given predictive problem. This is a simple corollary of the celebrated no free lunch theorem (Wolpert & Macready, 1997), which states (roughly) that there is no one-size-fits-all solution in ML. Any algorithm that performs well on one class of problems will necessarily perform poorly on another. Of course, this cuts both ways – black box algorithms are likewise guaranteed to fail on some datasets. If we value performance above all, which may well be the case for some especially important tasks, then we must be open to models of variable interpretability.

Second, the opacity of black box algorithms is not just a by-product of complex statistical techniques, but of institutional realities that are unlikely to change anytime soon. Pasquale (2015) offers a number of memorable case studies demonstrating how IP law is widely used to protect ML source code and training data not just from potential competitors but from any form of external scrutiny. Even if a firm were using an interpretable model to make its predictions, the model architecture and parameters would likely be subject to strict copyright protections. Some have argued for the creation of independent third-party groups tasked with the responsibility of auditing code under non-disclosure agreements (Floridi et al., 2018; Wachter et al., 2017), a proposal I personally support. However, until such legislation is enacted, anyone attempting to monitor the fairness, accountability, and transparency of algorithms will almost certainly have no choice but to treat the underlying technology as a black box.

§5.7.5 Double Standards

Zerilli et al. (2018) argue that proponents of iML place an unreasonable burden on algorithms by demanding that they not only perform better and faster than humans, but explain why they

[8] The name comes from Akira Kurosawa's celebrated 1950 film *Rashomon*, in which four characters give overlapping but inconsistent eyewitness accounts of a brutal crime in 8th century Kyoto.

do so as well. They point out that human decision-making is far from transparent, and that people are notoriously bad at justifying their actions. Why the double standard? We already have systems in place for accrediting human decision-makers in positions of authority (e.g., judges and doctors) based on their demonstrated track record of performance. Why should we expect anything more from machines? The authors conclude that requiring intelligibility of high-performing algorithms is not just unreasonable but potentially harmful if it hinders the implementation of models that could improve services for end users.

Zerilli et al. are right to point out that we are often unreliable narrators of our own internal reasoning. We are liable to rationalise irrational impulses, draw false inferences, and make decisions based on a host of well-documented heuristics and cognitive biases. But this is precisely what makes iML so promising: not that learning algorithms are somehow immune to human biases – they are not, at least not if those biases are manifested in the training data – but rather that, with the right tools, we may conclusively reveal the true reasoning behind consequential decisions. Kleinberg et al. (2019) make a strong case that increased automation will reduce discrimination by inaugurating rigorous, objective procedures for auditing and appealing algorithmic predictions. It is exceedingly difficult under current law to prove that a human has engaged in discriminatory behaviour, especially if they insist that they have not (which most people typically do, especially when threatened with legal sanction). For all the potential harms posed by algorithms, deliberate deception is not (yet) one of them.

I argue that the potential benefits of successful iML strategies are more varied and numerous than Kleinberg et al. acknowledge. To reiterate the motivations listed in §5.2, I see three areas of particular promise. In the case of algorithmic auditing, iML can help ensure the fair, accountable, and transparent application of complex statistical models in high-stakes applications like criminal justice and healthcare. In the case of validation, iML can be used to test algorithms before and during deployment to ensure that models are performing properly and not overfitting to uninformative patterns in the training data. In the case of discovery, iML can reveal heretofore unknown mechanisms in complex target systems, suggesting new theories and hypotheses for testing. Of course, there is no guarantee that such methods will work in every instance – iML is no panacea – but it would be foolish not to try. The double standard that Zerilli et al. caution against is in fact a welcome opportunity.

§5.8 Conclusion

Black box algorithms are here to stay. Private and public institutions already rely on ML to perform basic and complex functions with greater efficiency and accuracy than people. Growing datasets and ever-improving hardware, in combination with ongoing advances in computer science and statistics, ensure that these methods will only become more ubiquitous in the years to come.

There is less reason to believe that algorithms will become any more transparent or intelligible, at least not without the explicit and sustained effort of dedicated researchers in the burgeoning field of iML. I have argued that there are good reasons to value algorithmic interpretability on ethical, epistemological, and scientific grounds. I have outlined a formal framework in which agents can collaborate to explain the outputs of any supervised learner. The explanation game serves both a descriptive function – providing a common language in which to compare iML proposals – and a normative function – highlighting aspects that are underexplored in the current literature and pointing the way to new and improved solutions. Of course, important normative challenges remain. Thorny questions of algorithmic fairness, accountability, and transparency are not all so swiftly resolved. However, I am hopeful that the explanation game can inform these debates in a productive and principled manner.

Future work will relax the assumptions upon which this beta version of the game is based. Of special interest are adversarial alternatives in which Bob has his own utility function to maximise, or three-player versions in which Carol and Bob compete to find superior explanations from which Alice must choose. Other promising directions include implementing semi-automated explanation games with greedy algorithms that take turns maximising one explanatory desideratum at a time until convergence. Similar proposals have already been implemented for optimising mixed objectives in algorithmic fairness (Kearns et al., 2018), but I am unaware of any similar work in explainability. Finally, the scope of such games could be expanded to unsupervised learning algorithms, which pose a number of altogether different explanatory challenges.

Part III: *Poiesis*

Testing Conditional Independence in Supervised Learning Algorithms

§6 Abstract

I introduce the conditional predictive impact (CPI), a consistent and unbiased estimator of the association between one or several features and a given outcome, conditional on a reduced feature set. Building on the knockoff framework of Candès et al. (2018), I develop a novel testing procedure that works in conjunction with any valid knockoff sampler, supervised learning algorithm, and loss function. The CPI can be efficiently computed for high-dimensional data without any sparsity constraints. I demonstrate convergence criteria for the CPI and develop statistical inference procedures for evaluating its magnitude, significance, and precision. These tests aid in feature and model selection, extending traditional frequentist and Bayesian techniques to general supervised learning tasks. The CPI may also be applied in causal discovery to identify underlying multivariate graph structures. I test the method using various algorithms, including linear regression, neural networks, random forests, and support vector machines. Empirical results show that the CPI compares favourably to alternative variable importance measures and other nonparametric tests of conditional independence on a diverse array of real and simulated datasets. Simulations confirm that the inference procedures successfully control Type I error and achieve nominal coverage probability. The method has been implemented in an `R` package, `cpi`, which can be downloaded from https://github.com/dswatson/cpi.

§6.1 Introduction

Variable importance (VI) is a major topic in statistics and machine learning. It is the basis of most if not all feature selection methods, which analysts use to identify key drivers of variation in an outcome of interest and/or create more parsimonious models (Isabelle Guyon & Elisseeff, 2003; M. Kuhn & Johnson, 2019; Meinshausen & Bühlmann, 2010). Many importance measures have been proposed in recent years, either for specific algorithms or more general applications. Several different notions of VI – some overlapping, some inconsistent – have emerged from this literature. I examine these in greater detail in §6.2.1.

One fundamental difference between various importance measures is whether they test the marginal or conditional independence of features. To evaluate response variable Y's marginal dependence on predictor X_j, we test against the following hypothesis:

$$H_0^m: X_j \perp Y, \boldsymbol{X}_{-j}, \quad (6.1)$$

where $\boldsymbol{X}_{-j}$ denotes a set of covariates. A measure of conditional dependence, on the other hand, tests against a different null hypothesis:

$$H_0^c: X_j \perp Y | \boldsymbol{X}_{-j}. \quad (6.2)$$

Note that X_j's marginal VI may be high due to its association with either Y or $\mathbf{X}_{-j}$. This is why measures of marginal importance tend to favour correlated predictors. Often, however, the goal is to determine whether X_j adds any new information – in other words, whether Y is dependent on X_j even after conditioning on $\boldsymbol{X}_{-j}$. This becomes especially important when the assumption of feature independence is violated.

Tests of conditional independence (CI) are common in the causal modelling literature. For instance, the popular PC algorithm (Spirtes et al., 2000), which infers a set of underlying directed acyclic graphs (DAGs) consistent with some observational data, relies on the results of CI tests to recursively remove the edges between nodes. Common parametric examples include the partial correlation test for continuous variables or the χ^2 test for categorical data. A growing body of literature in recent years has examined nonparametric alternatives to these options. I review several such proposals in §6.2.2.

In this chapter, I introduce a new CI test to measure VI. The conditional predictive impact (CPI) quantifies the contribution of one or several features to a given algorithm's predictive performance, conditional on some other feature subset. This work relies on so-called "knockoff" variables (formally defined in §6.2.3) to provide negative controls for feature testing. Because knockoffs are, by construction, exchangeable with their observed counterparts and conditionally independent of the response, they enable a paired testing approach without any model refitting.

The CPI is extremely general. It can be used with any combination of knockoff sampler, supervised learner, and loss function. It can be efficiently computed in low or high dimensions without sparsity constraints. I demonstrate that the CPI is an unbiased estimator, provably consistent under minimal assumptions. I develop statistical inference procedures for evaluating its magnitude, precision, and significance. Finally, I demonstrate the measure's utility on a variety of real and simulated datasets.

The remainder of this paper is structured as follows. I review related work in §6.2. I present theoretical results in §6.3, where I also outline an efficient algorithm for estimating the CPI, along with corresponding p-values and confidence intervals. I test the procedure on real and simulated data in §6.4, comparing its performance with popular alternatives under a variety of regression and classification settings. Following a discussion in §6.5, I conclude in §6.6.

§6.2 Related Work

In this section, I survey the relevant literature on VI estimation, CI tests, and the knockoff filter.

§6.2.1 Variable Importance Measures

The notion of VI may feel fairly intuitive at first, but closer inspection reveals a number of underlying ambiguities. One important dichotomy is that between global and local measures, which respectively quantify the impact of features on all or particular predictions. This distinction has become especially important with the recent emergence of interpretable machine learning techniques designed to explain individual outputs of black box models (e.g., Datta, Sen, & Zick, 2016; Lundberg & Lee, 2017; Ribeiro, Singh, & Guestrin, 2016; Wachter, Mittelstadt, & Russell, 2018). In what follows, I restrict my focus to global importance measures.

Another important dichotomy is that between model-specific and model-agnostic approaches. For instance, a number of methods have been proposed for estimating importance in linear regression (Barber & Candès, 2015; Grömping, 2007; Lindeman, Merenda, & Gold, 1980), random forests (Breiman, 2001; Kursa & Rudnicki, 2010; Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008), and neural networks (Bach et al., 2015; Gevrey, Dimopoulos, & Lek, 2003; Shrikumar et al., 2017). These measures have the luxury of leveraging an algorithm's underlying assumptions and internal architecture for more precise and efficient VI estimation.

Other, more general techniques have also been developed. Van der Laan (2006) derives efficient influence curves and inference procedures for a variety of VI measures. Hubbard et al. (2018) build on this work, proposing a data-adaptive method for estimating the causal influence of variables within the targeted maximum likelihood framework (van der Laan & Rose, 2018). Williamson et al. (2020) describe an ANOVA-style decomposition of a regressor's R^2 into feature-wise contributions. Feng et al. (2018) design a neural network to efficiently compute this decomposition using multi-task learning. Fisher et al. (2019) propose a number of "reliance" statistics, calculated by integrating a loss function over the empirical distribution of covariates while holding a given feature vector constant.

Perhaps the most important distinction between various competing notions of VI is the aforementioned split between marginal and conditional measures. The topic has received considerable attention in the random forest literature, where Breiman's popular permutation technique (2001) has been criticised for failing to properly account for correlations between features (Gregorutti, Michel, & Saint-Pierre, 2015; Nicodemus et al., 2010). Conditional alternatives have been developed (Mentch & Hooker, 2016; Strobl et al., 2008), but I will not consider them here, as they are specific to tree ensembles.

The proposed measure resembles what Fisher et al. (2019) call "algorithm reliance" (AR). The authors do not have much to say about AR in their paper, the majority of which is instead devoted to two related statistics they term "model reliance" (MR) and "model class reliance" (MCR). These measure the marginal importance of a feature subset in particular

models or groups of models, respectively. Only AR measures the importance of the subset conditional on remaining covariates for a given supervised learner, which is my focus here. Fisher et al. derive probabilistic bounds for MR and MCR, but not AR. They do not develop hypothesis testing procedures for any of their reliance statistics.

§6.2.2 Conditional Independence Tests

CI tests are the cornerstone of constraint-based and hybrid methods for causal graph inference and Bayesian network learning (Koller & Friedman, 2009; Korb & Nicholson, 2009; Scutari & Denis, 2014). Assuming the causal Markov condition and faithfulness – which together state (roughly) that statistical independence implies graphical independence and vice versa – a number of algorithms have been developed that use CI tests to discover an equivalence class of DAGs consistent with a set of observational data (Maathuis, Kalisch, & Bühlmann, 2009; Spirtes et al., 2000; Verma & Pearl, 1991).

Shah & Peters (2020) have shown that there exists no uniformly valid CI test. Parametric assumptions are typically deployed to restrict the range of alternative hypotheses, which is default behaviour for most causal discovery software (e.g., Kalisch et al., 2012; Scutari, 2010). However, more flexible methods have been introduced. Much of this literature relies on techniques that embed the data in a reproducing kernel Hilbert space (RKHS). For instance, Fukumizu et al. (2008) use a normalised cross-covariance operator to test the association between features in the RKHS. A null distribution is approximated via permutation. Doran et al. (2014) build on Fukumizu et al.'s work with a modified permutation scheme intended to capture the effects of CI. Zhang et al. (2012) derive a test statistic from the traces of kernel matrices, using a gamma null distribution to compute statistical significance.

Because kernel methods do not scale well with sample size, several authors have proposed more efficient alternatives. For instance, Strobl et al. (2018) employ a fast Fourier transform to reduce the complexity of matrix operations. Methods have been developed for estimating regularised, nonlinear partial correlations (Ramsey, 2014; Shah & Peters, 2020). Lei et al. (2018) and Rinaldo et al. (2019) study the leave-one-covariate-out (LOCO) procedure, in which an algorithm is trained on data with and without the variable of interest. The predictive performance of nested models is compared to evaluate the conditional importance of the dropped feature.

The current proposal is conceptually similar to LOCO, which can in principle be extended to feature subsets of arbitrary dimension. The method enjoys some especially nice statistical properties when used in conjunction with sample splitting. For instance, Rinaldo et al. derive a central limit theorem for LOCO parameters, while Lei et al. prove finite sample error control using conformal inference. However, retraining an algorithm for each CI test

quickly becomes infeasible, especially for complex learners and/or large datasets. With knockoffs, we can directly import LOCO's statistical guarantees without any model refitting.

§6.2.3 The Knockoff Framework

My work builds on the knockoff procedure originally conceived by Barber and Candès (2015) and later refined by Candès et al. (2018). Central to this approach is the notion of a knockoff variable. Given an $n \times p$ input matrix $\boldsymbol{X}$, we may define a knockoff matrix of equal dimensionality $\tilde{\boldsymbol{X}}$ as any matrix that meets the following two criteria:

(a) *Pairwise exchangeability*. For any proper subset $S \subset [p] = (1, \dots, p)$:

$$(\boldsymbol{X}, \tilde{\boldsymbol{X}})_{swap(S)} =^{d} (\boldsymbol{X}, \tilde{\boldsymbol{X}}),$$

where $=^{d}$ represents equality in distribution and the swapping operation is defined below.

(b) *Conditional independence*. $\tilde{\boldsymbol{X}} \perp Y | \boldsymbol{X}$.

A swap is obtained by switching the entries X_j and $\tilde{X}_j$ for each $j \in S$. For example, with $p = 3$ and $S = \{1, 3\}$:

$$(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3)_{swap(S)} =^{d} (\tilde{X}_1, X_2, \tilde{X}_3, X_1, \tilde{X}_2, X_3).$$

Knockoffs provide negative controls for conditional independence testing. The intuition behind the method is that if X_j does not significantly outperform $\tilde{X}_j$ by some relevant importance measure, then the original feature may be safely removed from the final model.

Practical implementation requires both a method for generating knockoffs and a decision procedure for variable selection. The subject has quickly become a busy one in statistics and machine learning, with most authors focusing on the former task. In this chapter, I instead tackle the latter, developing a general framework for testing conditional variable importance.

Constructing nontrivial knockoffs is a considerable challenge. Numerous methods have been proposed, including but not limited to:

- Second-order Gaussian knockoffs (Candès et al., 2018)
- Conditional permutation sampling (Berrett et al., 2019)
- Hidden Markov model sampling (Sesia, Sabatti, & Candès, 2018)
- Conditional density estimation (Tansey et al., 2018)
- Generative deep neural networks (Jordon et al., 2019; Romano, Sesia, & Candès, 2019)
- Metropolis-Hastings sampling (Bates et al., 2020)

A complete review of these proposals is beyond the scope of this chapter. Bates et al. (2020) demonstrate that no efficient knockoff sampler exists for arbitrary probability distributions, suggesting that algorithms will have to make some assumptions about the data generating process to strike a reasonable balance between sensitivity and specificity.

The original knockoff papers introduce a novel algorithm for controlling the false discovery rate (FDR) in variable selection problems. The goal is to find the minimal subset $\mathcal{S} \subset [p]$ such that, conditional on $\{X_j\}_{j\in\mathcal{S}}$, Y is independent of all other variables. Call this the Markov blanket of Y (Pearl, 1988). Null features form a complementary set $\mathcal{R} = [p]\backslash\mathcal{S}$ such that $k \in \mathcal{R}$ if and only if $X_k \perp Y|\{X_j\}_{j\in\mathcal{S}}$. The FDR is given by the expected proportion of false positives among all declared positives:

$$\text{FDR} = \mathbb{E}\left[\frac{\left|\hat{\mathcal{S}} \cap \mathcal{R}\right|}{\left|\hat{\mathcal{S}} \vee 1\right|}\right], \quad (6.3)$$

where $\hat{\mathcal{S}}$ is the output of the decision procedure and the "$\vee$ 1" in the denominator enforces the convention that FDR = 0 when $\left|\hat{\mathcal{S}}\right| = 0$.

Barber & Candès (2015) demonstrate a method for guaranteed finite sample FDR control when (i) statistics for null variables are symmetric about zero and (ii) large positive statistics indicate strong evidence against the null. I will henceforth refer to this method as the adaptive thresholding test (ATT). Unlike other common techniques for controlling the FDR (e.g., Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001; Storey, 2002), the ATT does not require *p*-values as an intermediate step. Candès et al. (2018) argue that this is a benefit in high-dimensional settings, where p-value calculations can be unreliable.

Acknowledging that *p*-values may still be desired in some applications, however, Candès et al. also propose the conditional randomisation test (CRT), which provides one-sided Monte Carlo *p*-values by repeatedly sampling from the knockoff distribution. Experiments indicate that the CRT is slightly more powerful than the ATT, but the authors caution that the former is computationally intensive and do not recommend it for large datasets. That has not stopped other groups from advancing formally similar proposals (e.g., Berrett et al., 2018; Tansey et al., 2018).

I highlight several important shortcomings of the ATT: (1) Not all algorithms provide feature scoring statistics. (2) The ATT requires a large number of variables to reliably detect true positives. (3) Because the ATT does not perform individual hypothesis tests, it cannot provide confidence or credible intervals for particular variables. In the following sections, I present alternative inference procedures for conditional independence testing designed to address all three issues.

§6.3 Conditional Predictive Impact

The basic intuition behind my approach is that important features should be informative – that is, their inclusion should improve the predictive performance of an appropriate algorithm as measured by some preselected loss function. Moreover, the significance of improvement should be quantifiable so that error rates can be controlled at user-specified levels.

Consider an $n \times p$ feature matrix $\boldsymbol{X} \in \mathcal{X}$ and corresponding $n \times 1$ response variable $Y \in \mathcal{Y}$, which combine to form the dataset $\boldsymbol{Z} = (\boldsymbol{X}, Y) \in \mathcal{Z}$. Each observation $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ is an i.i.d. sample from a fixed but unknown joint probability distribution, $\mathbb{P}(\boldsymbol{Z}) = \mathbb{P}(\boldsymbol{X}, Y)$. Let $\boldsymbol{X}^S \subseteq (X_1, \ldots, X_p)$ denote some subset of features whose predictive impact we intend to quantify, conditional on the (possibly empty) set of remaining covariates $\boldsymbol{X}^R = \boldsymbol{X} \backslash \boldsymbol{X}^S$. Data can now be expressed as a triple, $\boldsymbol{Z} = (\boldsymbol{X}^S, \boldsymbol{X}^R, Y)$. We remove the predictive information in $\boldsymbol{X}^S$ while preserving the covariance structure of the predictors by replacing the submatrix with the corresponding knockoff variables, $\tilde{\boldsymbol{X}}^S$, rendering a new dataset, $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{X}}^S, \boldsymbol{X}^R, Y)$.

Define a function $f \in \mathcal{F}, \mathcal{F}: \mathcal{X} \to \mathcal{Y}$ as a mapping from features to outcomes. We evaluate a model's performance using some real-valued, non-negative loss function L. Define the risk of f with respect to $\boldsymbol{Z}$ as its expected loss over the joint probability distribution $\mathbb{P}(\boldsymbol{Z})$:

$$R(f, \boldsymbol{Z}) = \mathbb{E}[L(f, \boldsymbol{Z})]. \quad (6.4)$$

My strategy is to replace the conditional null hypothesis defined in §6.1 with the following:

$$H_0: R(f, \boldsymbol{Z}) \geq R\big(f, \tilde{\boldsymbol{Z}}\big). \quad (6.5)$$

In other words, I test whether the model performs better using the original or the knockoff data. Observe that this null hypothesis is weaker than H_0^C (Eq. 6.2), as it restricts attention to just the first moment. However, I argue that this is appropriate when the goal is simply to estimate a conditional mean rather than the complete conditional distribution. For instance, if X_j encodes information about higher moments (e.g., predictive variance or skewness) but does not improve the mean square error of a regressor, then a variable selection procedure for f should drop it, even though X_j is within the Markov blanket of Y (assuming no other covariates encode the same information).

§6.3.1 Consistency and Convergence

The CPI of submatrix $\boldsymbol{X}^S$ measures the extent to which the feature subset improves predictions made using model f. Assume that the loss function L can be evaluated for each sample i.[1] Define the following random variable:

$$\Delta_i = L(f, \tilde{\boldsymbol{z}}_i) - L(f, \boldsymbol{z}_i). \quad (6.6)$$

This vector represents the difference in sample-wise loss between predictions made using knockoff data and original data. The CPI is given by its expectation:

$$\mathrm{CPI}(\boldsymbol{X}^S) = \mathbb{E}[\Delta]. \quad (6.7)$$

Note that the CPI is always a function of some feature subset $\boldsymbol{X}^S$. I suppress the dependency for notational convenience moving forward.

[1] For loss functions that do not have this property, such as the area under the receiver operating characteristic curve, the following arguments can easily be modified to apply to each fold in a cross-validation.

To consistently estimate this statistic, it is necessary and sufficient to show that we can consistently estimate the risk of model f. The population parameter $R(f, \boldsymbol{Z})$ is estimated using the empirical risk formula:

$$R_{\text{emp}}(f, \boldsymbol{Z}) = \frac{1}{m}\sum_{i=1}^{m} L(f, \boldsymbol{z}_i). \quad (6.8)$$

The goal in estimating risk is to evaluate how well the model generalises beyond its training data, so the m samples in Eq. 6.8 constitute a test set drawn independently from $\boldsymbol{Z}$, distinct from the n samples used to fit f. In practice, this is typically achieved by some resampling procedure like cross-validation. In what follows, I presume that unit-level loss $L(f, \boldsymbol{z}_i)$ is always an out-of-sample evaluation, such that f was trained on data excluding $\boldsymbol{z}_i$.

The empirical risk minimisation (ERM) principle is a simple decision procedure in which we select the function f that minimises empirical risk in some function space $\mathcal{F}$. A celebrated result of Vapnik and Chervonenkis (1971), independently derived by Sauer (1972) and Shelah (1972), is that the ERM principle is consistent with respect to $\mathcal{F}$ if and only if the function space is of finite VC dimension. Thus, for any algorithm that meets this minimal criterion, the empirical risk $R_{\text{emp}}(f, \boldsymbol{Z})$ converges uniformly in probability to $R(f, \boldsymbol{Z})$ as $n \to \infty$, which means the estimate:

$$\begin{aligned}\widehat{\text{CPI}} &= \frac{1}{n}\sum_{i}^{n} L(f, \tilde{\boldsymbol{z}}_i) - L(f, \boldsymbol{z}_i) \quad (6.9)\\ &= R_{\text{emp}}\left(f, \tilde{\boldsymbol{Z}}\right) - R_{\text{emp}}(f, \boldsymbol{Z})\\ &= \frac{1}{n}\sum_{i}^{n} \Delta_i\end{aligned}$$

is likewise guaranteed to converge.

Though finite complexity thresholds have been derived for many algorithms – e.g., projective planes, decision trees, boosting machines, and neural networks (Shalev-Schwartz & Ben-David, 2014) – it is worth noting that some popular supervised learners do in fact have infinite VC dimension. This is the case, for instance, with methods that rely on the radial basis function kernel, widely used in support vector machines and Gaussian process regression. The learning theoretic properties of these algorithms are better described with other measures such as the Rademacher complexity and PAC-Bayes bounds (Guedj, 2019). However, as shown in §6.4, the CPI shows good convergence properties even when used with learners of infinite VC dimension.

Inference procedures for the CPI can be designed using any paired difference test. Familiar frequentist examples include the *t*-test and the Fisher exact test, which I use for large- and small-sample settings, respectively. Bayesian analogues can easily be implemented as well. Rouder et al. (2009) advocate an analytic strategy for calculating Bayes factors for *t*-tests.

Wetzels et al. (2009) and Kruschke (2013) propose more general methods based on Markov chain Monte Carlo sampling, although they differ in their proposed priors and decision procedures. Care should be taken when selecting a prior distribution in the Bayesian setting, especially with small sample sizes. Tools for Bayesian inference are implemented in the accompanying `cpi` package; however, for brevity's sake, I restrict the following sections to frequentist methods.

§6.3.2 Large Sample Inference: Paired *t*-Tests

By the central limit theorem, empirical risk estimates for functions of finite VC dimension will tend to be normally distributed around the true population parameter value. Thus I use paired, one-sided t-tests to evaluate statistical significance when samples are sufficiently large ($n \geq 30$ or thereabouts).

The variable Δ has mean $\widehat{\text{CPI}}$ and standard error SE = $s/\sqrt{n}$, where s denotes the sample standard deviation:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\Delta_i - \widehat{\text{CPI}})^2}\,. \quad (6.10)$$

The t-score for $\widehat{\text{CPI}}$ is given by $t = \widehat{\text{CPI}}/\text{SE}$, and we may compute p-values by comparing this statistic to the most tolerant distribution consistent with $H_0: R(f, \boldsymbol{Z}) \geq R(f, \widetilde{\boldsymbol{Z}})$, namely t_{n-1}. To control Type I error at level α, we reject H_0 for all t greater than or equal to the $(1-\alpha)$ quantile of t_{n-1}. This procedure can easily be modified to adjust for multiple testing.

We can relax the assumption of homoskedasticity if reliable estimates of predictive precision are available. Construct a $2n \times (n+1)$ feature matrix $\boldsymbol{X}$ with columns for each unit $i = \{1, \ldots, n\}$, as well as an indicator variable for data type D (original vs. knockoff). Let $\boldsymbol{W}$be a $2n \times 2n$ diagonal matrix such that $\boldsymbol{W}_{ii}$ denotes the weight assigned to the i^{th} prediction. For instance, in a regression setting, this could be the inverse of the expected residual variance for i. Then solve a weighted least squares regression, with the response variable $\boldsymbol{y}$ equal to the observed loss for each unit-data type combination:

$$\widehat{\boldsymbol{\gamma}} = (\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{y}$$

The t-statistic and p-value associated with coefficient γ_D can then be used to test the CPI of the substituted variable(s) under a heteroskedastic error model.

Confidence intervals around $\widehat{\text{CPI}}$ may be constructed in the typical manner. The lower bound is set by subtracting from our point estimate the product of SE and $F_{n-1}^{-1}(1-\alpha)$, where $F_{n-1}(\cdot)$ denotes the CDF of t_{n-1}. Using this formula, we obtain a 95% confidence interval for

$\widehat{\text{CPI}}$ by calculating $[\widehat{\text{CPI}} - \text{SE} \times F_{n-1}^{-1}(0.95), \infty)$. As n grows large, this interval converges to the Wald-type interval, $[\widehat{\text{CPI}} - \text{SE} \times \Phi^{-1}(0.95), \infty)$, where Φ represents the standard normal CDF.

The *t*-testing framework also allows for analytic power calculations. Let t^* denote the critical value $t^* = F_{n-1}^{-1}(1-\alpha)$. Then Type II error is given by the formula $\beta = F_{n-1}(t^* - \delta)$, where δ represents the postulated effect size. Statistical power is just the complement $1-\beta$, and rearranging this equation with simple algebra allows us to determine the sample size required to detect a given effect at some fixed Type I error α.

§6.3.3 Small Sample Inference: Fisher Exact Tests

The applicability of the central limit theorem is dubious when sample sizes are small. In such cases, exact *p*-values may be computed for a slightly modified null hypothesis using Fisher's method (1935). Rather than focusing on overall risk, this null hypothesis states that replacing $\boldsymbol{X}^S$ with the knockoff submatrix $\tilde{\boldsymbol{X}}^S$ has no impact on unit-level loss. More formally, we test against the following:

$$H_0^{\text{FEP}}: L(f, \boldsymbol{z}_i) \geq L(f, \tilde{\boldsymbol{z}}_i), \quad i = 1, \ldots, n. \quad (6.11)$$

Under this null hypothesis, which is sufficient but not necessary for the conditional predictive H_0, we may implement a permutation scheme in which the CPI is calculated for all possible assignments of data type *D*. Consider a $2n \times 3$ matrix with columns for unit index $U = \{1, 1, \ldots, n, n\}$, data type $D \in \{0, 1\}$, and loss *L*. We permute the rows of *D* subject to the constraint that every sample's loss is recorded under both original and knockoff predictions. For each possible vector *D*, compute the resulting CPI and compare the value of our observed statistics, $\widehat{\text{CPI}}$, to the complete distribution. Note that this paired setup dramatically diminishes the possible assignment space from an unmanageable $\binom{2n}{n}$, corresponding to a Bernoulli trial design, to a more reasonable 2^n. The one-tailed Fisher exact *p*-value (FEP) is given by the formula:

$$\text{FEP}\left(\widehat{\text{CPI}}\right) = \frac{1}{2^n} \sum_{b=1}^{2^n} \mathbb{I}(\widetilde{\text{CPI}}_b \geq \widehat{\text{CPI}}), \quad (6.12)$$

where $\mathbb{I}(\cdot)$ represents the indicator function and $\widetilde{\text{CPI}}_b$ is the CPI resulting from the b^{th} permutation of *D*.

To construct a confidence interval for $\widehat{\text{CPI}}$ at level $1-\alpha$, we use our empirical null distribution. Find the critical value CPI* such that FEP(CPI*) = α. Then a $(1-\alpha) \times 100\%$ confidence interval for $\widehat{\text{CPI}}$ is given by $[\widehat{\text{CPI}} - \text{CPI}^*, \infty)$. For n large, approximate calculations can be made by sampling from the set of 2^n permissible permutations. In this case, however, it is important to add 1 to both the numerator and denominator to ensure unbiased inference (Phipson & Smyth, 2010).

§6.3.4 Computational Complexity

To summarise, I outline the proposed algorithm for testing the conditional importance of feature subsets for supervised learners in pseudocode below. This algorithm executes in $\mathcal{O}(ak + g + h)$ time. I take the complexity of the learner a and knockoff sampler g to be given. The empirical risk estimator k can be made more or less complex depending on the resampling procedure. The most efficient option for evaluating generalisation error is the holdout method, in which a model is trained on a random subset of the available data and tested on the remainder. Unfortunately, this procedure can be unreliable with small sample sizes. Popular alternatives include the bootstrap and cross-validation. Both require considerable model refitting, which can be costly when a is complex.

The inference procedure h is quite efficient in the parametric case – on the order of $\mathcal{O}(n)$ for the t-test – but scales exponentially with the sample size when using the permutation-based approach. As noted above, the complexity of the Fisher test can be bounded by setting an upper limit on the number of permutations B used to approximate the empirical null distribution. The standard error of a p-value estimate made using such an approximation is $\sqrt{p^*(1-p^*)/B}$, where p^* represents the true p-value. This expression is maximised at $p^* = 0.5$, corresponding to a standard error of $1/(2\sqrt{B})$. Thus, to guarantee a standard error of at most 0.001, it would suffice to use $B = 250{,}000$ permutations, an eminently feasible computation on a modern laptop.

Algorithm 6.1: CPI Algorithm

Inputs: Dataset $\mathbf{Z}$, submatrix $\mathbf{X}^S$, supervised learner a, risk functional R, knockoff sampler g, risk estimator k, inference procedure h

1. Grow a model f on Z
2. Apply g to generate the knockoff matrix $\tilde{\mathbf{X}}^S$
3. Use risk estimator k to compute each $L(f, \boldsymbol{z}_i)$ and $L(f, \tilde{\boldsymbol{z}}_i)$
4. Compute $\widehat{\text{CPI}} = n^{-1} \sum_{i=1}^{n} L(f, \tilde{\boldsymbol{z}}_i) - L(f, \boldsymbol{z}_i)$
5. Apply inference procedure h to determine the associated p-value (p) and confidence interval (ci)

Output: $\widehat{\text{CPI}}$, p, ci

§6.4 Experiments

All experiments were conducted in the `R` statistical computing environment, version 3.6.2. Code for reproducing all results and figures can be found in the dedicated GitHub repository: https://github.com/dswatson/dissertation.

§6.4.1 Simulated Data

I report results from a number of simulation studies. First, I analyse the statistical properties of the proposed tests under null and alternative hypotheses. I proceed to compare the sensitivity and specificity of the CPI to those of several alternative measures.

Data were simulated under four scenarios, corresponding to all combinations of independent vs. correlated predictors and linear vs. nonlinear outcomes. Because conditional importance is most relevant in the case of correlated predictors, results for the two scenarios with independent features are left to Appendix A, §2. In the linear setting, 10 variables were drawn from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, with covariance matrix $\Sigma_{ij} = 0.5^{|i-j|}$. A continuous outcome Y was calculated as $Y = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, where $\boldsymbol{\beta} = (0.0, 0.1, \ldots, 0.9)'$ and $\varepsilon \sim \mathcal{N}(0,1)$. In the nonlinear scenario, I keep the same predictors but generate the response from a transformed matrix, $Y = \boldsymbol{X}^*\boldsymbol{\beta} + \varepsilon$, where

$$x_{ij}^* = \begin{cases} +1, & \text{if } \Phi^{-1}(0.25) \leq x_{ij} \leq \Phi^{-1}(0.75) \\ -1, & \text{else} \end{cases}$$

with the same $\boldsymbol{\beta}$ and ε as in the linear case. A similar simulation was performed for a classification outcome, where the response Y was drawn from a binomial distribution with probability $[1 + \exp(-\boldsymbol{X}\boldsymbol{\beta})]^{-1}$ and $[1 + \exp(-\boldsymbol{X}^*\boldsymbol{\beta})]^{-1}$ for the linear and nonlinear scenarios, respectively.

Knockoffs for all simulated data were generated using the second-order Gaussian technique described in (Candès et al., 2018) and implemented in the `knockoff` package, version 0.3.2 (Patterson & Sesia, 2018).

§6.4.1.1 Type I and Type II Errors

I simulate 10^4 datasets with $n = 1000$ observations and compute the CPI using four different learning algorithms: linear/logistic regression (LM), random forest (RF), artificial neural network (ANN), and support vector machine (SVM). Risk was estimated using holdout sampling with a train/test ratio of 2:1. For regression models, I used mean square error (MSE) and mean absolute error (MAE) loss functions; for classification, I used cross entropy (CE) and mean misclassification error (MMCE). I computed p-values via the inference procedures described in §6.3, i.e. paired t-tests and Fisher exact tests. For Fisher tests I used 10^4 permutations.

Linear and logistic regressions were built using the functions `lm()` and `glm()`, respectively, from the `R` package `stats` (R Core Team, 2020). RFs were built using the `ranger` package (Wright & Ziegler, 2017), with 500 trees. ANNs were built with the `nnet` package (Venables & Ripley, 2002), with 20 hidden units and a weight decay of 0.1. SVMs were built with the `e1071` package (Meyer et al., 2018), using a Gaussian radial basis function (RBF) kernel and $\sigma = 1$. Unless stated otherwise, all parameters were left to their default values. Resampling was performed with the `mlr` package (Bischl et al., 2016).

Significance levels for all tests were fixed at $\alpha = 0.05$. For each simulation, I recorded CPI values, Type I errors, Type II errors, empirical coverage, and t-statistics, where applicable.

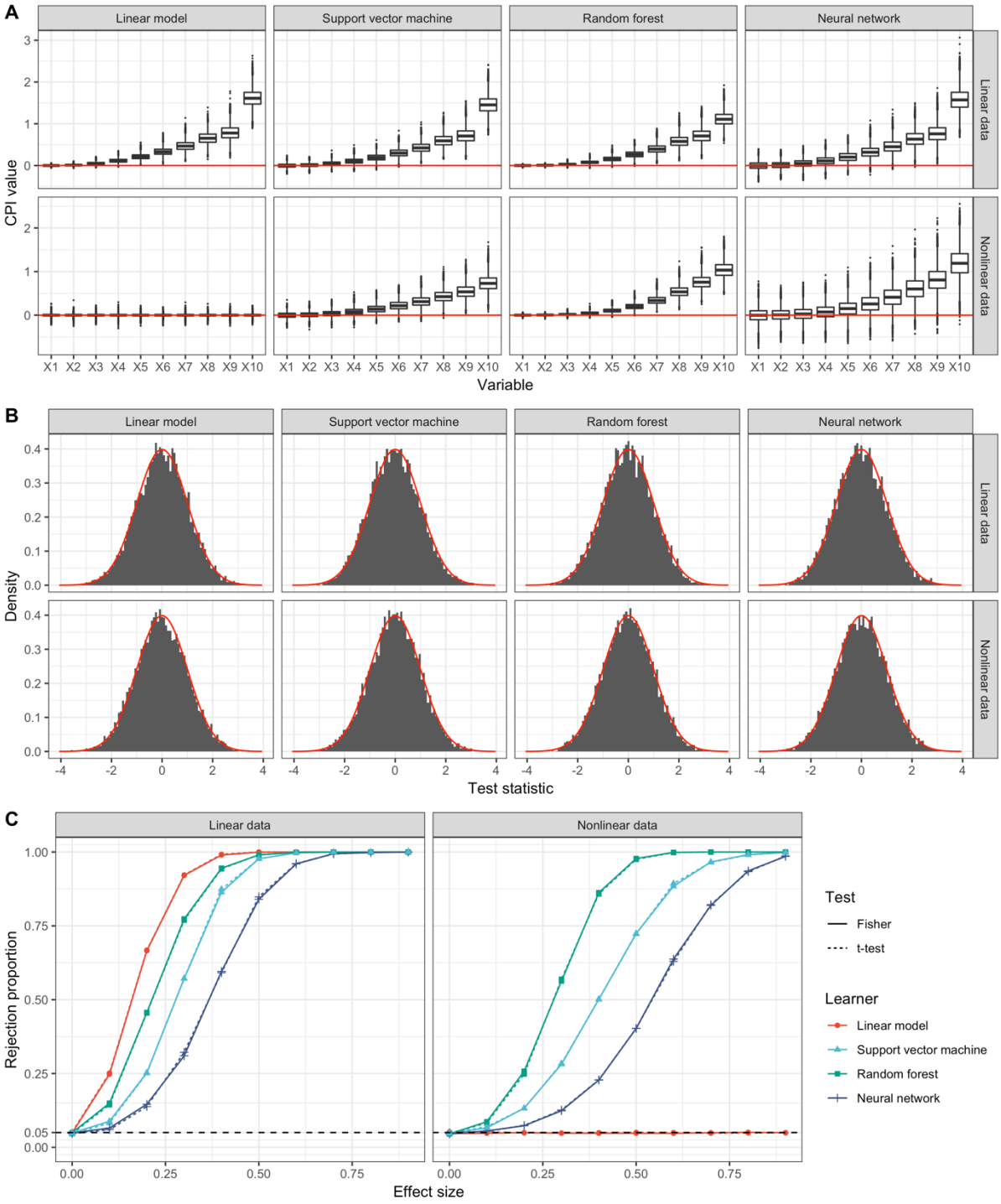

Figure 6.1. Simulation results for continuous outcome with MSE loss and correlated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

Results for MSE loss are shown in Fig. 6.1. Similar plots for MAE, CE, and MMCE loss functions are presented in Appendix A, along with tables of empirical coverage probabilities.

For continuous outcomes, CPI controlled Type I error with all four learners and reached 100% power under all settings, with the exception of the LM on nonlinear data. No observable differences are detected between the MSE and MAE loss functions. Similar results were found for categorical outcomes. The CPI controlled Type I error for the MMCE and CE loss functions with all four learners. The LM once again performed poorly on nonlinear data, as expected. The Fisher test had slightly increased power compared to the *t*-test. Statistical power was generally greater with CE loss than with MMCE loss.

§6.4.1.2 Comparative Performance

I use the same simulation setup to compare the CPI's performance to that of three other global, nonparametric, model-agnostic measures of CI:

- ANOVA: Williamson et al. (2020)'s nonparametric ANOVA-inspired VI, as implemented in the `R` package `vimp`, version 1.1.4.
- LOCO: Lei et al. (2018)'s leave-out-covariates measure, as implemented in the `R` package `conformalInference`, version 1.1.
- GCM: Shah & Peters (2020)'s generalised covariance measure, a nonparametric estimate of the partial correlation between two vectors.

Unfortunately, software for Hubbard et al. (2018)'s targeted maximum likelihood VI statistic was still under development at the time of testing, and beta versions generated errors. Candès et al.'s probabilistic knockoff procedure (2018) can be extended to nonparametric models, but requires an algorithm-specific VI measure, which not all learners provide. I consider this method separately in §6.4.1.3. Kernel methods do not work with arbitrary algorithms and were therefore excluded. I restrict this section to the regression setting, as only LOCO can be extended to classification problems.

Training and test sets are of equal size, with $n = \{100, 500, 1000\}$. In each case, I fit LM, RF, ANN, and SVM regressions, as described previously. I estimate the VI of all features on the test set for every model. This procedure was repeated 10^4 times. Results for $n = 1000$ are plotted in Fig. 6.2. Similar results for smaller sample sizes are included in Appendix A (§9.3).

All methods have high Type II error rates when fitting an LM to nonlinear data, highlighting the dangers of model misspecification. GCM appears to dominate in the linear setting but struggles to detect VI in nonlinear simulations. LOCO is somewhat conservative, often falling short of the nominal Type I error rate under the null hypothesis. However, the method fails to control Type I error in the case of an ANN trained on nonlinear data. The nonparametric ANOVA generally performs poorly, especially with RF regressions, where we may observe Type I error rates of up to 100%.

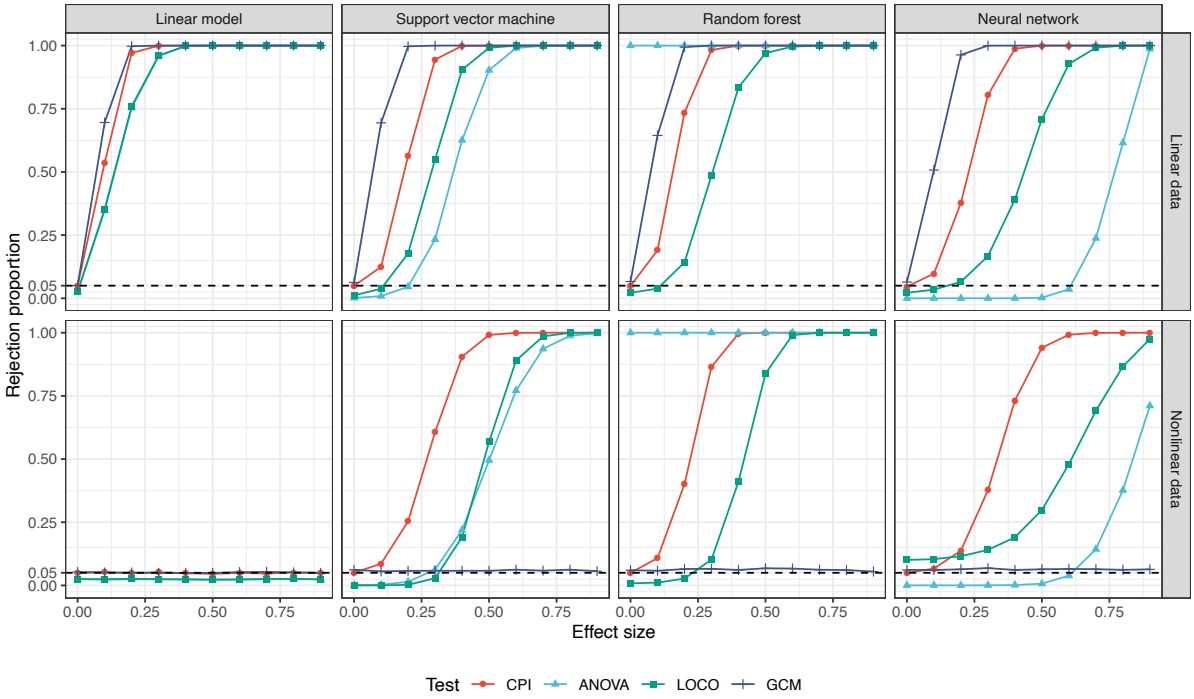

Figure 6.2. Comparative performance of VI measures across different simulations and algorithms. Plots depict the proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$. These results were computed using training and test samples of $n = 1000$ and $p = 10$. Similar results were obtained for sample sizes of $n = \{100, 500\}$ and $p = \{20, 50, 100\}$ (see Appendix A, §2).

The CPI outperforms all competitors with nonlinear data, and achieves greater power than ANOVA or LOCO in the linear case. GCM is the only other method to control Type I error under all simulation settings, but it has nearly zero power with nonlinear data.

§6.4.1.3 Knockoff Filter

To compare the performance of the CPI with that of the original knockoff filter, I followed the simulation procedure described in §6.4 of (Candès et al., 2018). A $n = 300 \times p = 1000$ matrix was sampled from a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$, with covariance matrix $\Sigma_{ij} = \rho^{|i-j|}$. A continuous outcome Y was calculated as $Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,1)$ and the coefficient vector $\boldsymbol{\beta}$ contains just 60 nonzero entries, with random signs and variable effect sizes. I vary ρ with fixed nonzero $|\beta| = 1$, and vary effect size with fixed $\rho = 0$.

I train a series of lasso regressions (Tibshirani, 1996) on the data using the original design matrix and 10-fold cross-validation to calculate the CPI, and the expanded $n \times 2p$ design matrix for the knockoff filter. VI for the latter was estimated using the difference statistic originally proposed by Barber and Candès (2015):

$$W_j = |\hat{\beta}_j| - |\hat{\beta}_{j+p}|, \quad (6.13)$$

where $|\hat{\beta}_j|$ and $|\hat{\beta}_{j+p}|$ represent coefficients associated with a feature and its knockoff, respectively, at some fixed value of the Lagrange multiplier λ. Variables are selected based on the ATT method described in §6.2.3. The hyperparameter λ is tuned via 10-fold cross-validation, per the default settings of the `glmnet` package (Friedman, Hastie, & Tibshirani, 2010). Power and FDR are averaged over 10^4 iterations for each combination of effect size and autocorrelation coefficient.

The CPI is more powerful than the original knockoff filter for all effect sizes at $\rho = 0$, but less powerful for high autocorrelation of $\rho = 0.5$ (see Fig. 6.3). Both methods generally control the FDR at the target rate of 10%. The only exceptions are under small effect sizes, where the knockoff filter shows slightly inflated errors.

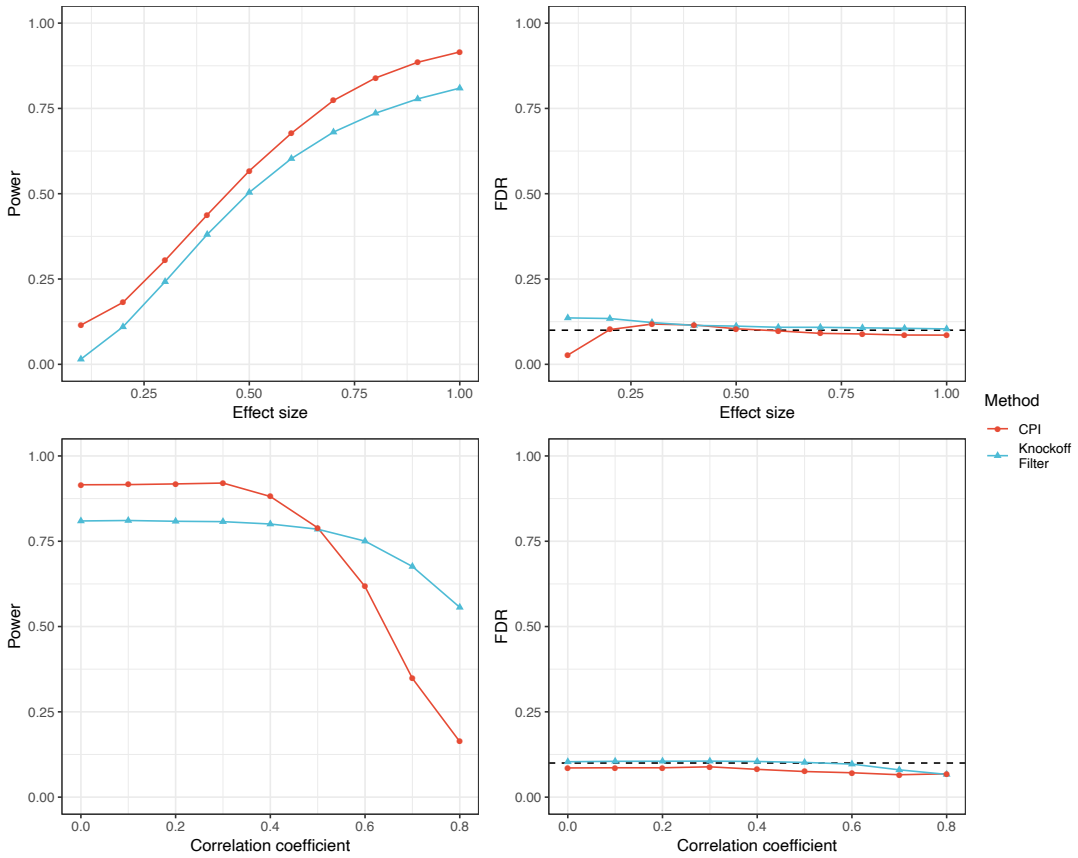

Figure 6.3. Power and FDR as a function of effect size and autocorrelation for CPI and knockoff filter. Target FDR is 10%. Results are from a lasso regression with $n = 300$ and $p = 1000$. Each point represents 10,000 replications. Similar results were obtained for $p = 2000$ (see Appendix A, §3).

Note that in addition to being a more powerful test under most conditions, the CPI has other important advantages over the ATT. Whereas the latter can only be applied to algorithms with inbuilt feature scoring statistics, the former requires nothing more than a valid loss function. Whereas the ATT struggles to select important variables in low-dimensional settings, the CPI imposes no dimensionality restraints. Finally, the CPI is more informative, insomuch as it provides feature-level p-values and confidence (or credible) intervals.

§6.4.2 Real Data

In this section, I apply the CPI to real datasets of low- and high-dimensionality.

§6.4.2.1 Boston Housing

I analysed the Boston housing data (Harrison & Rubinfeld, 1978), which consists of 506 observations and 14 variables. This benchmark dataset is available in the UCI Machine Learning Repository (Dua & Graff, 2017). The dependent variable is the median price of owner-occupied houses in census tracts in the Boston metropolitan area in 1970. The independent variables include the average number of rooms, crime rates, and air pollution.

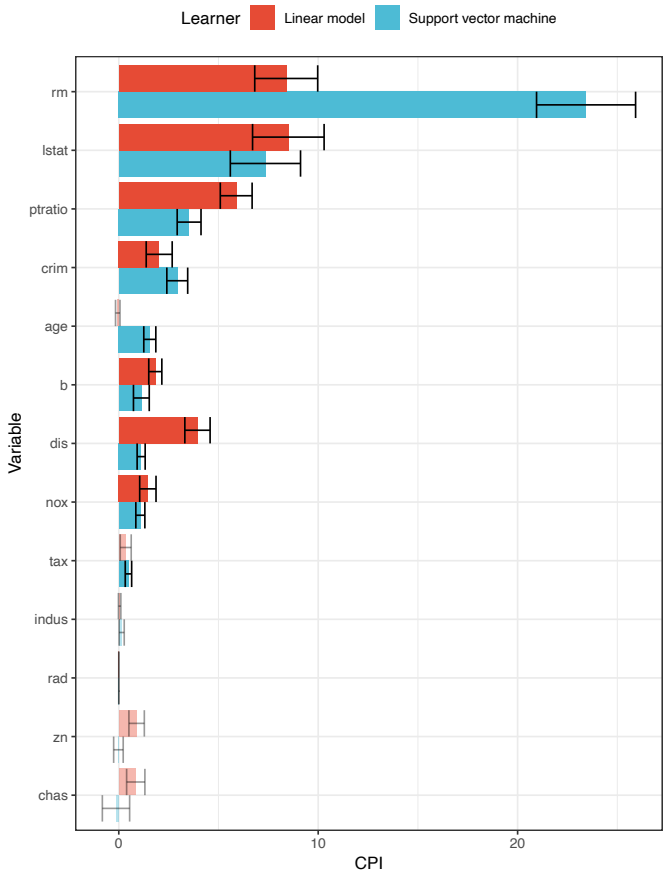

Figure 6.4. Results of the Boston housing experiment. For each variable in the data set, the CPI value is shown, computed with a linear model and a support vector machine. Whiskers represent standard errors. Non-significant variables at $\alpha = 0.05$ after adjustment for multiple testing are shaded.

Using LM and SVM regressions, I computed CPI, standard errors, and t-test p-values for each feature, adjusting for multiple testing using Holm's (1979) procedure. I used an RBF kernel for the SVM, measured performance via MSE, and used 5 subsampling iterations to evaluate empirical risk. The results are shown in Fig. 6.4. I found significant effects at $\alpha = 0.05$ for the average number of rooms (`rm`), percentage of lower status of the population (`lstat`), pupil-teacher ratio (`ptratio`), and several other variables with both LM and SVM, which is in line with previous analyses (Friedman & Popescu, 2008; Williamson et al., 2020).

Interestingly, the SVM assigned a higher CPI value to several variables compared to the LM. For example, the proportion of owner-occupied units built prior to 1940 (`age`) significantly increased the predictive performance of the SVM but had approximately zero impact on the LM. The reason for this difference might be a nonlinear interaction between `rm` and `age`, which was also observed by Friedman and Popescu (2008).

§6.4.2.2 Breast Cancer

I examined gene expression profiles of human breast cancer samples downloaded from GEO series GSE3165. Only the 94 arrays of platform GPL887 (Agilent Human 1A Microarray V2) were included. These data were originally analysed by Herschkowitz et al. (2007) and later studied by Lim et al. (2009). I follow their pre-processing pipeline, leaving 13,064 genes. All samples were taken from tumour tissue and classified into one of six molecular subtypes: basal-like, luminal A, luminal B, Her2, normal-like, and claudin-low.

Basal-like breast cancer (BLBC) is an especially aggressive form of the disease, and BLBC patients generally have a poor prognosis. Following Wu & Smyth (2012), I defined a binary response vector to indicate whether or not samples are BLBC. Gene sets were downloaded from the curated C2 collection of the MSigDB and tested for their association with this dichotomous outcome.

I trained an RF classifier with 10^4 trees to predict BLBC based on microarray data. Second-order knockoffs were sampled using an approximate semidefinite program with block-diagonal covariance matrices of maximum dimension 4000×4000. I test the CPI for each of the 2,609 gene sets in the C2 collection for which at least 25 genes were present in the expression matrix. Models were evaluated using the CE loss function on out-of-bag samples.

I calculate p-values for each CPI via the t-test and corresponding q-values using the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995). I identify 660 significantly enriched gene sets at $q \leq 0.05$, including 24 of 73 explicitly breast cancer derived gene sets and 6 of 13 gene sets indicative of basal signatures. Nearly all top results are from cancer studies or other biologically relevant research (see Fig. 6.5). These include 4 sets of BRCA1 targets, genetic mutations known to be associated with BLBC (Turner & Reis-Filho, 2006), and 4 sets of ESR1 targets, established markers for the luminal A subtype (Therese Sørlie et al., 2003).

By comparison, popular pathway enrichment tests like GSEA (Subramanian et al., 2005) and CAMERA (Wu and Smyth, 2012) respectively identify 137 and 74 differentially enriched pathways in this dataset at 5% FDR. These results are especially notable given that those methods rely on marginal associations between gene expression and clinical outcomes, whereas the CPI is a conditional test with a more restrictive null hypothesis, and should theoretically have less power to detect enrichment when features within a gene set are correlated with others outside it. Despite collinearity between genes, the CPI still identifies a large

number of biologically meaningful gene sets differentiating BLBC tumours from other breast cancer subtypes.

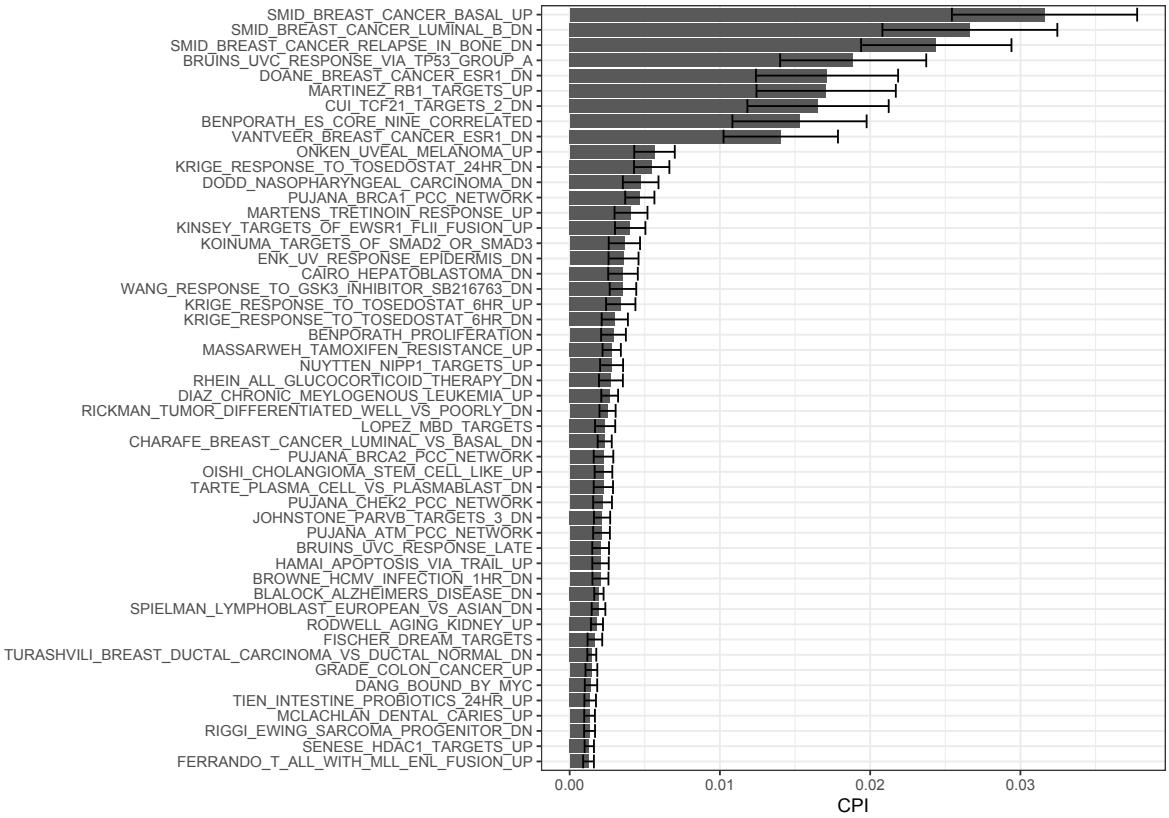

Figure 6.5. Results for the top 50 gene sets. For each gene set, the CPI value is shown, computed with a random forest. Whiskers represent standard errors.

§6.5 Discussion

Shah & Peters (2020) have demonstrated that no CI test can be uniformly valid against arbitrary alternatives, a sort of no-free-lunch (NFL) theorem for CI. Bates et al. (2020) prove a similar NFL theorem for constructing knockoff variables, showing that no algorithm can efficiently compute nontrivial knockoffs for arbitrary input distributions. The original NFL theorem for optimisation is well-known (Wolpert & Macready, 1997). Together, these results delimit the scope of the CPI. The method is extremely general, in the sense that it works with any well-chosen combination of supervised learner, loss function, and knockoff sampler. However, it is simultaneously constrained by these choices. The CPI cannot be guaranteed to control Type I error or have any power against the null when knockoffs are poorly constructed or models are misspecified.

In this chapter's experiments, I employed a variety of risk estimators, including cross-validation, subsampling, out-of-bag estimates, and the holdout method. Results did not depend on these choices, suggesting that analysts may use whichever is most efficient for the problem at hand.

Computational bottlenecks can complicate the use of this procedure for high-dimensional datasets. It took approximately 49 hours to generate second-order knockoffs for the gene expression matrix described in §6.4.2.2. However, as noted in §6.2.3, knockoff sampling is an active area of research, and it is reasonable to expect future advances to speed up the procedure considerably.

§6.6 Conclusion

I propose the conditional predictive impact (CPI), a global measure of variable importance and general test of conditional independence. It works for regression and classification problems using any combination of knockoff sampler, supervised learning algorithm, and loss function. It imposes no parametric or sparsity constraints, and can be efficiently computed on data with many observations and features. I propose a number of inference procedures that are fast and powerful, able to simultaneously control Type I error and achieve nominal coverage probability. I have shown that this approach is consistent and unbiased under minimal assumptions. Empirical results demonstrate that the method performs favourably against a number of alternatives for a range of supervised learners and data generating processes.

I envision several avenues for future research in this area. Localised versions of the CPI algorithm could be used to detect the conditional importance of features on particular predictions. Model-specific methods could be implemented to speed up the procedure. Potential applications to causal discovery and inference represent an especially promising direction for this approach.

Rational Feature Attributions:
Synthesising Shapley Values and Counterfactuals

§7 Abstract

Explaining the predictions of opaque machine learning algorithms is an important and challenging task, especially as complex models are increasingly used to assist in high-stakes decisions such as those arising in healthcare and finance. Most popular tools for post-hoc explainability are either insensitive to context (e.g., feature attributions) or difficult to interpret (e.g., counterfactuals). In either case, the default behaviour is to treat explanations as objective deliverables, mere functions of a supervised learning model and accompanying dataset. However, in many cases, agents will justifiably want and expect compact, custom explanations that explicitly account for their individual beliefs and interests. For instance, doctors and patients may seek different explanations for the same algorithmic diagnosis. In this chapter, I formalise and implement a pragmatic approach to explainability, leveraging tools from decision theory and causal modelling to synthesise and extend current methods in a rigorous, flexible manner. I review a range of possible reference distributions and compare their applicability across various explanation tasks. Through a series of examples and experiments, I demonstrate how user goals and knowledge can inform and constrain the range of satisfactory explanations for given model predictions.

§7.1 Introduction

Machine learning algorithms are increasingly used in a variety of high-stakes domains, from credit scoring to medical diagnosis. However, many of the most popular and powerful statistical methods are *opaque*, in the sense that end users cannot understand the reasoning that goes into particular predictions. The last few years have seen numerous proposals for how to mitigate this problem in a post-hoc, model-agnostic manner, most prominently via feature attributions and counterfactuals, which are amongst the most widely used methods in the subdiscipline of computational statistics known as interpretable machine learning (iML). These strands of research have so far been conducted in parallel, with little or no overlap.

Both approaches have recently come under fire for failing to properly handle dependencies between covariates. Most methods for computing Shapley values – a popular feature attribution technique, formally defined in §7.2 – implicitly treat predictors as mutually independent, assigning positive weight to potentially impossible data permutations (Aas, Jullum, & Løland, 2019; Frye et al., 2020; Kumar et al., 2020). Much recent work in counterfactual explanation has focused on how to model the impact of interventions on downstream variables in the system, a reformulation that requires an explicitly causal approach (Karimi, Schölkopf, et al., 2020; Mahajan et al., 2019).

Others have pushed back against these critiques. Sundararajan & Najmi (2019) demonstrate that conditioning on covariates can misleadingly assign importance to irrelevant features. Janzing et al. (2020) point out that standard supervised learning algorithms do not explicitly model dependencies between features, and so intervening to set predictors to some fixed value is not just permissible but proper when computing Shapley values.

Often lost in these debates is the seemingly obvious realisation that the "right" explanation depends crucially on who is asking and why. Current explainability methods tend to treat explanations as objective deliverables that can be computed from just two ingredients: a target function and an accompanying dataset of real or simulated samples. This, I argue, ignores a fundamental fact long acknowledged in philosophy and social science – that *explanations are fundamentally pragmatic*. To succeed, they must be tailored to the inquiring agent, who requests an explanation on the basis of certain beliefs and interests.

If pragmatic approaches are few and far between in the iML literature, perhaps this is because it is far from obvious how we should formalise abstract, subjective notions like beliefs and interests. In this chapter, I use tools from Bayesian decision theory to do just that. My main contributions are threefold: (1) I combine disparate work in feature attributions and counterfactuals, demonstrating how the two can join forces to create more useful and flexible explanations. (2) I extend the axiomatic guarantees of Shapley values, allowing user beliefs and preferences to guide the search for a relevant subspace of contrastive baselines. (3) I implement an expected utility maximisation procedure that optionally incorporates causal information where available and appropriate to compute optimal attribution sets for individual agents.

The remainder of this chapter is structured as follows. In §7.2, I review important background material in iML, with a focus on Shapley values and counterfactuals, as well as the essential formalisms of decision theory and structural causal models. In §7.3, I compare several different reference distributions and propose a new desideratum for Shapley values, consistent with the characteristic axioms yet better suited for explainability. I describe a novel algorithm for computing *rational* Shapley values in §7.4 and evaluate its empirical performance on a range of benchmark tasks. A discussion follows in §7.5, where I consider and respond to several objections. §7.6 concludes.

§7.2 Background

This section covers relevant literature in iML, decision theory, and causal analysis.

§7.2.1 iML Methods

As I established in Chapter 3, algorithmic explainability is a large and heterogeneous research programme. My focus in this chapter is restricted to post-hoc, model-agnostic local explana-

tions. These methods attempt to explain particular predictions from a target model f without making any assumptions about its form. Two popular approaches, which I build upon below, are Shapley values and counterfactuals.

Originally developed in the context of cooperative game theory, Shapley values (Shapley, 1953) have been adapted for model interpretability by numerous authors. They represent the contribution of each feature toward a particular prediction. Let $\boldsymbol{x}_i \in \boldsymbol{\mathcal{X}} \subset \mathbb{R}^d$ denote an input datapoint and $f(\boldsymbol{x}_i) = \hat{y}_i \in \mathcal{Y} \subset \mathbb{R}$ the corresponding output of function f. Shapley values aim to express this value as a sum:

$$f(\boldsymbol{x}_i) = \phi_0 + \sum_{j=1}^{d} \phi_j \,, \quad (7.1)$$

where ϕ_0 represents a baseline expectation and ϕ_j the weight assigned to feature X_j at point $\boldsymbol{x}_i$. Let $v\colon 2^d \to \mathbb{R}$ be a value function such that $v(S)$ is the payoff associated with feature subset $S \subseteq [d] = \{1, \dots, d\}$ and $v(\emptyset) = 0$. The Shapley value ϕ_j is given by j's average marginal contribution to all subsets that exclude it:

$$\phi_j = \frac{1}{d!} \sum_{S \subseteq [d] \setminus \{j\}} |S|!\,(d - |S| - 1)!\,[v(S \cup \{j\}) - v(S)]. \quad (7.2)$$

It can be shown that this is the unique value satisfying a number of desirable properties, including efficiency, linearity, sensitivity, and symmetry.[1] Computing exact Shapley values is NP-hard, although several efficient approximations have been proposed.

There is some ambiguity as to how one should calculate payoffs on a proper subset of features, since f requires d-dimensional input. Let S and R denote a partition of $[d]$, such that we can rewrite any $\boldsymbol{x}_i$ as a pair of subvectors $(\boldsymbol{x}_i^S, \boldsymbol{x}_i^R)$. Then the payoff for feature subset S takes the form of an expectation, with $\boldsymbol{x}_i^S$ held fixed while $\boldsymbol{X}^R$ varies. Following Merrick & Taly (2020), I consider a general formulation of the value function, indexed by a distribution $\mathcal{D}_R$:

$$v_{\mathcal{D}_R}(S) = \mathbb{E}_{\boldsymbol{X}^R \sim \mathcal{D}_R}\big[f\big(\boldsymbol{x}_i^S, \boldsymbol{X}^R\big)\big]. \quad (7.3)$$

Popular options for $\mathcal{D}_R$ include the marginal distribution $P(\boldsymbol{X}^R)$, as in Lundberg & Lee (2017); the conditional $P(\boldsymbol{X}^R|\boldsymbol{x}_i^S)$, as in Aas et al. (2019); and the interventional $P(\boldsymbol{X}^R|do(\boldsymbol{x}_i^S))$, as in Heskes et al. (2020). Each reference distribution offers certain advantages and disadvantages, but the choice of which to use is ultimately dependent upon one's analytical goals. One contribution of this chapter is to offer guidance on which values of $\mathcal{D}_R$ to plug in for different tasks.

Counterfactual explanations, unlike Shapley values, do not produce feature weights.[2] Instead, they identify one or several nearest neighbours with different outcomes – say, all

[1] See Appendix B, §1 for formal statements of the Shapley axioms.

[2] Somewhat confusingly, the term "counterfactual" has a different meaning in the causal modelling literature, where it refers to the probability of an alternative outcome conditional on the true outcome and an unrealised antecedent. The probability that I would still have a headache if I had not taken an aspirin, given the fact that I do not have a headache and did take an aspirin, is a canonical example of such a counterfactual probability.

datapoints $\boldsymbol{x}$ within an ε-ball of $\boldsymbol{x}_i$ such that labels $f(\boldsymbol{x})$ and $f(\boldsymbol{x}_i)$ differ (for classification) or $f(\boldsymbol{x}) > f(\boldsymbol{x}_i) + \tau$ (for regression). The optimisation problem can be expressed as

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \text{CF}(\boldsymbol{x}_i)}{\text{argmin}}\, cost(\boldsymbol{x}_i, \boldsymbol{x}), \quad (7.4)$$

where $\text{CF}(\boldsymbol{x}_i)$ denotes a counterfactual space such that $f(\boldsymbol{x}_i) \neq f(\boldsymbol{x})$ and $cost(\cdot)$ is a user-specified cost function, typically identified with some measure of distance. Wachter, Mittelstadt, & Russell (2018) recommend using generative adversarial networks (GANs) to solve Eq. 7.4, while others have proposed a variety of alternatives that constrain solutions in an effort to ensure that counterfactuals are coherent and actionable (Karimi et al., 2020; Mothilal, Sharma, & Tan, 2020; Poyiadzi et al., 2020; Russell, 2019; Ustun, Spangher, & Liu, 2019). A simplified version of this approach is implemented in Google's What-If Tool (Wexler et al., 2020), which selects counterfactuals by finding the nearest observed points on the opposite side of a decision boundary.

Though feature attributions and counterfactuals are probably the most popular algorithmic explainability tools on the market today (Bhatt et al., 2020), some interactive alternatives have been proposed. Lakkaraju et al. (2019) introduce MUSE, which allows users to specify particular features of interest. Explanations are computed as compact decision sets within this subspace. Sokol & Flach (2020) present LIMEtree, an interactive method in which users submit a series of "What if?" questions that guide the surrogate model training process. This work differs from these methods in several respects. First, I maintain and extend the axiomatic guarantees of Shapley values, providing feature attributions as opposed to rule lists. Second, I optionally incorporate causal information to accommodate both model- and system-level explanations. Finally, I explicitly encode user beliefs and interests via credence and utility functions that enable precise expected utility maximisation.

Within the Shapley value literature, some authors have recently begun to acknowledge the irreducible context-sensitivity of feature attributions. Merrick & Taly (2020) review a number of incompatible methods for estimating Shapley values, demonstrating how results vary with different baseline expectations and reference distributions. Chen et al. (2020) observe that explanations may be true to the model or true to the data, and argue that the choice of which to prefer depends on the desired application. The solutions proposed in this latter work are restricted to linear models and multivariate Gaussian distributions, two limitations not imposed here. In neither article do the authors consider connections to counterfactual explanations or causal inference.

This chapter builds on proposals first sketched in Chapter 5, where I laid out a pluralist approach to iML under the framework of "explanation games". However, I did not offer any

Counterfactuals in iML are what Dawid (2000) aptly calls "hypotheticals" – i.e., other possibilities *tout court*. I will for the most part stick with the terminological conventions of iML in this chapter, except where distinctions must be made for clarity.

practical algorithms there. In this chapter, I go beyond conceptual analysis by implementing an efficient algorithm and applying it on a range of examples from benchmark datasets, comparing performance against alternative methods in a series of experiments.

§7.2.2 Bayesian Decision Theory

Why do we use iML tools in the first place? In Chapter 5, I offered three reasons: (1) to audit for potential bias; (2) to validate performance, guarding against unexpected errors; and (3) to discover underlying mechanisms of the data generating process. Another use case, with elements of (1) and (3) yet distinctly analysed in the algorithmic recourse setting, is (4) to recommend actions so as to alter predicted outcomes. In all four cases, the aim is ultimately to make some sort of decision – be it about whether to sue a firm, deploy an algorithm, or perform an experiment. These motivations may overlap at the edges but they represent distinct tasks based on different assumptions and requiring their own explanatory methodologies. This heterogeneity is largely ignored by current feature attribution and counterfactual approaches, which implicitly assume a sort of *explanatory objectivism*. According to this view, the quality of an explanation is independent of its context.

Decades of research in philosophy and social science has made such objectivism untenable (Floridi, 2019; Legg & Hookway, 2020; Miller, 2019). If the view has any adherents today, perhaps it is because the alternative is often misconstrued as *relativism*, an anything-goes slippery slope that inevitably leads to what Feyerabend (1975) (approvingly) refers to as "epistemological anarchism". This, however, is a false dichotomy. I reject the polar extremes of objectivism and relativism in favour of *pragmatism*, which holds neither that there exists some single ideal explanation nor that all explanations are equally valid, but rather that various explanations may be more or less appropriate depending on who requests them and why.

To operationalise this insight, I rely on the formal apparatus of decision theory. Let A and H denote finite sets of actions and outcomes, respectively. An agent's preferences over action-outcome pairs induce a partial ordering that can be expressed as a utility function $u: A \times H \to \mathbb{R}$. Let $p(\cdot)$ denote a credence function over outcomes such that $p(H) = 1$ and $p(h|E)$ denotes the agent's subjective degree of belief in some $h \in H$ after conditioning on evidence E. The expected utility of each action $a \in A$ is given by a weighted average over hypotheses:

$$\mathbb{E}[u(a, H)|E] = \sum_j p\big(h_j\big|E\big)u\big(a, h_j\big). \quad (7.5)$$

If u satisfies the utility axioms (i.e., completeness, transitivity, continuity, and independence) and p satisfies the probability axioms (i.e., non-negativity, unit measure, and σ-additivity), then we say that the agent is *rational*, and it can be shown that she will tend to maximise

expected utility (von Neumann & Morgenstern, 1944).[3] That is, she will always choose at least one optimal action:

$$a^* = \underset{a \in A}{\operatorname{argmax}}\, \mathbb{E}[u(a, H)|E] \quad (7.6)$$

from a set of alternatives. The rationality in question is Bayesian insomuch as an agent operating in accordance with the axioms of the probability calculus will compute posteriors $p(h|E)$ from priors $p(h)$ and likelihoods $p(E)$ in accordance with Bayes's theorem.

This expected utility framework is very general, and applies to explanation instances of all four types cited above. For instance, in an auditing example like that described in Chapter 5, we might have $A = \{\texttt{sue}, \neg\texttt{sue}\}$ and $H = \{\texttt{biased}, \neg\texttt{biased}\}$, with evidence provided by some iML tool and fairness criteria (see §2.2.3). In algorithmic recourse, we typically consider more diverse action sets. For example, a credit scoring case may include $A = \{\texttt{education}, \texttt{income}, \texttt{credit}\}$, where increasing any subset of those variables incurs some cost but may improve outcomes $H = \{\texttt{approve}, \neg\texttt{approve}\}$.

A key consideration, often ignored by authors in iML but expanded upon at some length in Chapter 4, is the difference between model- and system-level explanations. The target level of abstraction (LoA) determines the relevant observables and assumptions underlying the analysis (Floridi, 2008a). At the model-LoA, for instance, it is typically appropriate to treat all features as mutually independent and observe outputs on real and/or synthetic inputs that vary in some systematic fashion. At the system-LoA, by contrast, we cannot afford to ignore causal relationships between features, as an intervention on one variable may have downstream effects on others. In these cases, nature constrains the input space in accordance with certain structural dependencies.

I formalise the distinction using methods from the causal modelling literature (Pearl, 2000; Peters et al., 2017). A structural causal model (SCM) is a tuple $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F} \rangle$, where $\boldsymbol{U}$ is a set of exogenous variables, i.e. background conditions; $\boldsymbol{V}$ is a set of endogenous variables, i.e. observed states; and $\mathcal{F}$ is a set of structural equations, one for each $V_j \in \boldsymbol{V}$. These functions map parents to children in a deterministic fashion, $f_j: pa(V_j) \rightarrow V_j$, where $pa(V_j) \in \boldsymbol{U} \cup \boldsymbol{V} \backslash \{V_j\}$ represents the set of all variables with direct causal influence on V_j. The model $\mathcal{M}$ implies an associated graph $\mathcal{G}$ that describes the topological ordering of variables. For instance, a child node is rendered independent of all its non-descendants after conditioning on its parent nodes, $V_j \perp \{\boldsymbol{U} \cup \boldsymbol{V} \backslash de(V_j)\} | pa(V_j)$. I restrict focus here to recursive models, which imply acyclic graphs. Pearl's *do*-calculus provides a provably complete set of rules for identifying causal effects from observational data where possible (Huang & Valtorta, 2006).[4]

[3] See Appendix B, §§2-3 for formal statements of the utility and probability axioms.
[4] See Appendix B, §4 for formal statements of the *do*-calculus axioms.

A deterministic SCM can be expanded to accommodate stochastic relationships by placing a joint probability distribution on background conditions, $p(\boldsymbol{U})$. By equating this with the credence function introduced above, we may encode an agent's uncertainty with respect to a causal system, which has implications for expected utility if $H \in \boldsymbol{V}$. This decision theoretic approach to causality is explicitly endorsed by Dawid (2012; 2015; 2020), who argues that treatment policies should be designed not by considering differences between so-called "potential outcomes" (Rubin, 1974), which involves counterfactual inference of allegedly dubious provenance, but instead by minimising expected loss as computed over a set of observed data. A similar premise underlies recent work in multi-armed bandit optimisation for SCMs (Lee & Bareinboim, 2018; 2019), where the goal is to find optimal interventions by maximising expected rewards over trials.

§7.3 Rational Shapley Values

My basic strategy is to allow user inputs to inform the generation of counterfactuals and estimation of Shapley values, thereby resulting in custom feature attributions that are more useful than either method on its own. Intuitively, this gives agents the tools to pose not just generic (i.e., objectivist) questions of the form "Why did model f predict outcome $\hat{y}_i$ for input $\boldsymbol{x}_i$?", but more targeted (i.e., pragmatic) questions such as:

- Why did f predict $\hat{y}_i$ as opposed to y' for $\boldsymbol{x}_i$?
- Why did f predict $\hat{y}_i$ as opposed to y' for $\boldsymbol{x}_i$, given $\boldsymbol{x}_i^S$?
- Why did f predict $\hat{y}_i$ as opposed to y' for $\boldsymbol{x}_i$, given $\boldsymbol{x}_i^S$ and certain constraints on $\boldsymbol{x}^R$?

Unlike counterfactuals, which may confuse users by generating a large number of synthetic datapoints with no clear takeaway message, the proposed method provides unambiguous weights for all variables of interest. Unlike feature attributions, which ignore pragmatic and causal information by construction, the present approach is *rational* in the decision theoretic sense. I will make these claims more precise and apparent below. First, however, I must introduce the notion of a *relevant subspace* to formalise the hierarchy of specification implied by the questions above.

Definition 7.1 (Relevant subspace). Let $\mathcal{Y}_0, \mathcal{Y}_1$ be a partition of $\mathcal{Y}$ into baseline and contrastive outcomes, respectively, with $\hat{y}_i \in \mathcal{Y}_0$. Let $S \subseteq [d]$ denote a (potentially empty) conditioning set with complement R. Let $\mathcal{D}_S$ be the space of permissible subvectors for the conditioning set, which reduces to the observed $\boldsymbol{x}_i^S$ in the limit. Alternatively, $\mathcal{D}_S$ may describe an ε-ball around $\boldsymbol{x}_i^S$, with perhaps some additional restrictions on the joint distribution of variables $\{X_j\}_{j \in S}$. Let $\mathcal{D}_R$ be a distribution encoding probable variation in $\boldsymbol{x}^R$ as a function of $\boldsymbol{x}^S$. We say that counterfactual sample $\tilde{\boldsymbol{z}} = (\tilde{\boldsymbol{x}}^S, \tilde{\boldsymbol{x}}^R, \tilde{y} = f(\tilde{\boldsymbol{x}}^S, \tilde{\boldsymbol{x}}^R))$ is *relevant* for the inquiring agent if and only if:

(i) $\tilde{y} \in \mathcal{Y}_1$,

(ii) $\tilde{\boldsymbol{x}}^S \sim \mathcal{D}_S$, and

(iii) $\tilde{\boldsymbol{x}}^R \sim \mathcal{D}_R$.

Any space $\tilde{\mathcal{Z}} \subseteq \mathcal{Z}$ that meets these criteria is a *relevant subspace.* ■

The issue with popular Shapley value approximators – e.g., SHAP (Lundberg & Lee, 2017) – is that baselines are fixed by the data centroid. Attributions are only designed to account for the discrepancy between $f(\boldsymbol{x}_i)$ and the empirical mean $\phi_0 = n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_i)$. Moreover, it is common to use the marginal $P(\boldsymbol{X}^R)$ as the reference distribution $\mathcal{D}_R$, thereby breaking associations between the features in S and R, resulting in a value function that implicitly compares each x_{ij} to the mean of the corresponding variable X_j. Yet these values might not fall within the relevant subspace for a given agent, who may, for instance, want to know why she received an average credit score instead of a better-than-average one. Or perhaps she wants to compare her predictions only to those of others in a similar age and income bracket, effectively setting attributions for those features to zero.

Counterfactuals naturally handle such cases, identifying real or synthetic datapoints that differ from the explanans in specific ways. However, there are several issues with the counterfactual approach. First, the cost function in Eq. 7.4 is almost always equated with some distance measure in practice, despite much discussion of more generic alternatives (Karimi et al., 2020; Ustun et al., 2019). This, I contend, is an unsatisfying notion of costs, which can and should reflect the preferences of inquiring agents much more flexibly. For instance, relevant action sets may range over interventions exogenous to the feature space, where the notion of distance is likely inapplicable. Second, there is the question of how many counterfactuals to provide. The optimisation problem expressed by Eq. 7.4 suggests that just a single sample would suffice; however, such a procedure overestimates the certainty with which such a datapoint is generated. Can we really be sure that counterfactual $\boldsymbol{x}^*$ is to be preferred to counterfactual $\boldsymbol{x}'$ just because the former is marginally closer in feature space to the input $\boldsymbol{x}_i$? Even if we are willing to accept a purely distance-based notion of cost, it is not at all obvious that very small distances should be taken seriously when there are so many potential sources of noise – in the measurements themselves, in the finite training data, and in whatever approximations were necessary to efficiently generate candidate counterfactuals. That is why some authors prefer to sample a large, diverse coalition of counterfactuals (Mothilal et al., 2020), so that users can survey the various paths through which a prediction may change. However, this approach raises new problems of intelligibility. How should an agent make sense of a relevant subspace with hundreds or thousands of samples within a tolerable distance of $\boldsymbol{x}_i$? How should she prioritise among different counterfactuals or integrate this information into a high-level summary?

Feature attribution methods are well suited to this task. Shapley values provide an efficient and intuitive summary of the information encoded in a large set of counterfactuals, regardless of sample size. I therefore propose a hybrid method in which data are sampled from a relevant subspace, and Shapley values computed only with respect to such samples. The crucial observation is that nothing in the definition of Shapley values (Eqs. 7.2 and 7.3) precludes us from shifting the reference distribution from an entire observational dataset to a particular region with certain desirable properties – e.g., a relevant subspace. The convention of using the empirical mean of all remaining points is just that – a convention. In fact, Shapley values can be computed on a pairwise basis, where attributions account for the difference between any two predictions, $f(\boldsymbol{x}_i), f(\boldsymbol{x}_{i'})$. Averaging over the full set of such points recovers the original Shapley values; conditioning on a particular subset, on the other hand, focuses the analysis.

§7.3.1 Selecting the Reference Distribution

I take it more or less for granted that agents requesting explanations for algorithmic predictions have some contrastive outcome $\mathcal{Y}_1$ in mind. They are also likely to have some intuition about the conditioning set S, which will presumably include immutable features and/or any variables that the agent does not want to enter into the explanation if it can be avoided. The precise distribution for these features may not be available *a priori*, but the observed $\boldsymbol{x}_i^S$ is probably a good start when the data allow for it. The trickiest component to define is then the reference distribution $\mathcal{D}_R$, which has implications both for how counterfactuals are computed (since this helps define the relevant subspace) and for how Shapley values are estimated (since this helps define the value function). I consider three possibilities of increasing complexity:

(1) Marginal: $\mathcal{D}_R = P(\boldsymbol{X}^R)$

(2) Conditional: $\mathcal{D}_R = P(\boldsymbol{X}^R|\boldsymbol{x}^S)$

(3) Interventional: $\mathcal{D}_R = P\big(\boldsymbol{X}^R\big|do(\boldsymbol{x}^S)\big)$

Note that option (1) is occasionally referred to as the "interventional" Shapley value (Janzing et al., 2020; Chen et al., 2020) – a misnomer based on the "simplifying assumption" that all features are mutually independent, in which case (3) reduces to (1). Since feature independence is a strong and often unjustified assumption, I will stick to the terminology above.

The first point to observe is that the three options are generally not equivalent. Consider a toy example, in which our goal is to evaluate the probability of recovery $\boldsymbol{X}^R$ in an observational study. The marginal $\mathcal{D}_R$ is just the base rate – what proportion of patients in the trial recover? Imagine, for the sake of concreteness, that exactly half of subjects recover, in which case $P(\boldsymbol{X}^R) = 0.5$. But some patients were given a placebo ($\boldsymbol{X}^S = 0$), and none of those subjects recovered. In that case, we would expect a higher value for the conditional $\mathcal{D}_R$, say

$P(\boldsymbol{X}^R|\boldsymbol{X}^S = 1) = 0.9$. However, treatment groups in this observational study were not randomly assigned. If some unobserved confounder has causal effects on both $\boldsymbol{X}^S$ and $\boldsymbol{X}^R$ – for instance, if only men were given the active treatment, and men have better odds of recovery in general – then the observed conditional probability may seriously overstate the causal efficacy of the treatment, which is given by the interventional $\mathcal{D}_R$, perhaps $P\left(\boldsymbol{X}^R\middle|do(\boldsymbol{x}^S)\right) = 0.75$.

I want to be clear that the differences between these distributions are in no sense normative. None is generally superior to the others, or more useful, or more broadly applicable. Rather, each is the right answer to a different question. What follows is a brief discussion of the advantages and disadvantages associated with each value of $\mathcal{D}_R$, as well some heuristic guidelines on when to use which.

Marginal. The marginal distribution is preferable when the goal is a model-LoA explanation. This may be the case in certain instances of auditing or validation, where the analyst seeks merely to discover what the model has learned without any further restrictions. In this case, we may not care too much about impossible data perturbations – e.g., teenage grandmothers or pregnant males – since we simply want to recreate a decision boundary in model space. The extent to which this model space matches reality is another question altogether, presumably one that data scientists should take seriously, both before and throughout deployment. Once the model is in the wild, however, it may be of independent interest to users and regulators how it performs on a wide range of hypothetical cases, including those off the true data manifold.

Conditional. The conditional distribution is better suited to explanations at the system-LoA, where impossible data permutations could lead to issues. This may be the case, for instance, in certain instances of validation, where we do not wish to punish a model for failing to extrapolate to data points far from its training set. Such methods are especially valuable when joint distributions may be estimated with reasonable accuracy and causal relationships are unknown or potentially inapplicable (e.g., in image classification tasks). Estimating joint densities from finite datasets is a hard problem in general, however parametric assumptions can ease the burden (Aas et al., 2019). For continuous data, several algorithms can achieve universal approximations – e.g., Gaussian processes (Rasmussen & Williams, 2006) and some kernel density estimators (Gramacki, 2018) – but the task grows exponentially more difficult as dimensionality increases. Moreover, mixed data types still pose problems for such methods. A simple, if somewhat crude option is to simply train a new supervised learner of any function class to estimate $\mathbb{E}\left[X_j\middle|\boldsymbol{X}^R\right]$ for each $j \in S$, but this requires $d(d-1)$ unique models, a feat that quickly becomes infeasible in even moderate dimensions.

Interventional. The interventional distribution is optimal when seeking explanations at the system-LoA for causal data generating processes. With access to the underlying SCM, there can be no more accurate estimator than that defined by (3). These methods are required

when seeking to use iML for discovery and/or planning, as both sorts of actions invariably rely on real-world mechanisms that cannot be approximated by either a purely marginal approach (which ignores all dependencies) or a conditional one (which fails to distinguish between correlation and causation). Of course, complete causal information is almost never available, which means that in practice analysts must either (a) rely on partial orderings, (b) place some distribution over an equivalence class of models consistent with the data, (c) estimate unknown causal relationships, (d) make causal assumptions that may not be testable, or (e) some combination of the above.

I employ marginal, conditional, and interventional value functions in all experiments below (§7.4.1). To compute interventional Shapley values, I take option (a) and rely on partial orderings, as these are often the most readily available to analysts at test time. A partial ordering of features implies a chain graph in which links are composed of all variables in a single causal group (Lauritzen & Richardson, 2002). For instance, users may not know the complete graph structure of the set {`age`, `sex`, `income`, `savings`, `credit`} but they may be more confident about the partial ordering {`age`, `sex`} → {`income`, `savings`} → {`credit`}. The probability distribution for a DAG of chain components can be factorised by iterative conditioning on parents:

$$P(\boldsymbol{X}) = \prod_{g \in G} P(\boldsymbol{X}_g | \boldsymbol{X}_{pa(g)}), \quad (7.7)$$

where g denotes a given subset of features constituting a single causal group and G is the total collection of such groups. With slight abuse of notation, we write $pa(g)$ for the index of all features that are parents of any variable in g. Computing causal effects in such a system requires some care with respect to intra-group dynamics. Specifically, we must distinguish between those chain components that do and do not contain common confounders, denoted hereafter by G_1 and G_0, respectively. When confounders are absent, any surplus dependencies are presumed to arise from the relationships between features within a given group. Now it can be shown with just a few steps of the *do*-calculus that causal effects in chain graphs of the sort implied by partial orderings can be computed from conditional dependencies. The following theorem builds on foundational work in chain graph theory (Lauritzen & Richardson, 2002) and closely tracks results independently derived by Heskes et al. (2020).

Theorem 7.1. *Interventions on feature subsets in causal chain graphs can be computed from observational data as follows:*

$$P\big(\boldsymbol{X}^R \big| do(\boldsymbol{x}^S)\big) = \prod_{g \in G_1} P\big(\boldsymbol{X}_{g \cap R} \big| \boldsymbol{X}_{pa(g) \cap R}, \boldsymbol{x}_{pa(g) \cap S}\big) \times \prod_{g \in G_0} P\big(\boldsymbol{X}_{g \cap R} \big| \boldsymbol{X}_{pa(g) \cap R}, \boldsymbol{x}_{pa(g) \cap S}, \boldsymbol{x}_{g \cap S}\big), \quad (7.8)$$

where $\boldsymbol{X}_{g \cap R}$ denotes the variables in group g that intersect with subset R.

Proof. The result follows from

$$P\left(\boldsymbol{X}^R \middle| do(\boldsymbol{x}^S)\right) =^{(1)} \prod_{g \in G} P\left(\boldsymbol{X}_{g \cap R} \middle| \boldsymbol{X}_{pa(g) \cap R}, do(\boldsymbol{x}^S)\right)$$

$$=^{(3)} \prod_{g \in G} P\left(\boldsymbol{X}_{g \cap R} \middle| \boldsymbol{X}_{pa(g) \cap R}, do\left(\boldsymbol{x}_{pa(g) \cap S}\right), do\left(\boldsymbol{x}_{g \cap S}\right)\right)$$

$$=^{(2)} \prod_{g \in G} P\left(\boldsymbol{X}_{g \cap R} \middle| \boldsymbol{X}_{pa(g) \cap R}, \boldsymbol{x}_{pa(g) \cap S}, do\left(\boldsymbol{x}_{g \cap S}\right)\right),$$

where the number associated with each equality sign denotes the corresponding *do*-calculus rule (see Appendix B, §4). For $g \in G_1$, we apply rule (3) to yield $P\left(\boldsymbol{X}_{g \cap R} \middle| \boldsymbol{X}_{pa(g) \cap R}, \boldsymbol{x}_{pa(g) \cap S}\right)$; for $g \in G_0$, we apply rule (2) to yield $P\left(\boldsymbol{X}_{g \cap R} \middle| \boldsymbol{X}_{pa(g) \cap R}, \boldsymbol{x}_{pa(g) \cap S}, \boldsymbol{x}_{g \cap S}\right)$. ■

§7.3.2 A New Axiom

Say two agents are identical along all recorded variables for some credit scoring function, and both are assigned lower scores than they had hoped. Counterfactuals identify two minimal perturbations sufficient to improve predictions: (1) increase education by one unit or (2) increase savings by one unit. Let us stipulate that these actions are equivalent in terms of time and money. However, they differ in another important respect – the first agent wants to go back to school and the second wants to save more each month. In this case, *preferences alone* determine which of two explanations is optimal.

We may devise a similar example to show the impact of differing credences. Say, for instance, that two labs are planning expensive gene knockout experiments to test the predictions of a gene regulatory network inference algorithm. Their utilities may be identical – the goal for both groups is to maximise discoveries while minimising false positives – but they are working with different sets of evidence after preliminary, as yet unpublished results indicated to the first lab that one set of pathways is a probable dead end. In this case, *beliefs alone* determine which of two explanations is optimal.

What these examples demonstrate is that utilities and credences matter when attempting to explain model predictions. With default software like SHAP, feature attributions for these credit loan applicants or research labs would not generally differ – same inputs, same outputs. However, by explicitly incorporating pragmatic information, we can secure custom explanations at no cost to the axiomatic guarantees of Shapley values. In fact, we can extend the current desiderata.

Observe that for fixed model f and input $\boldsymbol{x}_i$, feature attributions $\boldsymbol{\phi} = \{\phi_j\}_{j=1}^d$ vary only as a function of contrastive baseline $\tilde{\boldsymbol{z}}$. I make this dependence explicit moving forward, writing $\boldsymbol{\phi}(\tilde{\boldsymbol{z}})$. Define a reward function for the inquiring agent as the expected utility of the utility maximising action a^*, conditional on a given set of feature attributions:

$$r(\tilde{\boldsymbol{z}}) = \mathbb{E}[u(a^*, H)|\boldsymbol{\phi}(\tilde{\boldsymbol{z}})]. \quad (7.9)$$

The axiomatic constraint that characterises pragmatic feature attributions can now be stated as follows.

Rationality. Let $r(\cdot)$ be the reward function for a rational agent, i.e. one who behaves in accordance with the utility and probability axioms. Let $\boldsymbol{\phi}(\tilde{\boldsymbol{z}})$ be a feature attribution vector computed with respect to subspace $\tilde{\mathcal{Z}}$. Then $\boldsymbol{\phi}(\tilde{\boldsymbol{z}})$ is rational if and only if, for any alternative subspace $\mathcal{Z}'$, $\mathbb{E}_{\tilde{\boldsymbol{z}}\sim\tilde{\mathcal{Z}}}[r(\tilde{\boldsymbol{z}})] \geq \mathbb{E}_{\boldsymbol{z}'\sim\mathcal{Z}'}[r(\boldsymbol{z}')]$. In other words, feature attributions are rational to the extent that they tend to maximise expected rewards for the inquiring agent.

Theorem 7.2. *When the relevant subspace is nonempty, rational Shapley values are the unique additive feature attribution method that satisfies efficiency, linearity, sensitivity, symmetry, and rationality.*

Proof. Since uniqueness is already well established for the classical axioms, all that remains is to show that rational Shapley values (i) satisfy the additional rationality axiom and (ii) do not violate any of the classical axioms.

Take (i) first. Assume there exists some $\mathcal{Z}'$ that generates greater expected rewards for an agent than the relevant subspace $\tilde{\mathcal{Z}}$. Then conditioning on the corresponding Shapley values $\boldsymbol{\phi}(\boldsymbol{z}')$ must lead to an action of greater expected utility than the action recommended by $\boldsymbol{\phi}(\tilde{\boldsymbol{z}})$. This means that the values for some $\phi_j(\tilde{\boldsymbol{z}})$ are either too big or too small. But this can only arise from a misspecification of $\mathcal{D}_S$ or $\mathcal{D}_R$, which permits either too much variation in X_j (resulting in inflated values of $|\phi_j(\tilde{\boldsymbol{z}})|$), or too little (resulting in deflated values of $|\phi_j(\tilde{\boldsymbol{z}})|$). By definition, the relevant subspace properly specifies both distributions, as well as the contrastive outcome $\mathcal{Y}_1$. Thus we have a contradiction.

One can verify (ii) by confirming that rational Shapley values do not deviate from the classical formulae (see Eqs. 7.1, 7.2, 7.3). The only difference between the current proposal and more familiar alternatives is their respective methods of specifying baseline expectations ϕ_0 and reference distributions $\mathcal{D}_R$. Shifting these two parameters changes the interpretation of resulting Shapley vectors, but does not alter their fundamental properties. Any procedure that inputs valid values for ϕ_0 and $\mathcal{D}_R$ and conforms to the characteristic equations enjoys the same properties. ■

§7.4 Rational Shapley Value Algorithm

I reframe the objective function of Eq. 7.4 with a slight twist. Our goal is not to minimise the cost or distance between two individual datapoints, but rather to compute a subspace that maximises rewards. That is:

$$\mathcal{Z}^* = \underset{\tilde{\mathcal{Z}}\subseteq\mathcal{Z}}{\operatorname{argmax}}\, \mathbb{E}_{\tilde{\boldsymbol{z}}\sim\tilde{\mathcal{Z}}}[r(\tilde{\boldsymbol{z}})]. \quad (7.10)$$

This differs from (and improves upon) Eq. 7.4 in several respects. First, it results not in a single point but in a region of space, which is stabler and more informative. Second, it explicitly incorporates preferences and credences via the functions that define $r(\cdot)$. Finally, since rewards are conditioned upon a feature attribution vector, this target combines elements of both counterfactuals and feature attributions, which is preferable to either alternative alone for all the reasons argued above.

There are many ways one could go about computing this subspace in practice, and no single method is optimal in general. Popular generative examples from the counterfactual literature include GANs (Wachter et al., 2018), mixed integer linear programs (Ustun et al., 2019), and SAT-solvers (Karimi et al., 2020). With a bit of work, the problem could be reformulated to allow for other strategies, such as evolutionary algorithms (Yu & Gen, 2010), multi-armed bandits (Lattimore & Szepesvári, 2019), or reinforcement learning (Sutton & Barto, 2018). A simpler solution, which I use in all experiments below, is to rely on observational samples (Wexler et al., 2020). These subspaces have the advantage of being fast to compute and guaranteed to lie on the true data manifold (up to potential sampling or measurement errors). Moreover, they avoid the assumptions and approximations of the aforementioned optimisation techniques. The main disadvantage of this method is that it may provide low coverage for regions of the feature space where data are undersampled. However, this poses issues for other approaches as well, as no method can confidently draw realistic points from low-density regions of the feature space. In any event, the problem is not too vexing when datasets are sufficiently large and representative.

The basic strategy is schematised in pseudocode below. The user postulates a set of candidate subspaces, computing a sequence of rational Shapley vectors and comparing relative rewards. Evaluating $r(\tilde{\boldsymbol{z}})$ with precision requires explicit utilities and (conditional) credences, which may not be generally available. This can be done informally by comparing feature attribution vectors on an ordinal basis, such that rewards are ranked rather than directly quantified. This procedure is equivalent to the cardinal alternative under certain assumptions about agentive rationality, as noted above. The experiments below demonstrate how such heuristic behaviour may guide an exploratory analysis that is potentially more informative than a narrow optimisation approach.

Notably lacking from this algorithm are two components: (1) any guidance on how to select candidate subspaces; and (2) any guidance on how to update credences based on rational Shapley vectors. These omissions are by design. They allow a human-in-the-loop approach whereby agents may navigate subspaces and incorporate new information in accordance with their individual goals, a process that would be difficult or impossible to generally codify given the range of possible explanation tasks for which the method is applicable. Several examples are worked through below demonstrating how the method works in practice.

Algorithm 7.1: Rational Shapley Values

Inputs: Input datapoint $\boldsymbol{x}_i$, set of candidate subspaces $\{\langle \mathcal{D}_S, \mathcal{D}_R, \mathcal{Y}_1 \rangle\}_{k=1}^{m}$, utility function u, credence function p.

6. Compute classical Shapley values $\boldsymbol{\phi}(\boldsymbol{z})$ using mean $\phi_0 = \bar{y}$ and reference $\mathcal{D}_R$.
7. Initialise $r(\boldsymbol{z}) = \sum_j p\left(h_j \middle| \boldsymbol{\phi}(\boldsymbol{z})\right) u\left(a^*, h_j\right)$.
8. **for each** $k = \{1, \ldots, m\}$ **do**:
    4. Draw data $\tilde{\boldsymbol{z}}_k \sim \langle \mathcal{D}_S, \mathcal{D}_R, \mathcal{Y}_1 \rangle_k$.
    5. **if** $\tilde{\boldsymbol{z}}_k = \emptyset$ **then** FAIL. **else:**
    6. Compute $\boldsymbol{\phi}(\tilde{\boldsymbol{z}}_k)$.
    7. Record $r(\tilde{\boldsymbol{z}}_k) = \sum_j p\left(h_j \middle| \boldsymbol{\phi}(\tilde{\boldsymbol{z}}_k)\right) u\left(a^*, h_j\right)$.
8. Define $\tilde{\boldsymbol{z}}^* = \underset{\tilde{\boldsymbol{z}}}{\text{argmax}}\ r(\tilde{\boldsymbol{z}})$.

Output: $\boldsymbol{\phi}(\tilde{\boldsymbol{z}}^*)$.

A complexity analysis of Algorithm 7.1 cannot be achieved without making further assumptions about the range of candidate subspaces, which could be exponential in the number of features if left totally unconstrained. In practice, it is likely that agents will have some intuitions regarding membership in S and R. When the latter is multidimensional, graphical methods can be used to prune the search space for optimal actions, assuming an interventional reference distribution (Lee & Bareinboim, 2018; 2019). Parametric assumptions can further simplify matters and often work well in practice. Convergence conditions are likewise elusive without further restrictions on the shape of the reward function, the counterfactual sampling procedure, and/or the data generating process.

§7.4.1 Experiments

In this section, I describe results from a number of experiments on benchmark datasets. For reasons of space and clarity, I restrict the search for a relevant subspace to the empirical dataset. I compare performance against baselines using marginal, conditional, and interventional value functions, respectively labelled MSV, CSV, and ISV. Shapley values are computed using code adapted from the `fastshap`, `treeshap`, and `shapr` packages, which had to be extended to accommodate mixed data types and structural information encoded in partially ordered causal graphs. Complete code for reproducing all experiments and figures can be found at https://github.com/dswatson/dissertation.

§7.4.1.1 Auditing: COMPAS algorithm

The COMPAS algorithm is a statistical model used to assign risk scores to defendants awaiting trial. High risk scores are associated with elevated probability of recidivism, which judges in nine US states use to help decide whether to let defendants out on bail while awaiting trial. A 2016 report by ProPublica alleged that the COMPAS algorithm was racially biased against African Americans (Angwin et al., 2016); subsequent analysis by independent researchers has not always corroborated those findings (Fisher et al., 2019; Rudin, Wang, & Coker, 2020).

Though the creators of COMPAS have published technical reports defending their model, they have made neither their training data nor code publicly available, so attempts to recreate the algorithm and audit for racial bias are typically based on data gathered by ProPublica on some 12,000 individuals arrested in Broward County, Florida between 2013 and 2014. I follow the pre-processing steps on the ProPublica GitHub page,[5] which exclude a little over a third of defendants on various grounds, and additionally limit my analysis to the 5,278 individuals identified as African American or Caucasian to avoid sparsity in other racial categories. I regress violent risk scores on the following features: age, number of priors, and whether the present charge is a felony (all deemed "admissible"); as well as race and sex (deemed "inadmissible"). I use a random forest to recreate the regression surface, with default parameters from the `ranger` package (Wright & Ziegler, 2017). I log-transform the skewed features (age and priors) before modelling. This has no impact on predictions, as decision trees are invariant to monotone transformation, but improves conditional probability estimates for the reference distribution. The fit is good but imperfect – the model explains approximately 54% of the variance on out-of-bag samples, with a mean square error of 2.9 – a discrepancy that is likely attributable to the fact that the true COMPAS model uses covariates beyond those in the ProPublica dataset. However, for the purposes of this experiment, I take the trained model as the true target. Conclusions based on this analysis should be interpreted with caution, as they pertain not to the original COMPAS algorithm but to an approximation thereof.

MSV is computed via Monte Carlo (Štrumbelj & Kononenko, 2014), with 2000 simulations per attribution. CSV is computed using the TreeSHAP method (Lundberg et al., 2020), since random forests are composed of individual regression trees. For ISV, I presume a simple partial ordering in which demographic variables (age, sex, and race) are root nodes, upon which all other predictors depend. The mean predicted response for all defendants in the dataset is $\phi_0 = 3.69$.

I focus specifically on high-risk defendants, namely the 264 subjects with risk scores at or above the 95th percentile according to the COMPAS approximator. Some 94% of these defendants are African American, compared to just 60% in the general dataset, which immediately raises questions about racial bias. MSV, CSV, and ISV all look fairly similar in this case (see Fig. 7.1). Age and priors generally receive the highest feature attributions in this group, consistent with the findings of other researchers. Race also receives nonzero attribution throughout. Notably, race receives higher attribution under the ISV reference than MSV, since MSV can only detect direct effects and race is presumed to have both direct and indirect effects

[5] See https://github.com/propublica/compas-analysis.

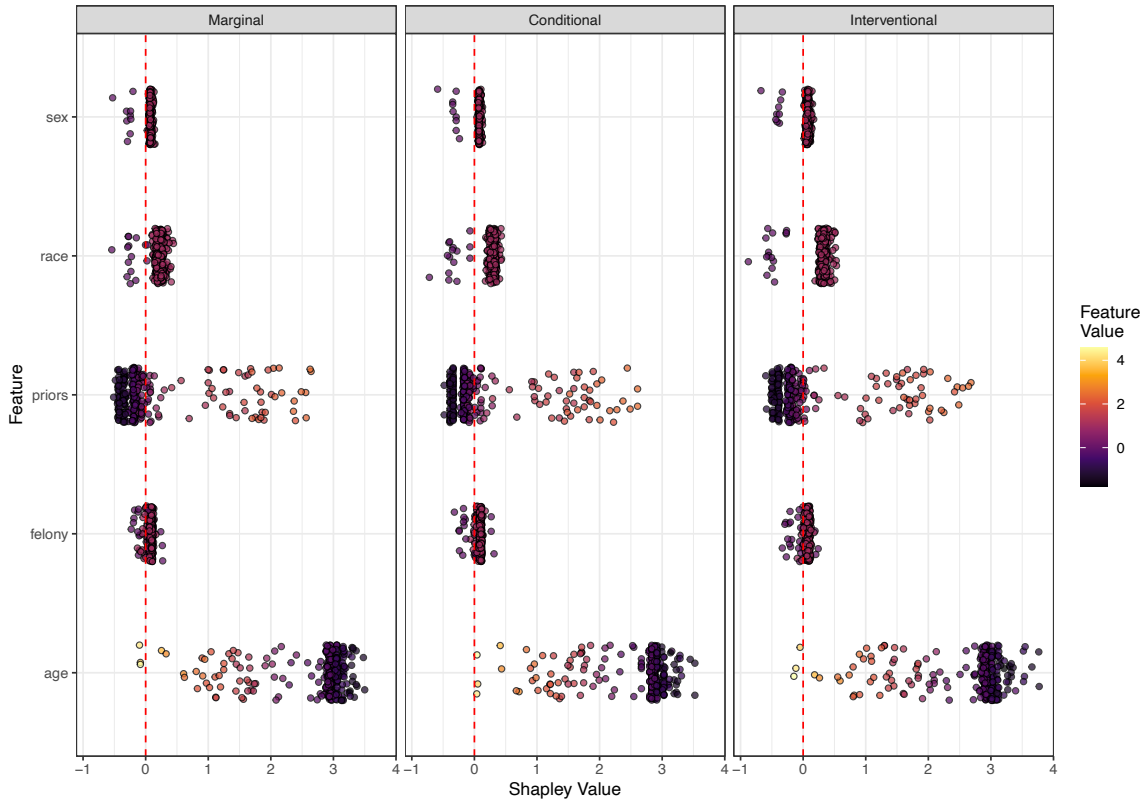

Figure 7.1. Classical Shapley values for the Broward County COMPAS dataset, computed with marginal, conditional, and interventional reference distributions. Continuous feature values are *z*-scored for visualisation.

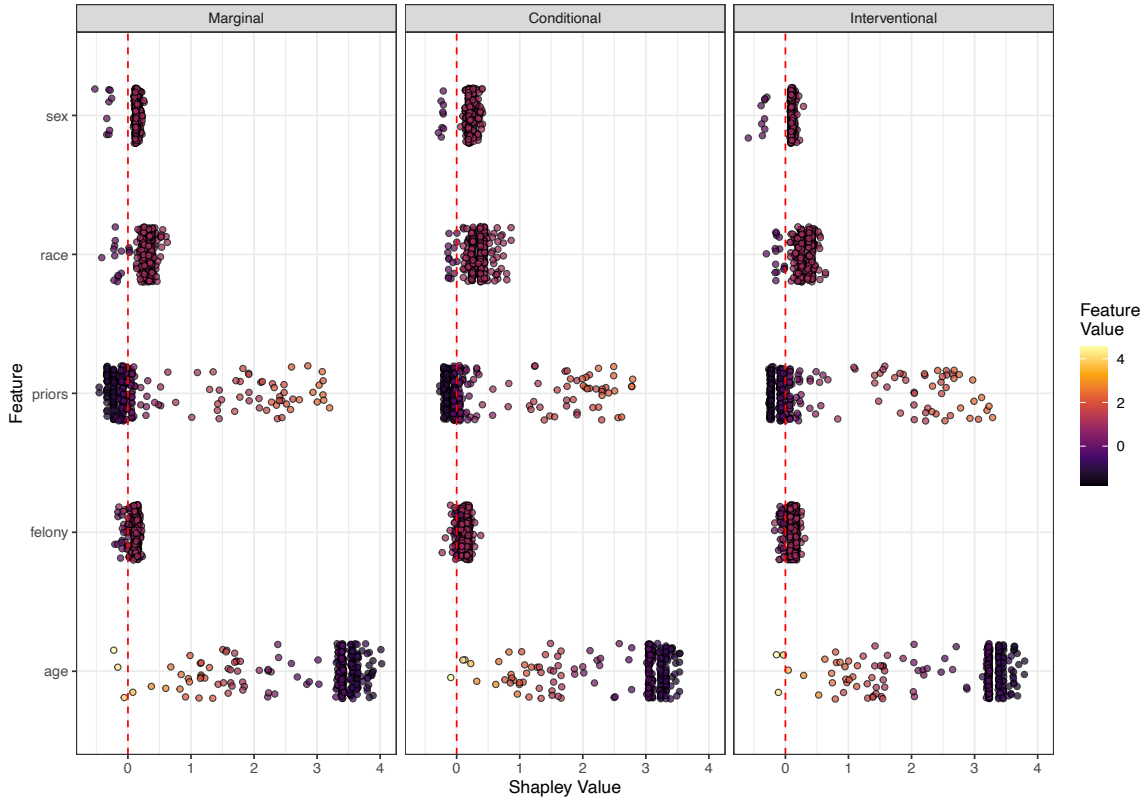

Figure 7.2. Rational Shapley values for the Broward County COMPAS dataset, computed with marginal, conditional, and interventional reference distributions. Continuous feature values are *z*-scored for visualisation.

given the partial ordering assumed for this analysis. In this example, I presume that auditors are working at the model-LoA – i.e., testing the extent to which the algorithm has learned certain patterns, independent of any natural systems – in which case the marginal distribution is best suited to their purpose.

To compute rational Shapley values, I consider a relevant subspace of defendants in the same age range (at most 33 years old) with at least one prior, yet who fell in the bottom half of defendants by predicted risk score. All other features are allowed to vary unconstrained. 356 subjects in the data meet these criteria, and therefore constitute the reference sample for this experiment. This is too large a group of counterfactuals to "explain" anything on their own, even if they were ranked by some measure (e.g., distance from the target input). What the task requires is a summary of differences across features, a job for which Shapley values are uniquely well-suited. The racial breakdown of defendants in the relevant subspace is markedly different from our high-risk group – just 48% are African American – further reinforcing concerns about potential bias in the COMPAS data. The resulting Shapley values look fairly similar to their classical counterparts, with a slight but notable increase in average attributions for race (see Fig. 7.2). Boxplots comparing Shapley values across racial groups confirm that African Americans consistently face positive attribution (i.e., increased risk) due to race, while Caucasian defendants generally receive negative Shapley values for this same variable. A series of Kruskal-Wallis tests confirms that the differences are highly significant in all six reference sets ($p < 0.001$). This is strong evidence of racial bias.

However, the evidence is stronger using rational Shapley values than it is under the classical alternative. Under the marginal reference, the racial attribution gap between Black and white defendants is significantly greater in the relevant subspace than it is across the complete dataset ($t = 3.874, p < 0.001$). Interestingly, this gap tends to increase as we shift from marginal to conditional, and conditional to interventional reference distributions in both classical and rational Shapley values; however, attributions for African Americans are more constant in the relevant subspace. See Figs. 7.3 and 7.4.

Consider the case of a young Black defendant deciding whether or not to file suit against the makers of COMPAS. Despite having just a single prior offense, this 21-year-old was placed in the highest risk group according to the algorithm. A plausible utility matrix for this individual is given in Table 7.1. Assume that the defendant in question assigns a uniform prior over H to begin with. We say that h_1 is corroborated to the extent that race receives greater positive attribution for African Americans than it does for defendants of other races. Then although both classical and rational Shapley values recommend the same action (a_1), the latter does so with higher expected reward since $p\big(h_1 \big| \boldsymbol{\phi}(\tilde{z})\big) > p\big(h_1 \big| \boldsymbol{\phi}(z)\big)$. Rational Shapley values are therefore preferable in this case, as theory suggests.

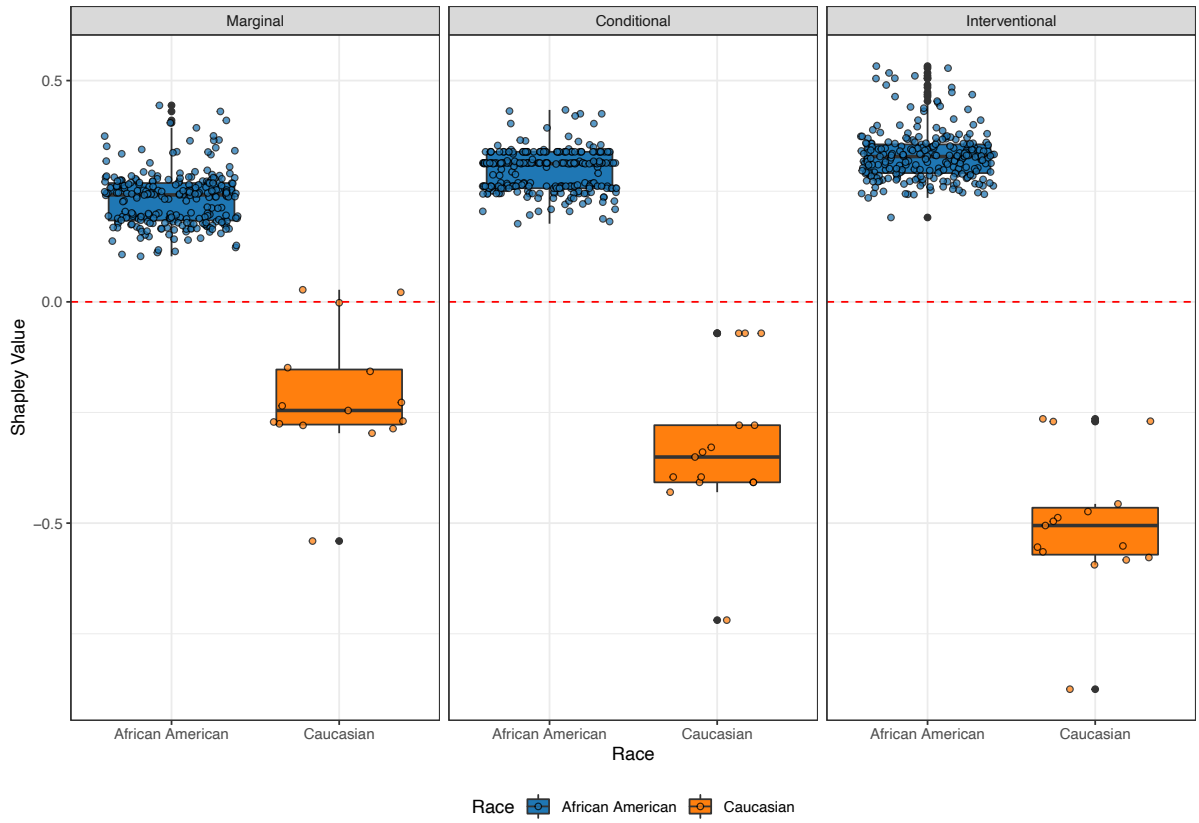

Figure 7.3. Boxplot of classical Shapley values by race for all three reference distributions.

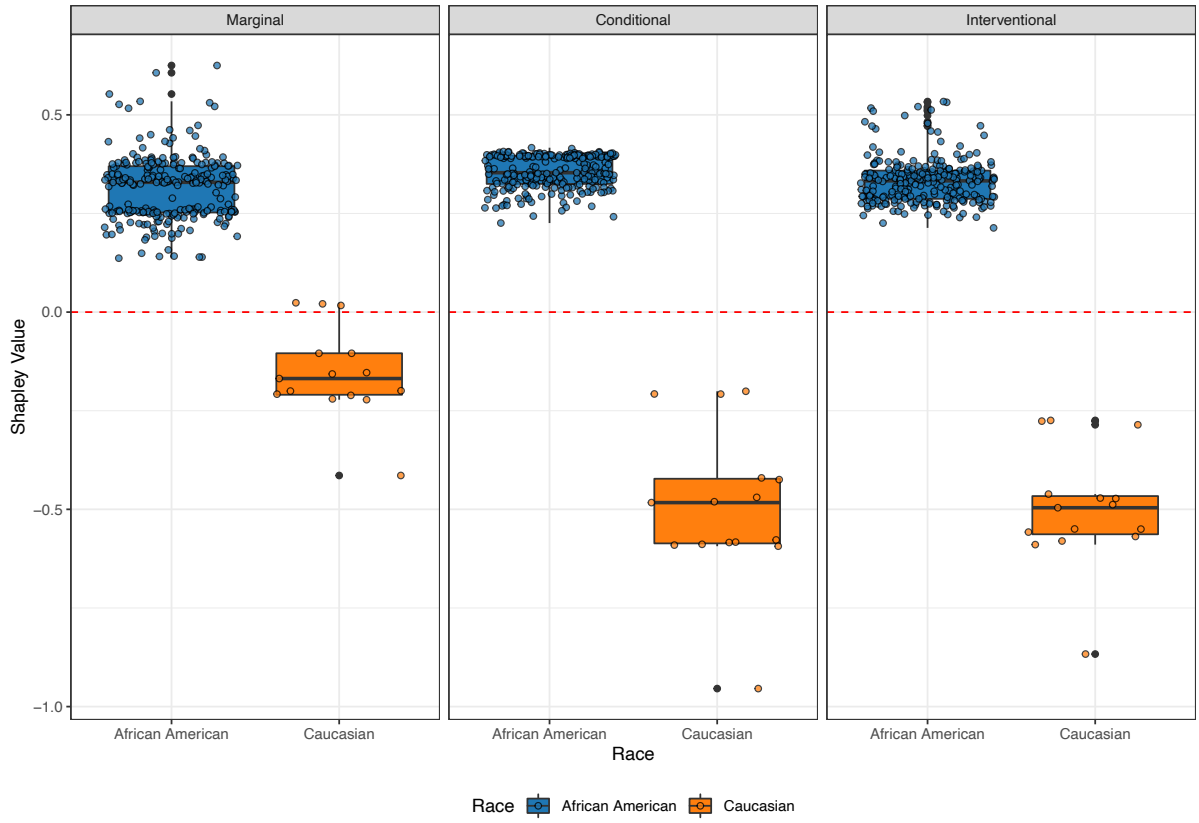

Figure 7.4. Boxplot of rational Shapley values by race for all three reference distributions.

	h_1: biased	h_2: ¬biased
a_1: sue	5	−1
a_2: ¬sue	0	0

Table 7.1. Utility matrix for a 21-year-old African American defendant with a single prior, predicted to be high risk and deciding whether to sue the makers of COMPAS.

§7.4.1.2 Discovery: Medical Diagnosis

As an example of iML for discovery, I consider the diabetes dataset originally described by (Efron et al., 2004). The data consist of 442 patients and 10 predictors – age, sex, body mass index (BMI), blood pressure (MAP), and six blood serum measurements, including three cholesterol-related variables (LDL, HDL, and TC). The goal is to predict disease progression after one year. Variables are normalised to mean zero and unit variance prior to modelling, per the design of Efron et al. I use the `glmnet` package (Friedman et al., 2010) to fit an elastic net regression to the data, a form of regularised linear model that combines L_1 and L_2 penalties on the coefficients (Zou & Hastie, 2005). Hyperparameters are tuned via 10-fold cross-validation. Note that exact Shapley values can be analytically calculated for linear models under the assumption of feature independence. However, exploratory data analysis reveals strong correlations between some features in the data (especially blood serum measurements), and causal dependencies between features such as BMI and MAP are well established. In such a case, the analytic formulae for local explanation in linear models are inapplicable.

MSV is estimated via Monte Carlo with 2000 simulations. CSV is computed via the multivariate Gaussian method of Aas et al. (2019) for all continuous variables; the lone binomial feature (sex) is estimated via the empirical distribution. For ISV, I assume the partial ordering {`age`, `sex`, `bmi`} → {`map`, `ldl`, `hdl`, `tc`} → {`tch`, `ltg`, `glt`}. This treats age, sex, and BMI as root nodes; MAP and cholesterol as causally intermediate; and remaining blood serum measurements are downstream. Such a DAG admittedly oversimplifies several complex biochemical processes but it is broadly consistent with known structural relationships.

I focus on patients with especially poor prognoses, as these are typically the subjects of greatest concern to clinicians. Specifically, I examine the top decile of patients by disease activity ($\hat{y} \geq 265$). Classical Shapley values for these patients are visualised under all three value functions in Fig. 7.5. Conditional and interventional Shapley values are fairly similar here, although both are quite different from the marginal alternative. For instance, TC receives relatively large attributions under the marginal reference distribution, but nearly none under conditional or interventional distributions, owing to its strong collinearity with other cholesterol measures.

For rational Shapley analysis, I consider a subspace consisting of older men (i.e., those in the top quartile by age), who are generally at higher risk than other groups in this dataset.

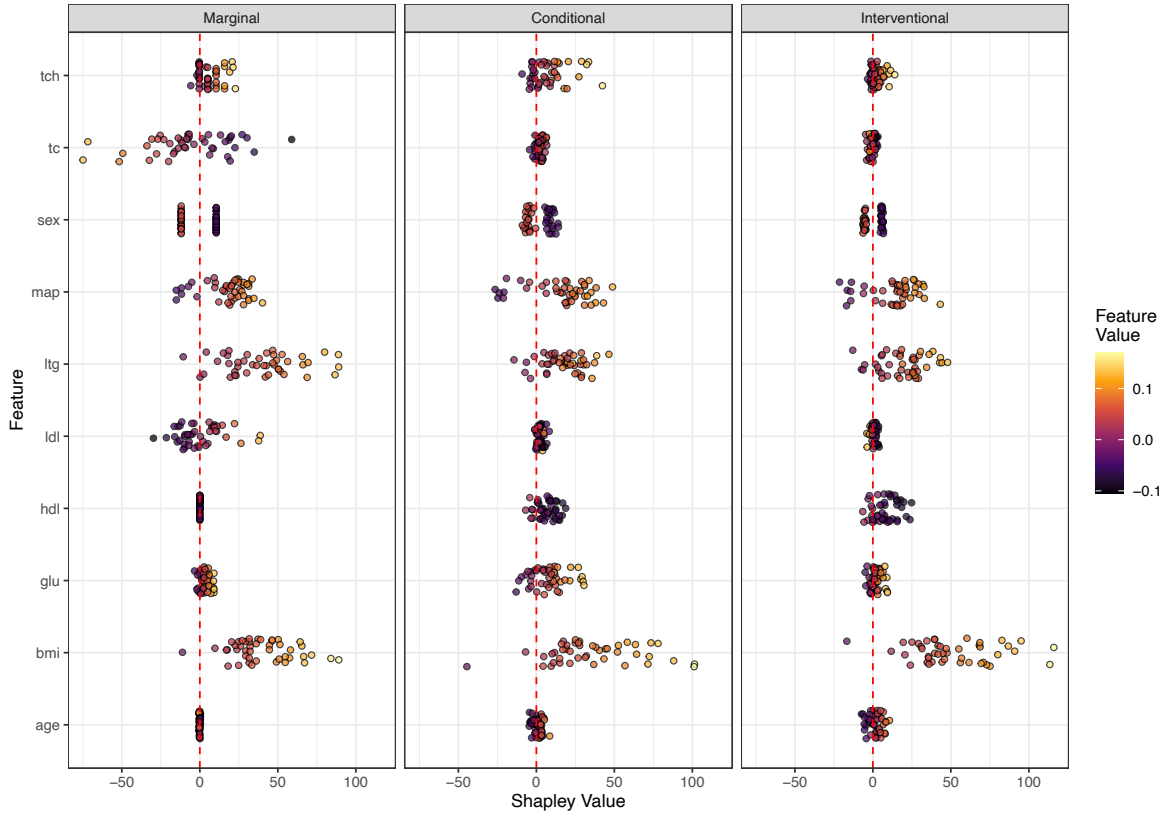

Figure 7.5. Classical Shapley values for the diabetes dataset, computed with marginal, conditional, and interventional reference distributions. Continuous feature values are z-scored for visualisation.

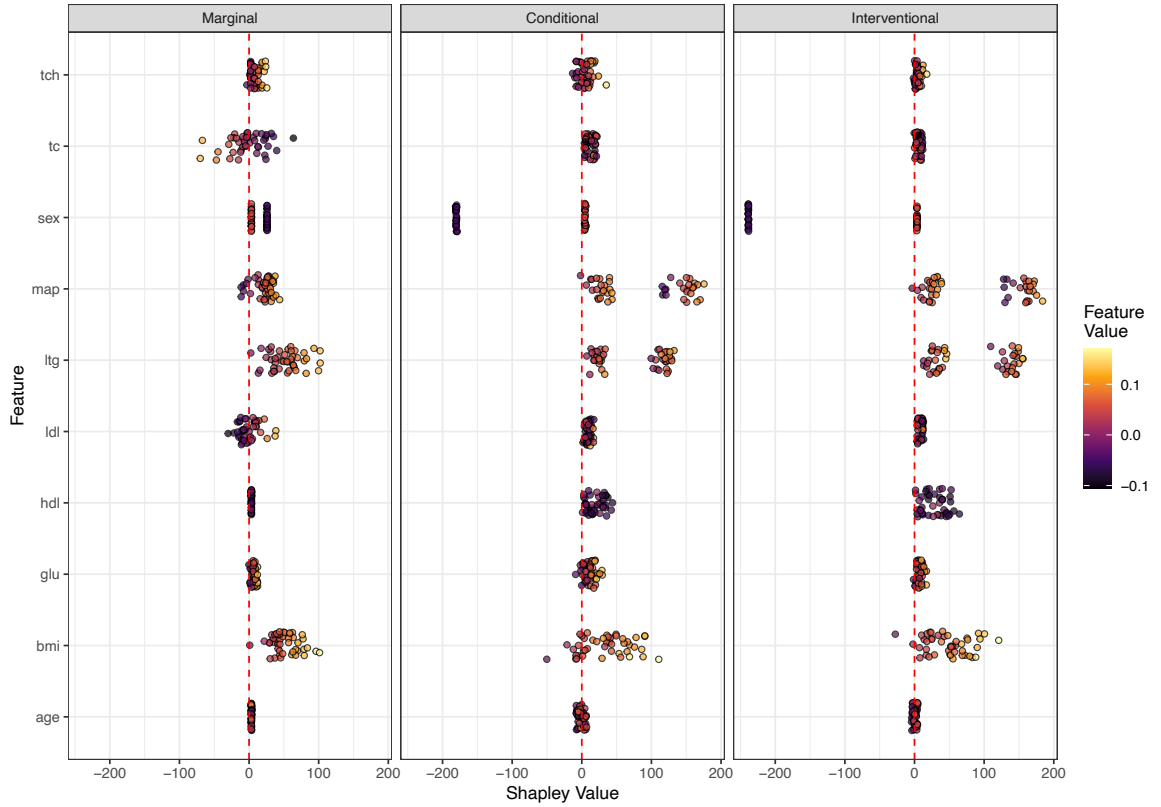

Figure 7.6. Rational Shapley values for the diabetes dataset, computed with marginal, conditional, and interventional reference distributions. Continuous feature values are z-scored for visualisation.

Age and sex are (more or less) unactionable, and so a rational agent will likely want prognostic explanations that place little or no weight on such features. Rational Shapley results are visualised in Fig. 7.6, where we find that BMI and MAP receive large attributions on average, reflecting their greater variance in the relevant subspace. Conditional and interventional reference distributions are especially valuable here, as the analytical goal almost certainly requires a system-LoA approach and feature independence cannot reasonably be assumed. Alternative structural assumptions may lead to somewhat different attributions, but focusing on a relevant subspace has effectively driven the Shapley values for unactionable variables like sex and age toward zero. The result is that explanations are provided primarily in terms of lifestyle variables like BMI and MAP, which agents can work to minimise in order to improve their disease prognosis.

sex	age	bmi	map	tc	ldl	hdl	tch	ltg	glu
1	−0.064	0.096	0.105	−0.003	−0.005	−0.007	-0.003	0.023	0.073

Table 7.2. Complete feature vector for both Bert and Ernie in the diabetes dataset.

Consider the case of two high-risk patients – say, Bert and Ernie – who share identical feature vectors (see Table 7.2. Note that the negative value of age is a quirk of the scaling procedure.) Despite their striking similarities, however, Bert and Ernie differ in their culinary tastes. Bert was raised on carbohydrate-rich foods, and cannot imagine living without bread, pasta, and potatoes; Ernie, though he indulges in similar meals from time to time, is a voracious carnivore who eats bacon for breakfast, cold cuts for lunch, and steak for dinner. Their respective utility matrices are given by Table 7.3, where we assume for the sake of this example that only two dietary interventions are under consideration: a low-carb diet and a low-fat diet.

Bert	h_1: $\hat{y} < 265$	h_2: $\hat{y} \geq 265$
a_1: low-fat	5	−1
a_2: low-carb	1	−6

Ernie	h_1: $\hat{y} < 265$	h_2: $\hat{y} \geq 265$
a_1: low-fat	1	−6
a_2: low-carb	5	−1

Table 7.3. Utility matrices for Bert (left) and Ernie (right). Bert would like to reduce his risk of diabetes without lowering his carb intake; Ernie wants the same outcome without lowering his fat intake.

Evidence suggests that the latter is slightly more effective for weight loss, while the former has the added benefit of reducing blood pressure (Brinkworth et al., 2009). Classical Shapley values, which place the greatest weight on BMI, would tend to favour a_1 regardless of agentive preferences. However, rational Shapley values can distinguish between optimal explanations based entirely on utilities, since changes to either BMI or MAP are sufficient to bring about the desired outcome, albeit at differing costs to Bert and Ernie. Thus we find the following two feature attribution vectors (see Table 7.5), computed using the conditional reference distribution on inputs that differ only with respect to utilities u:

Bert:

sex	age	bmi	map	tc	ldl	hdl	tch	ltg	glu
−3.789	−4.867	94.922	29.241	1.763	0.849	4.053	6.566	13.019	4.324

Ernie:

sex	age	bmi	map	tc	ldl	hdl	tch	ltg	glu
−1.291	0.114	42.604	66.942	2.187	1.129	−0.167	−3.532	11.136	9.903

Table 7.4. Rational Shapley vectors for Bert and Ernie, computed from the same input vector (see Table 7.2) but with respect to different relevant subspaces.

Observe that, among the many differences between these two vectors, they flip the relative importance of BMI and MAP, suggesting two alternative paths toward reducing disease risk. A counterfactual approach with a well-designed cost function could in principle also identify different explanations for these two agents, but it could not additionally provide complete feature attributions summarising the relative impact of other predictors. Such a synthesis is unique to the rational Shapley approach.

§7.4.1.3 Recourse: Credit Scoring

As a final example, I examine the recourse setting, where the goal is to advise an agent what interventions are necessary and/or sufficient to secure a desired prediction. As Karimi et al. (2020) have shown, this task requires causal information. More specifically, they prove that noncausal recommendations, such as those provided by Ustun et al. (2019), guarantee recourse if and only if the treatment variables have no descendants in the underlying causal graph. For this experiment, I use the German credit dataset from the UCI Machine Learning Repository (Dua & Graff, 2017). This data includes 1000 samples and 20 predictors, including a combination of demographic and financial variables. To simplify the presentation, I restrict focus to just seven of the most informative features: age A, gender G, marital status M, job J, savings S, loan amount L, and duration D. The outcome Y is binary (loan approved/denied), with a base rate of 70%. As a pre-processing step, I log-transform the skewed variables A, L, D, and S to better approximate normality, and add slight Gaussian noise to the ordinal features S and J to make them more nearly continuous. Both steps simplify conditional probability estimation at no cost to model performance. I proceed to train a support vector machine (SVM) to predict Y using a radial basis kernel with default hyperparameters provided by the `e1071` package (Meyer et al., 2018).

MSV is estimated via Monte Carlo with 2000 simulations. CSV is computed via the multivariate Gaussian method for continuous variables, while the binomial features G and M are estimated via the empirical distribution. For ISV, I assume a simple partial ordering in

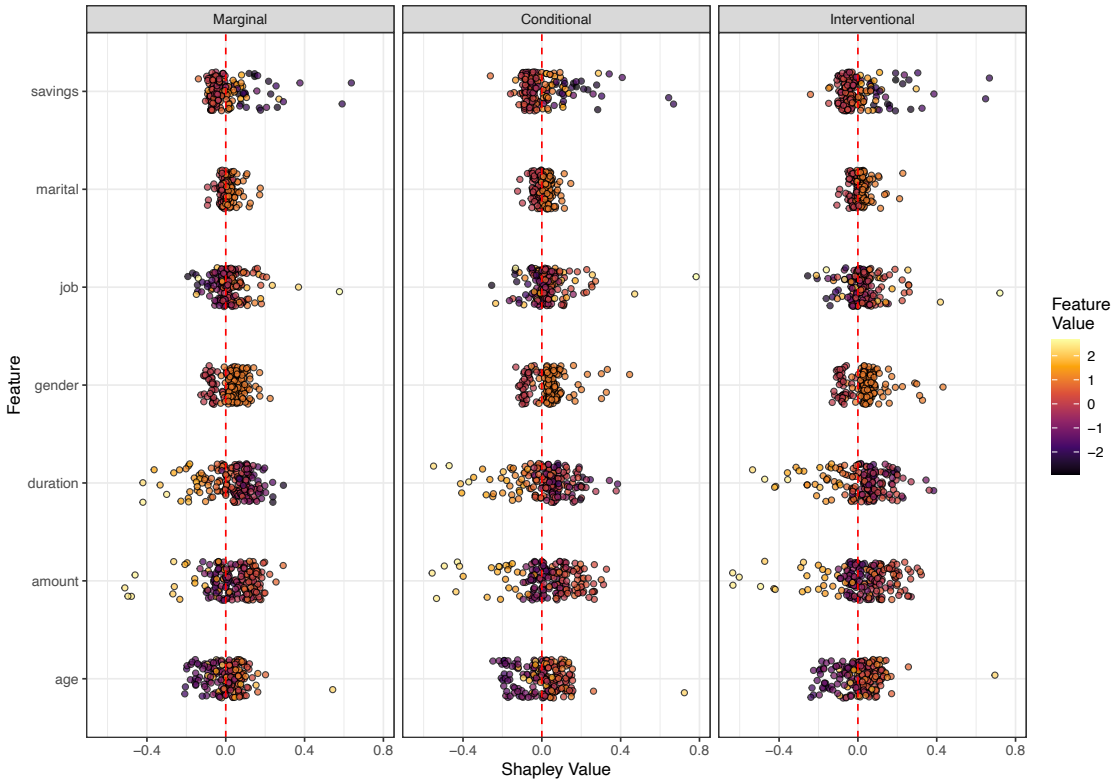

Figure 7.7. Classical Shapley values for the German credit dataset, computed with marginal, conditional, and interventional reference distributions. Continuous feature values are z-scored for visualisation.

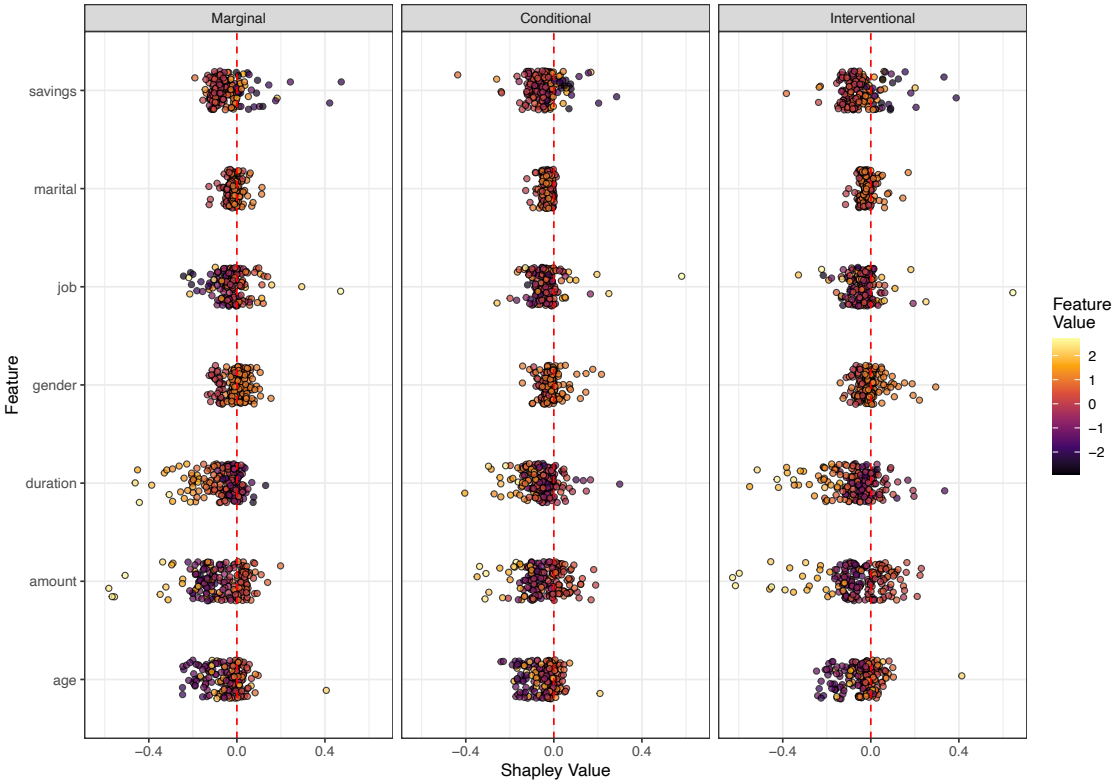

Figure 7.8. Rational Shapley values for the German credit dataset, computed with marginal, conditional, and interventional reference distributions. Continuous feature values are z-scored for visualisation.

which demographic variables are causally antecedent to both financial predictors and loan application details: $\{A, G, M\} \rightarrow \{L, D, J, S\}$.

The UCI website notes that false positives are more costly for banks than false negatives, and therefore recommends penalising the former at five times the rate of the latter. Post-processing the SVM, an optimal decision threshold is obtained at 0.74, and I therefore label all and only those points with predicted probabilities at or above this value as 1. The resulting model is about 69% accurate on the true positives and 73% accurate on the true negatives. I focus on the 189 borderline applicants with predicted success probabilities on the interval [0.7, 0.74). These are subjects who could plausibly benefit from even a slight improvement in their model predictions. Classical Shapley values are unable to help such applicants, since the mean response in the full dataset is $\phi_0 = 0.7$, i.e. the lower bound of the interval. (Irrational) feature attributions for these subjects are therefore guaranteed to sum to some value on [0, 0.04), explaining why they did slightly better than average instead of why they were ultimately unsuccessful.

The most important component of the relevant subspace is the contrastive outcome y_1, as the borderline applicants appear fairly representative along all predictors of interest. (There is, for instance, no clear overrepresentation of unmarried males or unemployed youngsters.) I therefore stratify by predicted response, zooming in on the top quartile of applicants ($\hat{y} \geq 0.76$). Bivariate analyses suggest that such applicants are for the most part financially secure and request small loans of relatively brief duration. The classical and rational Shapley vectors look fairly similar in this experiment, except for the crucial difference in offset. (See Figs. 7.8 and 7.9; note that the response variable is transformed to the logit scale for easier visualisation.)

Consider the case of a loan applicant, call her Ruth, whose predicted success probability is 0.73 – just under the decision boundary. Her input feature vector can be found in Table 7.7, where values are reported on the original (pre-transformation) scale.

gender	marital	age	amount	duration	savings	job
0	0	26	3181	26	290	2

Table 7.5. Complete feature vector for Ruth in the German credit dataset.

She is looking to buy a new home and believes her high-skilled job and decent savings (both well above their respective median values in the dataset) should make her a strong applicant. She is willing to make some changes if necessary – specifically, to increase her savings, decrease the size of her requested loan, or pay it back sooner – but in either case she would like to act quickly, since she fears the house she wants to buy will not be on the market for long. Ruth's utility matrix is given in Table 7.6. We find here that she prefers changing the terms of her application to increasing her savings in general, as this is likely the faster route to bank

approval. She is indifferent between changing the loan amount or duration. (We assume here that the amount of change in L or D is fixed; continuous extension based on distance measures could of course be devised.) As noted above, computing Shapley values from the entire dataset is pointless in this case, as Ruth's predicted outcome is already above the base rate. Counterfactuals could potentially aid her search for an explanation, but only by either providing an artificially narrow set of pathways to approval or overwhelming her with an unnecessarily large reference class of successful applicants.

	$h_1: \hat{y} \geq 0.74$	$h_2: \hat{y} < 0.74$
a_1: Increase S	1	−2
a_2: Decrease L	2	−1
a_3: Decrease D	2	−1

Table 7.6. Utility matrix for Ruth, whose demographic and financial information place her just below the loan approval threshold for a hypothetical bank.

Rational Shapley values, computed from a relevant subspace of successful loan applicants who share Ruth's demographic characteristics, provide a summary of feature attributions within this reference class (see Table 7.7. Note that calculations are performed in logit space, and that units should therefore not be interpreted as percentage points.)

gender	marital	age	amount	duration	savings	job
0	0	−0.168	0.061	−0.115	0.013	0.128

Table 7.7. Rational Shapley values for Ruth in the German credit dataset.

Interestingly, we find that her requested loan amount, though above average for the relevant subspace, did not hurt Ruth's application. On the contrary, it appears to boost her chances somewhat, a result we might not expect from surveying the counterfactuals directly. The loan duration, however – about double the average for the relevant subspace – brought her success probability down considerably. This is a highly actionable piece of information, insomuch as it guides Ruth toward algorithmic recourse that can push her application over the decision boundary. Of course, if the change is too onerous for Ruth, then she is free to resample from another subspace of successful applicants with longer loan duration on average. The procedure may continue like this indefinitely, with agents testing out new hypotheses in an iterative, exploratory fashion, updating their rewards accordingly.

§7.5 Discussion

In this section, I address two potential objections to the rational Shapley method.

§7.5.1 Scalability

The sceptical reader may plausibly object that the examples above are fairly neat and straightforward, with their small utility matrices of well-defined user preferences. Reality, of course,

is far messier. Complete sets of actions and outcomes may not be known upfront, let alone utility and credence functions defined thereon. Potential interventions may be far more numerous than these experiments permit, and associated outcomes entirely uncertain. Can rational Shapley values scale to larger, more complicated instances of algorithmic explanation?

In a word, yes. These cases are primarily illustrative, following on the back of theoretical results establishing the viability of a pragmatic synthesis between feature attributions and counterfactuals that works with various different reference distributions. The limiting factors in these experiments were complications around efficient Monte Carlo sampling and conditional probability estimation. The former can be entirely resolved with greater computational power, while the latter requires a bit more care to handle different data types and unique properties of certain joint distributions. With improvements in generative modelling – either through more thorough causal reasoning, more powerful unsupervised learning methods, or both – the rational Shapley method can easily extend to larger, more complex cases.

That said, I fully acknowledge that I have prioritised accuracy above expediency here. I presume that it is more important to get explanations right than it is to get them quickly – although both would of course be ideal. Faster, more familiar methods may be preferable in the early stages of model training, where data scientists are still experimenting and debugging. In a deployment scenario, however, where important decisions regarding criminal justice, healthcare, or personal finances hang in the balance, I believe it is worth taking the extra time to work through complex issues with user input to ensure understanding and consensus.

§7.5.2 Confirmation Bias

Another potential challenge to the proposed method is that it is vulnerable to confirmation bias. If users can scan the data in search of attribution vectors that make their preferred outcomes more likely, what's to stop them from finding such vectors even when do they do not exist? There are three answers to this charge. First, it is entirely possible that the relevant subspace be empty – that is, no observed or synthetic datapoint meets the threefold criteria specified by $\langle \mathcal{D}_S, \mathcal{D}_R, \mathcal{Y}_1 \rangle$. This may be the case, for instance, if there are no successful loan applicants with particular values for the target conditioning set. Such a failure is highly informative, as it demarcates a realm of (im)possibility for the inquiring agent. Thus it is simply false to allege that users will always find what they are looking for. Second, in cases where the relevant subspace is nonempty but sparsely populated, we should expect estimates to be unstable. Any good inference procedure should take such uncertainty into account, making it difficult for an opportunistic user's desired outcome to pass severe tests on the basis of just a small handful of outlying points. I have not said much about testing in this chapter, but it is simple in principle to extend the experiments above with frequentist or Bayesian methods

tailored to the particular task at hand. Third, there is no guarantee that agents will generally concur on proper values of $\mathcal{D}_S$, $\mathcal{D}_R$, and $\mathcal{Y}_1$. For instance, a loan applicant and bank manager may disagree as to whether one's credit score should be regarded as actionable. This makes it all the more important to be explicit about which values of each input go into any given explanation. Much like how Bayesians often defend their inferential approach by arguing that priors should be declared upfront and subject to critical scrutiny, the same could be said in this case about the components that define a relevant subspace. Such transparency will make it harder for adversarial agents to game the system and easier for those acting in good faith to come to consensus on particular cases.

§7.6 Conclusion

Local explanations for complex model predictions are sought in a wide variety of domains. The two most popular iML tools available for such explanations – feature attributions and counterfactuals – each have certain advantages, however they tend to provide different answers in particular cases that can confuse and/or mislead end users. In this chapter, I have shown how the two can be synthesised into a single method that preserves the best elements of each. Rational Shapley values preserve and extend the axiomatic guarantees of their classical forebears, giving users the flexibility to compute additive feature attributions in a fast, flexible manner. Marginal, conditional, and interventional value functions were all evaluated, and the applicability of each was compared across different use cases. By formalising the task in an expected utility framework, I was able to demonstrate how and why individual agents may rationally seek different explanations for the same model predictions. The resulting explanations are concise, intuitive, and thoroughly pragmatic.

There are several avenues for future research in this area. For instance, the value functions examined here could be extended to explicitly incorporate agentive utilities and credences so that user preferences and beliefs are used not just to weight given feature attribution vectors, but to compute them directly. I also did not explore generative methods in this chapter, however a great deal of work in counterfactual explanations is devoted to this task. It could be valuable to compare the relative merits of different generative algorithms for computing relevant subspaces, and the implications for rational Shapley values. I have also assumed that utility functions are more or less given, however it is possible to learn them gradually through user rankings when datasets are not too large. Extensions are especially complex and interesting in the case of causal systems, where unobserved confounders and/or uncertainty with respect to graph structure may enable more general solutions based on model averaging (Fragoso, Bertoli, & Louzada, 2018; Wasserman, 2000), invariant prediction sets (Bühlmann, 2020; Peters et al., 2016), or instrumental variables (Imbens, 2014; Kilbertus, Kusner, & Silva, 2020), where applicable.

Conclusion

Articulating a paradox that would later come to bear his name, Hans Moravec wrote in 1988 that "it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" (Moravec, 1988, p. 15). He argued that humans are poor judges of inherent complexity, tricked by our biological inheritance into believing that long division is hard and walking is easy. On the contrary, the apparent ease with which even small children navigate their way through a room without bumping into objects is the miraculous culmination of some billion years of evolution. Humans have not had nearly as long to perfect the art of arithmetic, which we have been practicing for at most 200 millennia. It is easy to see how natural selection might favour proficient walkers; those who struggle with long division, by contrast, seem to reproduce at prolific rates.

Moravec's paradox gets at the heart of what is philosophically rich and stimulating about computer science – not (just) its ability to *solve* problems per se, but its tendency to *reveal* their true nature. The vast majority of contemporary work on AI, both in academic and popular literature, is concerned with the former capacity. Thus, we may augment Moravec's checkers example with recent achievements in chess, shogi, and Go (Silver et al., 2018), to say nothing of advances in computer vision (Xie et al., 2019), natural language processing (Liu et al., 2019), and unstructured game play (Vinyals et al., 2019). Meanwhile, critics in the FAT ML community tend to highlight the failures of machine learning, especially as applied to high-stakes decisions in finance (Eubanks, 2018), education (O'Neil, 2016), and criminal justice (Berk et al., 2018). In all these cases, the emphasis is on the *performance* of the underlying algorithms – the extent to which they solve or exacerbate some problem, or potentially create one where there was none (Morozov, 2013). Yet perhaps the true lesson is not that AI can or cannot accomplish certain objectives to within a tolerable margin of error. Moravec's paradox reminds us that even our computational failures can clarify the structure and complexity of selected tasks, thereby forcing us to confront our own assumptions and biases – biological, psychological, and sociological.

The recent "boom" phase of the AI hype cycle, which arguably began in 2012 with the first successful deep neural network for image recognition (Krizhevsky, Sutskever, & Hinton, 2012), has made ML algorithms more ubiquitous than ever. In this thesis, I have focused on just one aspect of this emerging technology, namely *explainability*. Humans have strong (occasionally conflicting) intuitions about how explanations ought to work. However, just like the roboticist who struggles to program a machine to walk around a room, we run into a number of vexing obstacles as soon as we attempt to automate the process of explaining models and predictions. Some of these challenges are conceptual in nature; others are technical. I have critically engaged with both sorts of issues at length. I do not claim that the preceding chapters

have solved all the problems of iML – far from it, I believe that some of the most interesting ones will keep researchers busy for years to come – yet I would argue that "solutions" are not the only worthwhile terminus of intellectual discourse. More valuable than any formal or empirical result has been the rediscovery of certain fundamental truths about the nature of explanation, truths that are thrown into stark relief by the attempt to design iML tools.

I will review the contributions (both negative and positive) of this dissertation more fully in §8.1. For now, I want to call attention to three main themes that emerged from the preceding chapters:

(1) *Explanations are fundamentally causal*. The questions iML attempts to answer are all about how outcomes would differ under various sorts of interventions. However, there is an important and subtle ambiguity as to the target level of abstraction. Explanations may be causal with respect to the model but not the data generating process, and vice versa.

(2) *Explanations are fundamentally pragmatic*. They are the response to questions posed by agents with certain beliefs and interests. The same explanandum may warrant different explanantia depending on who is asking and why. An individual may seek several overlapping explanations emphasising different aspects of a given phenomenon before her curiosity is satisfied.

(3) *Explanations must be severely tested*. They are essentially a sort of scientific hypothesis, and as such should be subject to the same standards of rigour we have come to expect from published works of physics or biology. That means that error rates must be reported, uncertainty quantified, and testing procedures open to scrutiny.

I am not the first to stumble upon these points. Indeed, all three have been defended in different contexts by numerous authors who predate not just the current AI hype cycle, but digital computation altogether (see Chapter 2). What they may lack in novelty, however, they make up for in utility. For within them we find more or less everything we need to diagnose what ails iML and point the way toward better alternatives. Moreover, they exemplify a rare and fruitful sort of cross-pollination between philosophy and computer science. Just as careful attention to theoretical foundations can impose rigour and improve practice, practice itself can sharpen intuitions and inform theory. Putting this platitude to work is no small task. If I have done nothing else in this thesis, I hope to have demonstrated a convincing model of what interdisciplinary research can look like in this area.

The remainder of this chapter is structured as follows. In §8.1, I summarise the results of previous chapters, demonstrating how the heterogeneous components that comprise this dissertation cohere into an integrated thesis orbiting around this trio of themes. In §8.2, I briefly review the key takeaway lessons for readers from various audiences. I outline directions for future work in §8.3. Though this project may have advanced the state of the art in several

respects, it has hardly exhausted the conceptual or technical challenges of iML. Finally, I close with some concluding remarks in §8.4.

§8.1 Review

In Chapter 1, I set out to answer two research questions. To wit:

> RQ1. What constitutes a satisfactory explanation of a supervised learning model or prediction?
>
> RQ2. Can reliable methods be developed for generating model-agnostic algorithmic explanations?

RQ1 is essentially conceptual. It demands a theory of explanation, an awareness of the distinction between (global) models and (local) predictions, as well as an appreciation for the affordances and constraints that are characteristic of supervised learning algorithms. RQ2, on the other hand, is a bare bones prompt – almost a dare. It suggests a certain scepticism about existing iML tools, which are presumably unreliable in some respect(s). It builds naturally upon RQ1, leveraging the momentum established in Parts I and II to design and implement iML tools that users can apply to real-world problems. This was the goal of Part III, in which I developed methods for explaining the behaviour of arbitrary supervised learners at global and local resolutions.

In Chapter 2, I sought to map the landscape of the current discourse on iML. Critically examining literature from CDS scholars and philosophers, I grounded contemporary debates on explainability (and the closely related issue of algorithmic fairness) in a longstanding dialectic between technological dogmatism and scepticism. In the former camp, we found some potentially awkward intellectual bedfellows, including Marxist determinists, Weberian rationalists, and Hayekian capitalists. In the latter, we found a backlash among social constructivists and critical theorists who reject the assumptions and logic undergirding the modern sociotechnical order. I judged neither pole of the dialectic to be especially satisfying, advocating instead a pragmatic synthesis that rejects both the reductionism of the dogmatists and the puritanism of the sceptics. I analysed a range of theoretical approaches to explanation, embracing a broadly interventionist account that emphasises the role of trust and testing.

In Chapter 3, I surveyed a number of computational and statistical approaches to iML, including some of the most popular methods for generating feature attributions, rule lists, and case-based explanations. I traced the historical development of the subdiscipline within computer science, starting with Ribeiro et al.'s (2016) LIME algorithm, through to more recent optimality results for certain tree-based methods and ongoing debates regarding sampling procedures for counterfactuals. With an emphasis on post-hoc, model-agnostic approaches, this chapter provided key technical background that rounded out the literature review section and set the stage for the critical and constructive arguments that followed.

In Chapter 4, I focused on three conceptual challenges for iML. First, I pointed out a crucial ambiguity in feature attribution methods that struggle to differentiate between model- and system-level analyses. By being explicit about the intended target of analysis – be it the algorithm itself or the underlying data generating process – we can select the proper causal framework and derive more useful, relevant explanations. Second, I emphasised the central role of severe testing in science, and argued that existing iML methods do not devote sufficient attention to quantifying uncertainty or bounding error rates. A rigorous inferential framework for model explanations, be it in the frequentist or Bayesian tradition, is required to protect against spurious outputs and ensure user trust. Third, I found that current iML tools overwhelmingly value product over process. They treat explanations as objective deliverables, computed once and for all, instead of as dynamic synergies that emerge organically through the interactions of multiple agents, human or otherwise. I argued that all three points are underappreciated or absent in the current iML discourse, and that failure to directly confront these challenges will only serve to stymie progress.

In Chapter 5, I introduced a novel framework for analysing iML proposals. The explanation game is an iterative procedure of questions and answers between (at least) two agents, working in collaboration to find the best explanations of some given model prediction(s). Outputs are scored on three axes – accuracy, simplicity, and relevance – which together define a solution space. Optimality is defined in terms of a Pareto frontier, such that no explanation along the frontier can improve by any measure without thereby degrading on at least one other measure. Agents are free to pick their favourite explanation(s) from among those on the Pareto frontier. I described convergence conditions for the game, which reaches a conditionally optimal equilibrium in polynomial time under reasonable assumptions. The explanation game serves a descriptive and a normative function, creating a common vocabulary with which to compare different iML tools and pointing the way toward novel solutions that improve upon the state of the art.

Chapters 2 and 3 together form Part I of this thesis, an interdisciplinary literature review that situates this project at the intersection of CDS, philosophy, and computer science. I labelled this part *Theoria* to emphasise its preparatory role, the thought that must precede any meaningful action. Chapters 4 and 5 constitute Part II: *Praxis*, where thought turns into motion, as I subject contemporary iML methods to epistemological critique. In Part III: *Poiesis*, I design new tools that distil the lessons of both theory and practice – naturally progressing from thinking and doing to *making*.

I began in Chapter 6 with a novel test of conditional independence, which doubles as a global measure of variable (subset) importance. This is much harder to test than marginal importance, which can be directly computed using parametric (e.g., partial correlation) or nonparametric (e.g., permutation) techniques. Building on the notion of knockoff variables,

which share the covariance structure of the original data but are independent of the response, I implemented an efficient procedure for assessing the conditional importance of any feature subset without resampling or model refitting. The conditional predictive impact (CPI) can be computed with any choice of supervised learner, loss function, and knockoff sampler, inheriting its consistency properties from those inputs. I showed that the CPI controls Type I error at the nominal level, and studied Type II error rates as a function of the data generating process (covariance structure, non/linearity of response, number of features and sample size) as well as learning algorithm and loss function. The test achieves high power throughout compared to a range of alternative baselines.

In Chapter 7, I considered the problem of local explanation. I used counterfactual methods to compute baseline expectations and Shapley values to estimate feature attributions, thereby combining the two main strands of contemporary iML research, which are generally considered orthogonal to each other. Furthermore, I ensured the relevance of resulting explanations by adopting a Bayesian decision theoretic framework in which agentive beliefs and interests are explicitly encoded via credence and utility functions, thereby enabling precise expected utility maximisation. I demonstrated how this allows users the flexibility to customise explanations to their specific needs, generating different outputs depending on who is asking and why. The method works with any valid procedure for computing counterfactuals and Shapley values – the choice of how much causal information to include depends upon an agent's goals – as well as any credence and utility functions that satisfy a small number of well-known axioms.

These chapters may differ markedly in their style and target audience, but they all share a common emphasis on causality, pragmatics, and severe testing. The importance of all three themes is philosophically motivated – I draw especially on Woodward (2003)'s interventionism, Floridi (2011)'s levelism, and Mayo (2018)'s notion of severity – yet each point has a formal analogue that aids in implementation. Specifically, I use Pearl (2000)'s notion of structural causal models to reason about causal systems, Bayesian decision theory (Savage, 1972) to express subjective beliefs and preferences, and frequentist inference (Lehmann & Romano, 2005) to quantify errors of the first and second kind. Thus the aforementioned themes form the theoretical and practical core of this integrated thesis. Their insights have helped better conceptualise the task of iML, identify shortcomings in existing methods, and design new solutions with more solid foundations.

§8.2 Lessons

I envision five main audiences for this work: CDS scholars, philosophers, data scientists, policymakers, and end users. Each group is considered in turn.

CDS scholars. CDS scholarship was a primary focus of Chapter 2, and constitutes one of the three theoretical pillars upon which this thesis stands. However, I suspect few self-professed practitioners will be especially pleased with my analysis in §2.1. Despite my profound respect for the foundational work of several key luminaries in this field – especially the legal and policy wing, spearheaded by Solon Barocas, Andrew Selbst, and Sandra Wachter, among others – I fear that the contemporary literature is becoming increasingly monolithic in its reflexive disdain for technology. Conflating tools with structures and eager to oversimplify complex issues for popular audiences, a raft of authors has begun to drift into decidedly unnuanced territory, dismissing the very notion of quantification as a misguided exercise in neoliberal oppression. I have argued that such reductionism is uncritical and self-defeating. The main lesson for this group is to rigorously benchmark proposed technologies against their analogue counterparts. The potential dangers of automation are real, especially in socially sensitive settings – however, that does not mean we should not consider the merits on a case-by-case basis. This is an explicitly pragmatic undertaking, relying on context and judgment over maxims and dogmas.

Philosophers. I surveyed a number of philosophical approaches to explanation in §2.2, and drew on several contemporary debates in philosophy of science and statistics in subsequent chapters. Beyond my aforementioned endorsement of particular theories and frameworks, I want to emphasise once again how applied AI research can inform and advance philosophical discourse. Just as philosophers with backgrounds in mathematics (e.g., Russell) and physics (e.g., Kuhn) were uniquely positioned to tackle the conceptual challenges associated with those areas in the twentieth century, I believe that greater experience in ML methods will provide twenty-first century philosophers with new perspectives on important issues in formal epistemology, inductive inference, probability theory, and other areas of mutual interest between the disciplines. This appreciation for practice goes hand-in-hand with the pragmatic approach I have defended in this thesis. Note, however, that the traffic runs both ways. Conceptual analysis can clarify issues in AI – as we have seen in the trio of core themes above – just as applied work in AI can inform philosophical debates. For instance, Vapnik (1995; 1998)'s foundational results in statistical learning theory led to refinements in falsificationism (Corfield et al., 2009) and reliabilism (Harman & Kulkarni, 2007; 2011), just as graphical theories of causality (Pearl, 2000; Spirtes et al., 2000) directly inspired modern interventionism (Woodward, 2003). The lesson, above all, is to keep channels of communication open, to stay intellectually curious and methodologically humble. Insight comes in many forms.

Data scientists. Studies indicate that data scientists are by far the most likely group to actually interact with iML tools, which they use to troubleshoot models during training and monitor performance after deployment (Bhatt et al., 2020). Judging by the adoption patterns

of current methods – SHAP and LIME are especially popular, while empirical counterfactuals are also widespread – the primary goal for this group appears to be computational efficiency, with axiomatic guarantees perhaps serving as an extra theoretical incentive. This makes some sense given the practical needs of professional data scientists, who must wrangle large datasets on a daily basis, building and debugging potentially massive systems with high frequency. Such a workflow places a natural premium on speed. Shortcuts are especially welcome when they come with the added reassurance of "axiomatic guarantees". However reasonable this may be during the model building phase, the logic does not extend well to deployment. First of all, the purported guarantees of, say, Shapley values, are only approximately satisfied by fast programs like SHAP, which help themselves to strong and often unreasonable assumptions such as the joint independence of all predictors. Relying on observational data for counterfactuals, which effectively amounts to matching, is only advisable with large and representative datasets. Even then, the method can fail when the sample in question comes from a low-density region of the feature space, precisely where predictions are most likely to err. More importantly, data scientists must recognise that the role of iML evolves as we shift from model training (during which it is probably used for validation) to model testing and deployment (during which it is more likely used to audit or discover). In the latter case, the stakes may be very high indeed. It is therefore essential that we get such explanations right, even if that means using computational methods that are more expensive than plug-and-play solutions like LIME and SHAP.

Policymakers. Though I have not devoted much time to the topic in this thesis, there are clear policy implications of the present research. Local explanation in particular has become a topic of intense interest in Europe following the passage of the 2018 GDPR and the ensuing debates surrounding the so-called "right to explanation". As several authors have argued (Barocas et al., 2020; Wachter et al., 2018), there are major problems with mandating current iML tools as rough and ready legal standards. This project has hopefully helped to clarify how and why that is the case for several popular proposals, and ameliorated those concerns with novel solutions better fit to purpose. Large scale implementation would likely require a distinct regulatory body tasked with auditing algorithms under nondisclosure agreements, as originally envisaged by Wachter, Mittelstadt, & Floridi (2017). The logistics of such a proposal are beyond the scope of this project, however it is worth noting that better explanations – in terms of accuracy, simplicity, and relevance – are almost surely attainable with greater access to the underlying models and data. The post-hoc, model-agnostic methods analysed in this project are advanced under the assumption that powerful forces will succeed in preventing lawmakers from gaining unfettered access to their intellectual property. If, on the contrary, regulators have broad powers to scrutinise code and data, then more efficient alternatives may be pursued on a case-by-case basis. This runs the risk of multiplying accept-

able standards – at least one for each popular class of algorithms – but those challenges would likely be preferable to the imprecision inherent to post-hoc, model-agnostic explanations.

End users. Ultimately, the goal of all this work has been to empower end users to interrogate and better understand the behaviour of arbitrary black box models. This is a difficult task, not least because "end users" is a vague and arguably unhelpful label, denoting a heterogeneous group that can safely be presumed to vary widely in their background knowledge and motivations. Evidence suggests that despite much rhetoric to the contrary, end users rarely engage directly with iML tools (Bhatt et al., 2020). This may be for several reasons – graphical user interfaces (GUIs) are not especially well developed for most iML software, although some tech companies are attempting to change this – but it certainly does not help that very few options currently on the market explicitly allow users to customise solutions. The lesson for this group (if indeed such a haphazard assemblage could be so called) is that *their needs matter*. If someone cannot understand an explanation provided by an iML tool, then our first assumption should be that the fault lies with the software itself, not the individual in question. If Alice cannot put an explanation to use – if it fails to advance her comprehension of the explanandum or aid her in predicting outputs on similar points – then something has gone awry. The solution is not to blame her for some perceived lack of mathematical nous, but to build better tools that will adapt to her inquiry. If two agents inquire about the same prediction with different priorities in mind – say, one seeks to validate the algorithm, which she doubts works correctly, while the other assumes it is accurate and wants to learn about the underlying system – then the explanations delivered to both should reflect that difference. Any iML tool that fails to distinguish between these cases does not take its end users seriously. It implicitly assumes they are a monolithic group with no variation in their goals or background knowledge. This thesis has argued that end users deserve better, and demonstrated how computational methods can achieve this.

§8.3 Future Directions

One of the biggest challenges in writing this thesis was avoiding the temptation to pursue various possible extensions of this work that naturally emerged throughout the course of my research. In an effort to bound the scope of this dissertation and stick (more or less) to my original outline, I filed away several ideas for future projects that would probably have doubled or tripled the length of this manuscript (and the duration of my DPhil) had I actually seen them through. While I have every intention of picking these topics up in future, I am grateful to my advisors for cautioning me against tackling them all herein. I note several below that I think hold particular promise.

Adversarial explanation games. A central assumption of the explanation game framework laid out in Chapter 5 is that Alice and Bob are on the same side. They cooperate in a

collaborative effort to establish the best possible explanation. But what if, instead of accomplices, they were adversaries? Under what conditions could Bob systematically deceive Alice into believing false explanations, or explanations that deliberately thwart her interests to Bob's advantage? How, if at all, could Alice protect herself against such efforts? Several authors have recently described how feature attribution methods like LIME and SHAP can be fooled into explaining discriminatory predictions as unbiased (Lakkaraju & Bastani, 2020; Slack et al., 2020). Their analyses have been primarily empirical, however. By contrast, I am interested in studying the theoretical properties of such scenarios. Specifically, I intend to characterise the features of data generating processes that allow for such manipulations, which are invariably the result of high mutual information between variables. I conjecture that an impossibility result may be proved, to the effect that, with a sufficiently large advantage in computational resources, Bob can always fool Alice into believing a suboptimal explanation.

Local CPI. The local explanation method described in Chapter 7 represents a significant departure from the global approach taken in Chapter 6. However, there are several possible ways to extend the knockoff-based CPI to local estimates. The conceptual obstacle to overcome here is that there is no clear sampling distribution associated with the scalar value Δ_i, i.e. the difference in error between null and alternative models evaluated at a single point i. We may nevertheless estimate the uncertainty of this parameter via several brute force methods, such as repeated refitting on data bootstraps or subsamples, potentially in combination with repeated knockoff samples, as in the conditional randomisation test described by Candès et al. (2018). These methods would be computationally expensive. Faster alternatives include adding kernel weights to the original CPI test, such that local estimates are a weighted average with factors proportional to the distance between points; or covariate-adjusted approaches in the spirit of false discovery rate regression (Korthauer et al., 2019). An empirical evaluation of these methods could demonstrate the trade-offs associated with each.

Fairness. Fairness has played a supporting role in this thesis, occasionally serving to motivate and focus the discussion yet rarely receiving much explicit attention, aside from a relatively brief subsection of Chapter 2 (§2.1.3). The relationship between algorithmic bias and interpretability is more complicated than a simple subset relation, although I believe that iML tools are an essential part of actionable auditing schemes, as argued above. A somewhat separate question, however, is how to learn fair representations in the first place. This ties into a number of related issues regarding ML approaches for causal inference, where it is often essential to learn unbiased estimators in the face of confounding signals (e.g., from non-randomised designs). One direction for future work, inspired by recent results in invariant causal prediction (Arjovsky et al., 2019; Bühlmann, 2020; Peters, Bühlmann, & Meinshausen, 2016), would be to directly optimise for a minimax loss function such that the maximal risk for any given value of some protected variable is minimised. This would implement a fairness

criterion inspired by Rawls's theory of justice, which holds that a society is "just when the prospects of the least fortunate are as great as they can be" (Rawls, 1971, p. 328). I believe the proposed learning procedure can be efficiently computed in a greedy fashion with regression trees, although I have yet to work out the details.

Abstraction. In attempting to implement pragmatic solutions for iML, I have focused primarily on beliefs and preferences over given action-outcome pairs. However, agents may seek to *learn* about outcomes incrementally in a data-driven manner over the course of investigation. Causal feature learning, in which low-level input data are coarsened into high-level summaries, has been studied in a series of papers by Chalupka et al. (2015; 2016; 2017) and formalised more generally by Beckers et al. (2019a, 2019b). I have argued elsewhere that learning causal features requires pragmatic information (Kinney & Watson, 2020), but this analysis could be extended to more general mapping methods through dimensionality reduction techniques, in the vein of so-called "disentangled representations", which are popular in the deep learning community (Locatello et al., 2019). A fully general solution – one that reflects both pragmatic and causal information, without the limits imposed by discretisation – remains elusive. An important future direction for iML is to learn these mappings in an efficient and flexible manner.

Unsupervised learning. I limited myself in this project to supervised learning algorithms, the typical focus of iML research. I observed in a footnote in Chapter 1 that *unsupervised* learning raises a number of unique philosophical questions that go beyond the remit of this thesis. Whereas the philosophical content of supervised learning is primarily *epistemological* in character – pertaining to issues of induction, inference, and uncertainty – the philosophical content of unsupervised learning is more *metaphysical*. This aspect of computational statistics has received comparatively little attention among analytic philosophers, though it is every bit as fundamental. Clustering algorithms, for instance, effectively implement some theory of natural kinds, using data-driven techniques to try and carve nature at the joints. Embedding methods such as autoencoders could be interpreted as a computational attempt to learn essences, like reconstructing the fire of Plato's cave using only data from the shadows on the wall. Generative models, meanwhile, learn contingencies, sampling shadows using only data from the fire. Studying these analogies in more depth and probing their lessons for philosophy and ML alike constitutes one of the most exciting future directions for my research.

§8.4 Concluding Remarks

I opened this dissertation with the twin cases of AlphaGo and COMPAS, two algorithms that made headlines in 2016 exemplifying the promise and challenge of ML in our contemporary information society. I began this chapter with reference to Moravec's paradox, which reminds

us that AI is not just a toy or a tool, but above all a mirror on our own cognition. These examples all serve to highlight the irreducibly complex and multifaceted nature of contemporary research in computational statistics, which inevitably touches on issues of sociopolitical significance and philosophical import.

It is worth remembering that the field of iML – and, indeed, of statistical learning more generally – is still in its infancy. Lessons from sociology and public policy will be necessary to steer this exciting and powerful technology in the right direction. Philosophical insight will be required to conceptualise the ethical, epistemological, and even metaphysical assumptions and implications of these models. I have argued for an inclusive, rigorous approach to these challenges, one that embraces interdisciplinary methodologies, combining theory and practice in a single intellectual framework. The horizon for such research is broad and bright. This doctoral project may have traced the landscape, but it remains merely a sketch – detailed in some parts but loose in others, perhaps a study for a larger canvas still years off yet. The next step is to fill in the outlines in all their vibrant colours, to add perspective and layer textures. I look forward to taking up the brush.

Bibliography

Aas, K., Jullum, M., & Løland, A. (2019). Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv* preprint, 1903.10464.

Achinstein, P. (1983). *The Nature of Explanation*. New York: Oxford University Press.

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

Al-Ghazali. (2000). *The Incoherence of the Philosophers* (M. E. Marmura, Trans.). Provo, UT: Brigham Young University Press.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347.

Ananny, M., & Crawford, K. (2016). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media Soc.*, *20*(3), 973–989.

Anderson, C. (June, 2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*. Retrieved from https://www.wired.com/2008/06/pb-theory/.

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning certifiably optimal rule lists for categorical data. *J. Mach. Learn. Res.*, *18*(234), 1–78.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. Technical report, *ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Anselm. (2002). *Anselm: Three Philosophical Dialogues* (T. Williams, Ed. & Trans.). Indianapolis: Hackett.

Aristotle. (1984). *The Complete Works of Aristotle* (J. Barnes, Ed.). Princeton: Princeton University Press.

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

Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations: A survey. *arXiv* preprint, 1911.07749.

Austin, J. L. (1961). *Philosophical Papers* (J. O. Urmson & G. J. Warnock, Eds.). Oxford: Clarendon Press.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K. R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, *10*(7), 1–46.

Baker, A. (2016). Simplicity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Bandyopadhyay, P. S., & Forster, M. R. (Eds.). (2011). *Philosophy of Statistics*. Oxford:

Elsevier.

Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.*, *43*(5), 2055–2085.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning*. Retrieved from fairmlbook.org.

Barocas, S., & Selbst, A. (2016). Big data's disparate impact. *Calif. L. Rev.*, *104*(1), 671–729.

Barocas, S., Selbst, A. D., & Raghavan, M. (2020). The hidden assumptions behind counterfactual explanations and principal reasons. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 80–89. New York: ACM Press.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion*, *58*, 82–115.

Bates, S., Candès, E., Janson, L., & Wang, W. (2020). Metropolized knockoff sampling. *J. Am. Stat. Assoc.*, 1–15.

Beer, D. (2017). The social power of algorithms. *Inform. Commun. Soc.*, *20*(1), 1–13.

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix prize challenge. *SIGKDD Explor. Newsl.*, *9*(2), 75–79.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate : A practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*, *57*(1), 289–300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, *29*(4), 1165–1188.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Anal.*, *1*(3), 385–402.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.*, 0049124118782533.

Berkeley, G. (1979). *Three Dialogues between Hylas and Philonous* (R. M. Adams, Ed.). Indianapolis: Hackett.

Berrett, T. B., Wang, Y., Barber, R. F., & Samworth, R. J. (2019). The conditional permutation test for independence while controlling for confounders. *J. Roy. Stat. Soc. B.*, *82*(1), 175–197.

Bett, R. (2012). *Sextus Empiricus: Against the Physicists*. Cambridge: Cambridge University Press.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., … Eckersley, P. (2020). Explainable machine learning in deployment. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657. New York: ACM Press.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Bien, J., & Tibshirani, R. (2011). Prototype selection for interpretable classification. *Ann. Appl.*

Stat., *5*(4), 2403–2424.

Bijker, W. E., Hughes, T. P., & Pinch, T. (Eds.). (1987). *The Social Construction of Technology Systems: New Directions in the Sociology and History of Technology*. Cambridge, MA: The MIT Press.

Bimber, B. (1990). Karl Marx and the three faces of technological determinism. *Soc. Stud. Sci.*, *20*(2), 333–351.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., … Jones, Z. M. (2016). mlr: Machine Learning in R. *J. Mach. Learn. Res.*, *17*(170), 1–5.

Blaauw, M. (Ed.). (2013). *Contrastivism in Philosophy*. New York: Routledge.

Bloor, D. (1976). *Knowledge and Social Imagery*. Chicago: University of Chicago Press.

Boca, P. P., Bowen, J. P., & Siddiqi, J. I. (2010). *Formal Methods: State of the Art and New Directions*. London: Springer.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. New York: Oxford University Press.

boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Inform. Commun. Soc.*, *15*(5), 662–679.

Breiman, L. (2001a). Random forests. *Mach. Learn.*, *45*(1), 1–33.

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statist. Sci.*, *16*(3), 199–231.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Boca Raton, FL: Taylor & Francis.

Briggs, R. (2012). Interventionist counterfactuals. *Philos. Stud.*, *160*(1), 139–166.

Brinkworth, G. D., Noakes, M., Buckley, J. D., Keogh, J. B., & Clifton, P. M. (2009). Long-term effects of a very-low-carbohydrate weight loss diet compared with an isocaloric low-fat diet after 12 mo. *Am. J. Clin. Nutr.*, *90*(1), 23–32.

Bromberger, S. (1966). Why Questions. In R. Colodny (Ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy*. Pittsburgh: University of Pittsburgh Press.

Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World*. Cambridge, MA: The MIT Press.

Bühlmann, P. (2020). Invariance, causality and robustness. *Statist. Sci.*, *35*(3), 404–426.

Bühlmann, P., Drineas, P., Kane, M., & van der Laan, M. (Eds.). (2016). *Handbook of Big Data*. Boca Raton, FL: Chapman and Hall/CRC.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Appl. Comput. Inform.*, *15*(1), 27–33.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of the Conference on Fairness, Accountability and Transparency*, 77–91. New York: ACM Press.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning

algorithms. *Big Data Soc.*, *3*(1), 1–12.

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *J. Roy. Stat. Soc. B*, *80*(3), 551–577.

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., & Smith, A. (2019). From soft classifiers to hard decisions: How fair can we be? *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 309–318. New York: ACM Press.

Carnap, R. (1950). *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carnap, R. (1952). *The Continuum of Inductive Methods*. Chicago: University of Chicago Press.

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *Br. J. Philos. Sci.*, *53*(3), 411–453.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge: Cambridge University Press.

Caruana, R. (1997). Multitask learning. *Mach. Learn.*, *28*(1), 41–75.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthCare. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1721–1730. New York: ACM Press.

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480.

Chalupka, K., Bischoff, T., Perona, P., & Eberhardt, F. (2016). Unsupervised discovery of El Niño using causal feature learning on microlevel climate data. *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, 72–81. Arlington, VA: AUAI Press.

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: an overview. *Behaviormetrika*, *44*(1), 137–164.

Chalupka, K., Perona, P., & Eberhardt, F. (2015). Visual causal feature learning. *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., & Su, J. K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems 32*, 8930–8941.

Chen, C., & Rudin, C. (2018). An optimization approach to learning falling rule lists. *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 604–612. Playa Blanca, Lanzarote, Canary Islands: PMLR.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. New York: ACM Press.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163.

Chouldechova, A., & Roth, A. (2018). The Frontiers of Fairness in Machine Learning. *arXiv* preprint, 1810.08810.

Clough, E., & Barrett, T. (2016). The gene expression omnibus database. *Methods Mol. Biol.*, *1418*, 93–110.

Corbett-Davies, S., Goel, S., Morgenstern, J., & Cummings, R. (2018). Defining and designing fair algorithms. *Proceedings of the 2018 ACM Conference on Economics and Computation*, 705. New York: ACM Press.

Corfield, D., Schölkopf, B., & Vapnik, V. (2009). Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *J. Gen. Philos. Sci.*, *40*(1), 51–58.

Correa, J., & Bareinboim, E. (2020). A calculus for stochastic interventions: Causal effect identification and surrogate experiments. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. New York: AAAI Press.

Datta, Amit, Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings. *Proc. Priv. Enh. Technol.*, (1), 92–112.

Datta, Anupam, Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy non-discrimination in data-driven systems. *arXiv* preprint, 1707.08120.

Datta, Anupam, Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *Proceedings of the 2016 IEEE Symposium on Security and Privacy*, 598–617.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

Dawid, A P. (2000). Causal inference without counterfactuals. *J. Am. Stat. Assoc.*, *95*(450), 407–424.

Dawid, A Philip. (2015). Statistical causality from a decision-theoretic perspective. *Annu. Rev. Stat. Appl.*, *2*(1), 273–303.

Dawid, A Philip. (2020). Decision-theoretic foundations for statistical causality. *arXiv* preprint, 2004.12493.

Dawid, P. (2012). The decision-theoretic approach to causal inference. In C. Berzuini, P. Dawid, & L. Bernardinelli (Eds.), *Causality: Statistical Perspectives and Applications* (pp. 25–42). London: Wiley.

de Regt, H. W., Leonelli, S., & Eigner, K. (Eds.). (2009). *Scientific Understanding: Philosophical Perspectives*. Pittsburgh: University of Pittsburgh Press.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.*, *210*, 2–21.

Dennis, M. J. (2018). Artificial intelligence and recruitment, admission, progression, and retention. *Enroll. Manag. Rep.*, *22*(9), 1–3.

Dewey, J. (1999). *The Essential Dewey* (L. Hickman & T. Alexander, Eds.). Bloomington, IN: Indiana Univesity Press.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe. Retreived from https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., & Pontil, M. (2018). Empirical risk minimization under fairness constraints. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2796–2806.

Doran, G., Muandet, K., Zhang, K., & Schölkopf, B. (2014). A Permutation-based kernel conditional independence test. *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 132–141. Arlington, VA: AUAI Press.

Doshi-Velez, F. (2017). A roadmap for the rigorous science of interpretability. *Talks at Google*. Retreived from https://www.youtube.com/watch?v=MMxZlr_L6YE.

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

Doshi-Velez, F., & Kortz, M. (2017). Accountability of AI under the law: The role of explanation. Technical report, *Berkman Klein Center for Internet & Society*. Retreived from https://dash.harvard.edu/handle/1/34372584.

Dowe, P. (2000). *Physical Causation*. Cambridge: Cambridge University Press.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.*, *4*(1), eaao5580.

Dua, D., & Graff, C. (2017). *UCI Machine Learning Repository*. Irvine, CA: University of California, School of Information and Computer Science.

Duhem, P. (1954). *The Aim and Structure of Physical Theory* (P. W. Wiener, Ed.). Princeton, NJ: Princeton University Press.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: The MIT Press.

Eberhardt, F. (2010). Causal Discovery as a Game. *Proceedings of the NIPS Workshop on Causality*, 87–96. Whistler, Canada: PMLR.

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a "right to explanation" is probably not the remedy you are looking for. *Duke Law Technol. Rev.*, *16*(1), 18–84.

Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. New York: Cambridge University Press.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, *32*(2), 407–499.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Feng, J., Williamson, B., Simon, N., & Carone, M. (2018). Nonparametric variable importance using an augmented neural network with multi-task learning. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (pp. 1496–1505). Stockholmsmässan, Stockholm Sweden: PMLR.

Feyerabend, P. (1975). *Against Method*. London: New Left Books.

Fine, K. (2012). Counterfactuals without possible worlds. *J. Philos.*, *109*(3), 221–246.

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

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). *The Design of Experiments*. London: Oliver & Boyd.

Floridi, L. (2002). *Sextus Empiricus: The Transmission and Recovery of Pyrrhonism*. New York: Oxford University Press.

Floridi, L. (2004). On the logical unsolvability of the Gettier problem. *Synthese*, *142*(1), 61–79.

Floridi, L. (2006). The logic of being informed. *Logique & Analyse*, *49*(196), 433–460.

Floridi, L. (2008a). The method of levels of abstraction. *Minds Mach.*, *18*(3).

Floridi, L. (2008b). Understanding epistemic relevance. *Erkenntnis*, *69*(1), 69–92.

Floridi, L. (2010). Information, possible worlds and the cooptation of scepticism. *Synthese*, *175*, 63–88.

Floridi, L. (2011a). Semantic information and the correctness theory of truth. *Erkenntnis*, *74*(2), 147–175.

Floridi, L. (2011b). *The Philosophy of Information*. Oxford: Oxford University Press.

Floridi, L. (2012). Semantic information and the network theory of account. *Synthese*, *184*(3), 431–454.

Floridi, L. (2013). *The Ethics of Information*. Oxford: Oxford University Press.

Floridi, L. (2014). Open data, data protection, and group privacy. *Philos. Technol.*, *27*(1), 1–3.

Floridi, L. (2017). The logic of design as a conceptual logic of information. *Minds Mach.*, *27*(3), 495–519.

Floridi, L. (2019). *The Logic of Information*. Oxford: Oxford University Press.

Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Sci. Rev. 1*(1)

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.*, *28*(4), 689–707.

Forgy, E. (1965). Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, *21*(3), 768–769.

Fragoso, T. M., Bertoli, W., & Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *Int. Stat. Rev.*, *86*(1), 1–28.

Franklin-Hall, L. R. (2014). High-level explanation and the interventionist's 'variables problem.' *Br. J. Philos. Sci.*, *67*(2), 553–577.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *arXiv* preprint, 1609.07236.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, *29*(5), 1189–1232.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *Ann. Statist.*, *2*(3), 916–954.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, *33*(1), 1–41.

Frosst, N., & Hinton, G. E. (2017). Distilling a neural network into a soft decision tree. *arXiv* preprint, 1711.09784.

Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., & Feige, I. (2020). Shapley explainability on the data manifold. *arXiv* preprint, 2006.01272.

Frye, C., Feige, I., & Rowat, C. (2019). Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems 34*.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems 20*, 489–496.

Galles, D., & Pearl, J. (1995). Testing identifiability of causal effects. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 185–195. San Francisco, CA: Morgan Kaufmann Publishers Inc.

Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspect. Psychol. Sci.*, *9*(6), 641–651.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (Third Edit). Boca Raton, FL: Chapman and Hall/CRC.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *Br. J. Math. Stat. Psychol.*, *66*(1), 8–38.

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. *Ecol. Model.*, *160*(3), 249–264.

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. Boczkowski, & K. Foot (Eds.), *Media Technologies: Essays on Communication, Materiality, and Society* (pp. 167–193). Cambridge, MA: The MIT Press.

Gimenez, J. R., & Zou, J. (2019). Discovering conditionally salient features with statistical guarantees. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (pp. 2290–2298). Long Beach, California, USA: PMLR.

Goldman, A. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and Knowledge* (pp. 1–25). Dordrecht: Reidel.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.*, *24*(1), 44–65.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems 27* (pp. 2672–2680).

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *Artif. Intell. Mag.*, *38*(3), 76–99.

Gramacki, A. (2018). *Nonparametric Kernel Density Estimation and its Computational Aspects*. New York: Springer.

Greenland, S. (2019). Valid P-values behave exactly as they should: Some misleading criticisms of P-values and their resolution with S-values. *Am. Stat.*, *73*(sup1), 106–114.

Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *Calif. L. Rev.*, *94*(4), 945–967.

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Comput. Stat. Data Anal.*, *90*, 15–35.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. *Advances in Neural Information Processing Systems 19*, 513–520.

Grice, P. (1989). *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Griffel, F. (2020). al-Ghazali. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 202). Metaphysics Research Lab, Stanford University.

Grimm, S. R. (2006). Is understanding a species of knowledge? *Br. J. Philos. Sci.*, *57*(3), 515–535.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition. *Am. Stat.*, *61*(2), 139–147.

Guedj, B. (2019). A primer on PAC-Bayesian learning. *arXiv* preprint, 1901.05353.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local rule-based explanations of black box decision systems. *arXiv* preprint, 1805.10820.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, *51*(5), 1–42.

Gunning, D. (2017). *Explainable Artificial Intelligence (XAI)*. DARPA. Retrieved from https:// www.darpa.mil/attachments/XAIProgramUpdate.pdf.

Guyon, Isabelle, & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*(7/8), 1157–1182.

Haavelmo, T. (1944). The probability approach in econometrics. *Econometrica*, *12*, iii–115.

Habermas, J. (1981). *Theory of Communicative Action* (T. McCarthy, Trans.). Cambridge: Polity Press.

Hall, P. (2018). Building explainable machine learning systems: The good, the bad, and the ugly. Retrieved from: https://www.youtube.com/watch?v=Q8rTrmqUQsU.

Halpern, J. Y. (2016). *Actual Causality*. Cambridge, MA: MIT Press.

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 501–512. New York: ACM Press.

Hansson, S. O. (2017). Science and pseudo-science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University.

Harman, G., & Kulkarni, S. (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge, MA: The MIT Press.

Harman, G., & Kulkarni, S. R. (2011). Statistical learning theory as a framework for the philosophy of induction. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Philosophy of Statistics* (pp. 833–847). Oxford: Elsevier.

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *J. Environ. Econ. Manag.*, *5*(1), 81–102.

Hasani, R. (2019). A journey inside a neural network. *TED*. Retreived from https://www.ted.com/talks/ramin_hasani_a_journey_inside_a_neural_network.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL: Chapman and Hall/CRC.

Hausman, D. M., & Woodward, J. (2004). Modularity and the causal Markov condition: A restatement. *Br. J. Philos. Sci.*, *55*(1), 147–161.

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Appl. Stoch. Models Bus. Ind.*, *33*(1), 3–12.

Heinze-Deml, C., Peters, J., & Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *J. Causal Inference*, *6*(2).

Hempel, C., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philos. Sci.*, *15*, 135–175.

Hempel, Carl. (1965). *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Herschkowitz, J. I., Simin, K., Weigman, V. J., Mikaelian, I., Usary, J., Hu, Z., … Perou, C. M. (2007). Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. *Genome Biol.*, *8*(5), R76.

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Research.

Hintikka, J. (1999). *Inquiry as Inquiry: A Logic of Scientific Discovery*. New York: Springer.

Hitchcock, C. (1999). Contrastive explanation and the demons of determinism. *Br. J. Philos. Sci.*, *50*(4), 585–612.

HLEGAI. (2019). *Ethics guidelines for trustworthy AI*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

Hobsbawm, E. J. (1952). The machine breakers. *Past & Present*, *1*(1), 57–70.

Hodges, W., & Väänänen, J. (2019). Logic and games. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019). Metaphysics Research Lab, Stanford University.

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Inf. Commun. Soc.*, *22*(7), 900–915.

Holland, P. W. (1986). Statistics and causal inference. *J. Am. Stat. Assoc.*, *81*(396), 945–960.

Holm, S. (1979). A Simple sequentially rejective multiple test procedure. *Scan. J. Stat.*, *6*(2), 65–70.

Holzinger, A. (2019). From explainable AI to human-centered AI. *TED*. Retrieved from: https://www.ted.com/talks/andreas_holzinger_from_explainable_ai_to_human_centered_ai.

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

Horkheimer, M., & Adorno, T. (1947). *Dialectic of Enlightenment* (G. S. Noerr, Ed.; E. Jephcott, Trans.). Stanford, CA: Stanford University Press.

Hu, X., Rudin, C., & Seltzer, M. (2019). Optimal sparse decision trees. *Advances in Neural Information Processing Systems 32*, 7267–7275.

Huang, Y., & Valtorta, M. (2006). Pearl's Calculus of Intervention is Complete. *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, 217–224. Arlington, VA: AUAI Press.

Huang, Y., & Valtorta, M. (2008). On the completeness of an identifiability algorithm for semi-Markovian models. *Ann. Math. Artif. Intell.*, *54*(4), 363–408.

Hubbard, A. E., Kennedy, C. J., & van der Laan, M. J. (2018). Data-adaptive target parameters. In Mark J. van der Laan & S. Rose (Eds.), *Targeted Learning in Data Science* (pp. 125–142). New York: Springer.

Hume, D. (1739/2000). *A Treatise of Human Nature* (L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford: Clarendon Press.

Hume, D. (1748/1993). *An Enquiry Concerning Human Understanding*. (E. Steinberg, Ed.). Indianapolis, IN: Hackett Publishing.

Hyafil, L., & Rivest, R. L. (1976). Constructing optimal binary decision trees is NP-complete. *Inf. Process. Lett.*, *5*(1), 15–17.

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data Soc.*, *3*(2), 1–16.

Imbens, G. W. (2014). Instrumental variables: An econometrician's perspective. *Statist. Sci.*, *29*(3), 323–358.

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

James, W. (1975). *Pragmatism: A New Name for some Old Ways of Thinking*. Cambridge, MA: Harvard University Press.

Janzing, D., Minorics, L., & Bloebaum, P. (2020). Feature relevance quantification in `explainable AI: A causal problem. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 2907–2916.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science* (G. L. Bretthorst, Ed.). Cambridge: Cambridge University Press.

John, C. R., Watson, D., Barnes, M., Pitzalis, C., & Lewis, M. J. (2020). Spectrum: Fast density-aware spectral clustering for single and multi-omic data. *Bioinformatics*, *36*(4), 1159–1166.

Jones, S. E. (2006). *Against Technology: From the Luddites to Neo-Luddism*. New York: Routledge.

Jordon, J., Yoon, J., & van der Schaar, M. (2019). KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks. *International Conference on Learning Representations*. New Orleans, USA.

Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., & Wu, Z. S. (2019). Eliciting and enforcing subjective individual fairness. *arXiv* preprint, 1905.10660.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Penguin.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *J. Stat. Softw.*, *47*(11), 1–26.

Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *J. Clin. Epidemiol.*, *54*(6), 541–549.

Karimi, A.-H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv* preprint, 2010.04050.

Karimi, A.-H., Schölkopf, B., & Valera, I. (2020). Algorithmic recourse: From counterfactual explanations to interventions. *arXiv* preprint, 2002.06278.

Karimi, A.-H., von Kügelgen, J., Schölkopf, B., & Valera, I. (2020). Algorithmic recourse under imperfect causal knowledge: A probabilistic approach. *Advances in Neural Information Processing Systems 34*.

Kaufman, L., & Rousseeuw, P. (1990). *Finding Groups in Data*. Hoboken, NJ: John Wiley & Sons.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *Proceedings of the 35th International Conference on Machine Learning* (pp. 2564–2572).

Kearns, M., & Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. New York: Oxford University Press.

Kearns, M., & Vazirani, U. (1994). *An Introduction to Computational Learning Theory*. Cambridge, MA: The MIT Press.

Keiff, L. (2011). Dialogical logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University.

Kelly, K., Genin, K., & Lin, H. (2016). Realism, rhetoric, and reliability. *Synthese*, *193*(4), 1191–1223.

Khalifa, K. (2012). Inaugurating understanding or repackaging explanation? *Philos. Sci.*, *79*(1), 15–37.

Khuller, S., Moss, A., & Naor, J. Seffi. (1999). The budgeted maximum coverage problem. *Inf. Process. Lett.*, *70*(1), 39–45.

Kilbertus, N., Kusner, M. J., & Silva, R. (2020). A class of algorithms for general instrumental variable models. *Advances in Neural Information Processing Systems 34*.

Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! Criticism for interpretability. *Advances in Neural Information Processing Systems 29*, 2280–2288.

Kim, B., Rudin, C., & Shah, J. (2014). The Bayesian case model: A generative approach for

case-based reasoning and prototype classification. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, 1952–1960.

Kim, M., Reingold, O., & Rothblum, G. (2018). Fairness through computationally-bounded awareness. *Advances in Neural Information Processing Systems 31*, 4842–4852.

Kinney, D. (2018). On the explanatory depth and pragmatic value of coarse-grained, probabilistic, causal explanations. *Philos. Sci., 86*(1), 145–167.

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

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *Q. J. Econ.*, *133*(1), 237–293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2018). Discrimination in the age of algorithms. *J. Leg. Anal., 10*, 113–174.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (pp. 43.1-43.23).

Koller, D., & Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability* (N. Morrison, Ed. & Trans.). New York: Chelsea Publishing Company.

Kontschieder, P., Fiterau, M., Criminisi, A., & Bulò, S. R. (2015). Deep neural decision forests. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1467–1475.

Korb, K. B., & Nicholson, A. E. (2009). *Bayesian Artificial Intelligence* (2nd Edition). Boca Raton, FL: Chapman and Hall/CRC.

Korthauer, K., Kimes, P. K., Duvallet, C., Reyes, A., Subramanian, A., Teng, M., … Hicks, S. C. (2019). A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.*, *20*(1), 118.

Krishnan, M. (2019). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philos. Technol.*, *33*, 487-502.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Conference on Neural Information Processing Systems 25*, 1097–1105.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *J. Exp. Psychol. Gen.*, Vol. 142, pp. 573–603.

Kuang, C. (2017, November). Can AI be taught to explain itself? *The New York Times Magazine*. Retreived from https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html.

Kuhn, M., & Johnson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Boca Raton, FL: Chapman and Hall/CRC.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press.

Kumar, I., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In H. Daumé & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 1–10).

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the Boruta package. *J. Stat. Softw. 36*(11), 1–13.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems 30*, 4066–4076.

Kyburg, H. (1992). The scope of Bayesian reasoning. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *2*, 139–152.

Lage, I., Chen, E., He, J., Narayanan, M., Gershman, S., Kim, B., & Doshi-Velez, F. (2018). An evaluation of the human-interpretability of explanation. *Conference on Neural Information Processing Systems Workshop on Correcting and Critiquing Trends in Machine Learning*.

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading black box explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 79–85.

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 131–138.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K. R., & Samek, W. (2016). Analyzing classifiers: Fisher vectors and deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912–2920.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. Technical report, *ProPublica*. Retreived from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Latour, B., & Woolgar, S. (1979). *Laboratory Life: The Construction of Scientific Facts*. Princeton, NJ: Princeton University Press.

Lattimore, T., & Szepesvári, C. (2019). *Bandit Algorithms*. Cambridge: Cambridge University Press.

Lauritzen, S. L., & Richardson, T. S. (2002). Chain graph models and their causal interpretations. *J. Roy. Stat. Soc. B*, *64*(3), 321–348.

Lee, S., & Bareinboim, E. (2018). Structural causal bandits: Where to intervene? *Advances in Neural Information Processing Systems*, 2568–2578.

Lee, S., & Bareinboim, E. (2019). Structural causal bandits with non-manipulable variables. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(1), 4164–4172.

Legg, C., & Hookway, C. (2020). Pragmatism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University.

Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd Edition). New York: Springer.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.*, *113*(523), 1094–1111.

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *Eur. J. Oper. Res.*, *247*(1), 124–136.

Letham, B., Rudin, C., McCormick, T. H., & Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann. Appl. Stat.*, *9*(3), 1350–1371.

Levy, D. (Ed.). (1988). *Computer Chess Compendium*. New York: Springer.

Lewis, D. (1973a). Causation. *J. Philos.*, *70*, 556–567.

Lewis, D. (1973b). *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, *13*(4), 455–476.

Lewis, D. (1986). *Philosophical Papers, Volume II*. Oxford: Oxford University Press.

Lewis, D. (2000). Causation as influence. *J. Philos.*, *97*, 182–197.

Lim, E., Vaillant, F., Wu, D., Forrest, N. C., Pal, B., Hart, A. H., … Lindeman, G. J. (2009). Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. *Nat. Med.*, *15*, 907.

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *IEEE Access*, *5*, 16568–16575.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to Bivariate and Multivariate Analysis*. Glenview, IL: Longman.

Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, *61*(10), 36–43.

Lipton, Z. C. (2017). The doctor just won't accept that! *arXiv* preprint, 1711.08037.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv* preprint, 1907.11692.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. *Proceedings of the 36th International Conference on Machine Learning*,

4114–4124.

Luciano, F., & Taddeo, M. (2016). What is data ethics? *Philos. Trans. Royal Soc. A*, *374*(2083), 20160360.

Lundberg, S. (2019). Explainable AI for science and medicine. *Microsoft Research*. Retreived from https://www.microsoft.com/en-us/research/video/explainable-ai-for-science-and-medicine/.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., … Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, *2*(1), 56–67.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, 4765–4774.

Maathuis, M. H., Kalisch, M., & Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *Ann. Statist.*, *37*(6A), 3133–3164.

Mahajan, D., Tan, C., & Sharma, A. (2019). Preserving causal constraints in counterfactual explanations for machine learning classifiers. *CausalML: Machine Learning and Causal Inference for Improved Decision Making Workshop, NeurIPS 2019*.

Marx, K. (1867/1990). *Capital, Vol. I* (B. Fowkes, Trans.). London: Penguin.

Marx, K. (1885/1992a). *Capital, Vol. II* (D. Fernbach, Trans.). London: Penguin.

Marx, K. (1894/1992b). *Capital, Vol. III* (D. Fernbach, Trans.). London: Penguin.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. New York: Cambridge University Press.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philos. Sci.*, *71*(5), 1007–1025.

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., … Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94.

McQuillan, D. (2018). Data science as machinic neoplatonism. *Philos. Technol.*, *31*, 253–272.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *J. Roy. Stat. Soc. B*, *72*(4), 417–473.

Mentch, L., & Hooker, G. (2016). Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *J. Mach. Learn. Res.*, *17*(1), 841–881.

Menzies, P., & Beebee, H. (2020). Counterfactual theories of causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 202). Metaphysics Research Lab, Stanford University.

Merrick, L., & Taly, A. (2020). The explanation game: Explaining machine learning models

using Shapley values. *Machine Learning and Knowledge Extraction, 4th International Cross-Domain Conference (CD-MAKE)* (pp. 17–38).

Merton, R. (1973). The normative structure of science. In N. Storer (Ed.), *The Sociology of Science: Theoretical and Empirical Investigations* (pp. 267–278). Chicago: University of Chicago Press.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2018). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group*. CRAN. R package version 1.7-0.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, *267*, 1–38.

Mittelstadt, B. (2017). From individual to group privacy in big data analytics. *Philos. Technol.*, *30*(4), 475–494.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data Soc.*.

Mittelstadt, B., Russel, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of FAT* '19: Conference on Fairness, Accountability, and Transparency*. New York: ACM Press.

Molnar, C. (2020). *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*. Münich: Christoph Molnar. Retreived from https://christophm.github.io/interpretable-ml-book/.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., & Müller, K. R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit.*, *65*, 211–222.

Moravec, H. (1988). *Mind Children*. Cambridge, MA: Harvard University Press.

Morozov, E. (2013). *To Save Everything, Click Here: The Folly of Technological Solutionism*. New York: Penguin.

Morris, J. W. (2015). Curation by code: Infomediaries and the data mining of taste. *Eur. J. Cult. Stud.*, *18*(4–5), 446–463.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 607–617. New York: ACM Press.

Moulin, B., Irandoust, H., Bélanger, M., & Desbordes, G. (2002). Explanation and argumentation capabilities: towards the creation of more persuasive agents. *Artif. Intell. Rev.*, *17*(3), 169–222.

Mukherjee, S. (2017, April). A.I. versus M.D. *The New Yorker*. Retreived from https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.

Munkhdalai, L., Munkhdalai, T., Namsrai, O.-E., Lee, Y. J., & Ryu, H. K. (2019). An empirical

comparison of machine-learning methods on bank client credit assessments. *Sustainability* Vol. 11.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.*, *116*(44), 22071–22080.

Nalenz, M., & Villani, M. (2018). Tree ensembles with rule structured horseshoe regularization. *Ann. Appl. Stat., 12*(4), 2379-2408.

Narayanan, A. (2018). Tutorial: 21 fairness definitions and their politics. Retrieved from https://www.youtube.com/watch?v=jIXIuYdnyyk.

Nasrabadi, N. (2014). Hyperspectral target detection: An overview of current and future challenges. *IEEE Signal Process. Mag.*, *31*(1), 34–44.

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics*, *34*(21), 3711–3718.

Newman, N., Fletcher, R., Kalogeropoulos, A., & Nielsen, R. (2019). *Reuters Institute Digital News Report 2019* (Vol. 2019). Reuters Institute for the Study of Journalism.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Royal Soc. A*, *231*(694–706), 289–337.

Nicodemus, K. K., Malley, J. D., Strobl, C., & Ziegler, A. (2010). The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, *11*(1), 110.

Noble, S. U. (2018). *Algorithms of Oppression*. New York: New York University Press.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*. Retreived from https://www.oecd.org/going-digital/ai/principles/.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds Mach.*, *29*(3), 441–459.

Pareto, V. (1935). *The Mind and Society* (A. Livingston, Trans.). New York: Harcourt, Brace.

Parry, R. (2020). Episteme and Techne. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2020). Metaphysics Research Lab, Stanford University.

Pasquale, F. (2015). *The Black Box Society*. Cambridge, MA: Harvard University Press.

Patterson, E., & Sesia, M. (2018). *knockoff: The knockoff filter for controlled variable selection*. CRAN. R package, version 0.3.2.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Pearl, J. (2018). Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright. *Soc. Sci. Med., 210*, 60–62.

Peirce, C. S. (1999). *The Essential Peirce* (The Peirce Edition Project, Ed.). Bloomington, IN: Indiana Univesity Press.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Washington, D.C.: RAND Corporation.

Peters, J., Bühlmann, P., & Meinshausen, N. (2016). Causal inference by using invariant prediction: Identification and confidence intervals. *J. Roy. Stat. Soc. B*, *78*(5), 947–1012.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *The Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA: The MIT Press.

Peters, M. A. (2018). Deep learning, education and the final stage of automation. *Educ. Philos. Theory*, *50*(6–7), 549–553.

Phipson, B., & Smyth, G. (2010). Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Stat. Appl. Genet. Mol. Biol.*, *9*(1).

Plato. (1997). *Plato: Complete Works* (J. M. Cooper & D. S. Hutchison, Eds.). New York: Hackett.

Popper, K. (1934/1959). *The Logic of Scientific Discovery*. London: Routledge.

Popper, K. (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. New York: Routledge.

Popper, K. (1972). *Objective Knowledge: An Evolutionary Approach*. Oxford: Clarendon Press.

Potochnik, A. (2015). Causal patterns and adequate explanations. *Phil.. Stud.*, *172*(5), 1163–1182.

Potochnik, A. (2017). *Idealization and the aims of science*. Chicago: University of Chicago Press.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: Feasible and actionable counterfactual explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 344–350.

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., & Lipton, Z. C. (2020). Learning to deceive with attention-based explanations. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4782–4793.

Quine, W. van O. (1951). Two dogmas of empiricism. *Philos. Rev.*, *60*(1), 20–43.

Quine, W. van O. (1960). *Word and Object*. Cambridge, MA: The MIT Press.

Quine, W. van O. (1980). *Methods of Logic* (4th Edition). Cambridge, MA: Harvard University Press.

R Akula, A., Todorovic, S., Y Chai, J., & Zhu, S.-C. (2019). Natural language interaction with explainable AI models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramsey, J. D. (2014). A scalable conditional independence test for nonlinear, non-Gaussian data. *arXiv* preprint, 1401.5031.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press.

Rawls, J. (1971). *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Reichenbach, H. (1956). *The Direction of Time*. Los Angeles: University of California Press.

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

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: high-precision model-agnostic explanations. *AAAI*, 1527–1535.

Rinaldo, A., Wasserman, L., & G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.*, *47*(6), 3438–3469.

Robins, J. M. (1997). Causal inference from complex longitudinal data. In M. Berkane (Ed.), *Latent Variable Modeling and Applications to Causality* (pp. 69–117). New York, NY: Springer New York.

Romano, Y., Barber, R. F., Sabatti, C., & Candès, E. (2020). With malice toward none: Assessing uncertainty via equalized coverage. *Harvard Data Sci. Rev.*, *2*(2).

Romano, Y., Sesia, M., & Candès, E. (2019). Deep knockoffs. *J. Am. Stat. Assoc.*, 1–12.

Romeijn, J.-W. (2017). Philosophy of statistics. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 201). Metaphysics Research Lab, Stanford University.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychon. Bull. Rev.*, *16*(2), 225–237.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, *66*(5), 688–701.

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

Rudin, C., Wang, C., & Coker, B. (2020). The age of secrecy and unfairness in recidivism prediction. *Harvard Data Sci. Rev.*, *2*(1).

Russell, C. (2019). Efficient search for diverse coherent explanations. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 20–28. New York: ACM Press.

Sale, K. (1996). *Rebels Against the Future*. New York: Basic Books.

Salmon, W. (1971). Statistical explanation. In W. Salmon (Ed.), *Statistical Explanation and Statistical Relevance* (pp. 29–87). Pittsburgh: University of Pittsburgh Press.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). *Explainable AI: Interpreting, Explaining, and Visualizing Deep Learning*. New York: Springer.

Sanguinetti, G., & Huynh-Thu, V. A. (2018). *Gene Regulatory Networks: Methods and Protocols*. New York: Springer.

Sauer, N. (1972). On the density of families of sets. *J. Comb. Theory Ser. A., 13*(1), 145–147.

Savage, L. J. (1972). *The Foundations of Statistics*. New York: Dover Publications.

Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and Algorithms*. Camridge, MA: The MIT Press.

Schölkopf, Bernhard. (2003). Statistical learning theory, capacity, and complexity. *Complexity*, *8*(4), 87–94.

Schölkopf, Bernhard, & Smola, A. (2017). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond* (2nd Edition). Cambridge, MA: The MIT Press.

Schwab, P., & Karlen, W. (2019). CXPlain: Causal explanations for model interpretation under uncertainty.*Advances in Neural Information Processing Systems 32*, 10220–10230.

Scriven, M. (1962). Explanations, predictions, and laws. In H. Feigl & G. Maxwell (Eds.), *Scientific Explanation, Space, and Time* (pp. 170–230). Minneapolis: University of Minnesota Press.

Scutari, M. (2010). Learning Bayesian networks with the bnlearn R Package. *J. Stat. Softw.*, *35*(3), 1–22.

Scutari, M., & Denis, J.-B. (2014). *Bayesian Networks: With Examples in R*. Boca Raton, FL: Chapman and Hall/CRC.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*(7698), 604–610.

Selbst, A., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham L. Rev.*, *87*(3), 1085–1139.

Selbst, A., & Powles, J. (2017). Meaningful information and the right to explanation. *Int. Data Priv. Law*, *7*(4), 233–242.

Semenova, L., Rudin, C., & Parr, R. (2019). A study in Rashomon curves and volumes: A new perspective on generalization and model simplicity in machine learning. *arXiv* preprint, 1908.01755.

Sesia, M., Sabatti, C., & Candès, E. J. (2018). Gene hunting with hidden Markov model knockoffs. *Biometrika*, *106*(1), 1–18.

Shah, R., & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Ann. Statist.*, *48*(3), 1514–1538.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge: Cambridge University Press.

Shannon, C E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shannon, Claude E. (1950). Programming a computer for playing chess. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *41*(314), 256–275.

Shapley, L. (1953). A value for n-person games. In *Contributions to the Theory of Games* (pp. 307–317).

Sharifi-Malvajerdi, S., Kearns, M., & Roth, A. (2019). Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems 32*, 8242–8251.

Shelah, S. (1972). A combinatorial problem: Stability and orders for models and theories in infinitariy languages. *Pac. J. Math.*, *41*(1), 247–261.

Shmueli, G. (2010). To explain or to predict? *Statist. Sci.*, *25*(3), 289–310.

Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. *J. Mach. Learn. Res.*, *9*, 1941–1979.

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *Proceedings of the 34th International Conference on Machine Learning*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140 LP – 1144.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., … Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–

359.

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

Sokol, K., & Flach, P. (2020a). LIMEtree: Interactively Customisable Explanations Based on Local Surrogate Multi-output Regression Trees. *arXiv* preprint, 2005.01427.

Sokol, K., & Flach, P. (2020b). One explanation does not fit all. *Künstliche Intelligenz*, *34*(2), 235–250.

Sørlie, T, Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., … Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.*, *98*(19), 10869–10874.

Sørlie, Therese, Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., … Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci.*, *100*(14), 8418 LP – 8423.

Spanos, A. (2010). Theory testing in economics and the error-statistical perspective. In D.G. Mayo & A. Spanos (Eds.), *Error and Inference* (pp. 202–246). New York: Cambridge University Press

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, Prediction, and Search* (2nd Edition). Cambridge, MA: The MIT Press.

Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., … Collins, J. J. (2020). A deep learning approach to antibiotic discovery. *Cell*, *180*(4), 688-702.e13.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. Roy. Stat. Soc. B*, *64*(3), 479–498.

Strawson, P. F. (1964). Intention and convention in speech acts. *Philos. Rev.*, *73*(4), 439–460.

Strevens, M. (2010). *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Strevens, M. (2013). No understanding without explanation. *Stud. Hist. Philos. Sci.*, *44*(3), 510–515.

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*(1), 307.

Strobl, E. V., Zhang, K., & Visweswaran, S. (2018). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *J. Causal Inference*, *7*(1), 200180017.

Štrumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, *41*(3), 647–665.

Stutz, D., Hermans, A., & Leibe, B. (2018). Superpixels: An evaluation of the state-of-the-art. *Comput. Vis. Image Underst.*, *166*, 1–27.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., & Gillette, M. A. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.*, *102*(43), 15545–15550.

Sundararajan, M., & Najmi, A. (2019). The many Shapley values for model explanation. *Proceedings of the ACM Conference*. New York: ACM.

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning*, 3319–3328.

Sutton, R., & Barto, A. (2018). *Reinforcement Learning: An Introductionq* (2nd Edition). Cambridge, MA: The MIT Press.

Taddeo, M. (2010a). An information-based solution for the puzzle of testimony and trust. *Soc. Epistemol.*, *24*(4), 285–299.

Taddeo, M. (2010b). Modelling trust in artificial agents: A first step toward the analysis of e-trust. *Minds Mach.*, *20*(2), 243–257.

Taddeo, M. (2019). Three ethical challenges of applications of artificial intelligence in cybersecurity. *Minds Mach.*, *29*(2), 187–191.

Taddeo, M., McCutcheon, T., & Floridi, L. (2019). Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nat. Mach. Intell.*, *1*(12), 557–560.

Talbott, W. (2016). Bayesian Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., & Blei, D. M. (2018). The holdout randomization test: Principled and easy black box feature selection. *arXiv* preprint 1811.00645.

Tarski, A. (1983). The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics* (2nd Edtion, pp. 152–278). Indianapolis: Hackett.

Thornton, S. (2019). Karl Popper. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Tian, J., & Pearl, J. (2002). A general identification condition for causal effects. *Eighteenth National Conference on Artificial Intelligence*, 567–573. Menlo Park, CA: AAAI.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B*, *58*(1), 267–288.

Toloşi, L., & Lengauer, T. (2011). Classification with correlated features: Enreliability of feature ranking and solutions. *Bioinformatics*, *27*(14), 1986–1994.

Topol, E. J. (2019a). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

Topol, E. J. (2019b). High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.*, *25*(1), 44–56.

Tromp, J., & Farnebäck, G. (2007). Combinatorics of Go. In H. J. van den Herik, P. Ciancarini,

& H. H. L. M. Donkers (Eds.), *Computers and Games* (pp. 84–99). Berlin, Heidelberg: Springer Berlin Heidelberg.

Turkle, S. (2017). *Alone Together: Why We Expect More from Technology and Less from Each Other* (2nd Edition). New York: Basic Books.

Turner, N. C., & Reis-Filho, J. S. (2006). Basal-like breast cancer and the BRCA1 phenotype. *Oncogene*, *25*, 5846.

Upadhyay, A., & Khandelwal, K. (2018). Applying artificial intelligence: Implications for recruitment. *Strategic HR Review*, *17*(5), 255–258.

Ustun, B., & Rudin, C. (2019). Learning optimized risk scores. *J. Mach. Learn. Res.*, *20*(150), 1–75.

Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 10–19. New York: ACM Press.

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134–1142.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., … Bernards, R. (2002). A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, *347*(25), 1999–2009.

van der Laan, M.J., & Rose, S. (Eds.). (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.

van der Laan, Mark J. (2006). Statistical inference for variable importance. *The International Journal of Biostatistics*, *2*(1).

van der Laan, Mark J., & Rose, S. (Eds.). (2018). *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. New York: Springer.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: John Wiley & Sons.

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies to their probabilities. *Theory Probab. Appl.*, *16*(2), 264–280.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th Edition). New York: Springer.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, 1–7.

Verma, T., & Pearl, J. (1991). Equivalence and synthesis of causal models. *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, 255–270. New York,

NY: Elsevier Science Inc.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., … Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, *575*(7782), 350–354.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wachter, S., & Mittelstadt, B. D. (2019). A right to reasonable inferences: Re-thinking data protection law in the age of Big Data and AI. *Colum. Bus. L. Rev.*, *2*, 443–493.

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

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.*, *31*(2), 841–887.

Walton, D. (2004). A new dialectical theory of explanation. *Philos. Explorations*, *7*(1), 71–89.

Walton, D. (2006). Examination dialogue: An argumentation framework for critically questioning an expert opinion. *J. Pragmat.*, *38*(5), 745–777.

Walton, D. (2011). A dialogue system specification for explanation. *Synthese*, *182*(3), 349–374.

Wasserman, L. (2000). Bayesian model selection and model averaging. *J. Math. Psychol.*, *44*(1), 92–107.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on P-values: Context, process, and purpose. *Am. Stat.*, *70*(2), 129–133.

Waters, A., & Miikkulainen, R. (2014). GRADE: Machine-learning support for graduate admissions. *Artif. Intell. Mag.*, *35*(1), 64–75.

Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds Mach.*, *29*(3), 417–440.

Watson, D., & Floridi, L. (2018). Crowdsourced science: sociotechnical epistemology in the e-research paradigm. *Synthese*, *195*(2), 741–764.

Watson, D., Krutzinna, J., Bruce, I. N., Griffiths, C. E. M., McInnes, I. B., Barnes, M. R., & Floridi, L. (2019). Clinical applications of machine learning algorithms: Beyond the black box. *Br. Med. J.*, *364*, 446-448.

Weber, M. (1930/2002). *The Protestant Ethic and the Spirit of Capitalism* (T. Parsons, Trans.). London: Routledge.

Weinberger, N. (2018). Faithfulness, coordination and causal coincidences. *Erkenntnis*, *83*(2), 113–133.

Weslake, B. (2010). Explanatory depth. *Philos. Sci.*, *77*(2), 273–294.

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t-test. *Psychon. Bull. Rev.*, *16*(4), 752–760.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Trans. Vis. Comput. Gr.*, *26*(1), 56–65.

Williams, M. (2016). Internalism, reliabilism, and deontology. In B. McLaughlin & H. Kornblith (Eds.), *Goldman and His Critics* (pp. 1–21). Oxford: John Wiley & Sons.

Williamson, B., Gilbert, P., Simon, N., & Carone, M. (2020). Nonparametric variable importance assessment using machine learning techniques. *Biometrics*.

Wittgenstein, L. (1953). *Philosophical Investigations* (R. Rhees & G. E. M. Anscombe, Eds.; G. E. M. Anscombe, Trans.). Oxford: Blackwell.

Wolpert, D H, & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.*, *1*(1), 67–82.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. (2008). Cause and Explanation in psychiatry: An interventionist perspective. In K. Kendler & J. Parnas (Eds.), *Philosophical Issues in Psychiatry* (pp. 287–318). Baltimore: Johns Hopkins University Press.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biol. Philos.*, *25*(3), 287–318.

Woodward, J. (2015). Interventionism and causal exclusion. *Philos. Phenomol. Res.*, *91*(2), 303–347.

Woodward, J. (2019). Scientific Explanation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Woodward, J., & Hitchcock, C. (2003). Explanatory generalizations, part I: A counterfactual account. *Noûs*, *37*(1), 1–24.

Worrall, J. (2007). Why there's no cause to randomize. *Br. J. Philos. Sci.*, *58*(3), 451–488.

Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.*, *77*(1).

Wu, D., & Smyth, G. K. (2012). Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.*, *40*(17), e133.

Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2019). Self-training with noisy student improves ImageNet classification. *arXiv* preprint, 1911.04252.

Yang, H., Rudin, C., & Seltzer, M. (2017). Scalable Bayesian rule lists. *Proceedings of the 34th International Conference on Machine Learning*, 3921–3930.

Yu, X., & Gen, M. (2010). *Introduction to Evolutionary Algorithms*. Berlin, Heidelberg:

Springer-Verlag.

Završnik, A. (2019). Algorithmic justice: Algorithms and big data in criminal justice settings. *Eur. J. Crimonol.*, 1477370819876762.

Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philos. Technol.*

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philos. Technol.*, *32*(4), 661–683.

Zhang, K., Peters, J., Janzing, D., & Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 804–813. Arlington, VA: AUAI Press.

Zhao, Q., & Hastie, T. (2019). Causal interpretations of black-box models. *J. Bus. Econ. Stat.*, 1–10.

Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives*. Ann Arbor, MI: University of Michigan Press.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Stat. Soc. B*, *67*(2), 301–320.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nat. Genet., 51*(1), 12–18.

Zuboff, S. (2019). *The Age of Surveillance Capitalism*. London: Profile Books.

Appendix A:
Chapter 6, Supplemental Materials

This appendix includes tables with coverage probabilities for all experiments, as well as plots depicting empirical Type I and Type II errors.

§1 Coverage

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Linear model	0.9514	0.9485	0.9533	0.9479
Support vector machine	0.9516	0.9550	0.9528	0.9527
Random forest	0.9505	0.9537	0.9546	0.9529
Neural network	0.9518	0.9533	0.9548	0.9501

Table A.1. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; continuous outcome with MSE loss; correlated predictors.

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Linear model	0.9523	0.9496	0.9517	0.9531
Support vector machine	0.9517	0.9491	0.9521	0.9539
Random forest	0.9500	0.9516	0.9498	0.9495
Neural network	0.9557	0.9544	0.9521	0.9532

Table A.2. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; continuous outcome with MAE loss; correlated predictors.

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Linear model	0.9494	0.9484	0.9549	0.9514
Support vector machine	0.9514	0.9544	0.9518	0.9542
Random forest	0.9533	0.9544	0.9503	0.9537
Neural network	0.9537	0.9525	0.9493	0.9538

Table A.3. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; continuous outcome with MSE loss; uncorrelated predictors.

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Linear model	0.9522	0.9486	0.9525	0.9502
Support vector machine	0.9511	0.9515	0.9513	0.9534
Random forest	0.9518	0.9511	0.9509	0.9517
Neural network	0.9506	0.9470	0.9508	0.9538

Table A.4. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; continuous outcome with MAE loss; uncorrelated predictors.

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Logistic regression	0.7405	0.9576	0.9029	0.9493
Support vector machine	0.9034	0.9525	0.9260	0.9503
Random forest	0.8585	0.9473	0.9080	0.9454
Neural network	0.9303	0.9529	0.9357	0.9519

Table A.5. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; classification outcome with MMCE loss; correlated predictors.

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Logistic regression	0.9511	0.9516	0.9510	0.9511
Support vector machine	0.9514	0.9520	0.9534	0.9495
Random forest	0.9503	0.9499	0.9521	0.9503
Neural network	0.9493	0.9522	0.9517	0.9518

Table A.6. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; classification outcome with CE loss; correlated predictors.

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Logistic regression	0.7651	0.9525	0.8968	0.9523
Support vector machine	0.9164	0.9503	0.9262	0.9532
Random forest	0.9023	0.9439	0.9110	0.9507
Neural network	0.9345	0.9504	0.9369	0.9514

Table A.7. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; classification outcome with MMCE loss; uncorrelated predictors.

Learner	Linear data		Non-linear data	
	t-Test	Fisher	t-Test	Fisher
Logistic regression	0.9510	0.9480	0.9487	0.9469
Support vector machine	0.9488	0.9492	0.9527	0.9493
Random forest	0.9532	0.9524	0.9499	0.9516
Neural network	0.9522	0.9523	0.9494	0.9522

Table A.8. Empirical coverage probabilities of 95% confidence intervals in the simulation study, calculated from 10^4 simulation replicates; classification outcome with CE loss; uncorrelated predictors.

§2 Errors

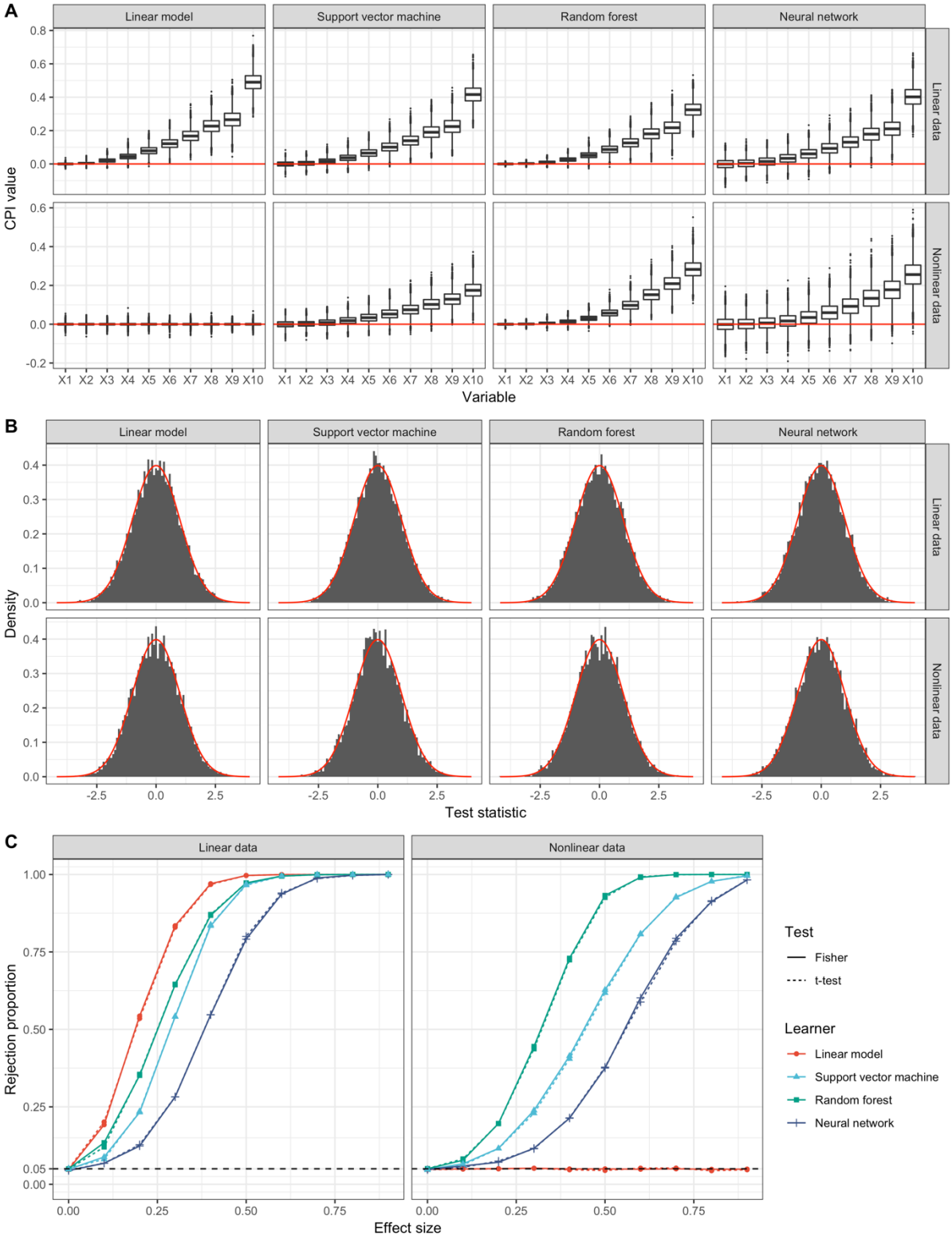

Figure A.1. Simulation results for continuous outcome with MAE loss and correlated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

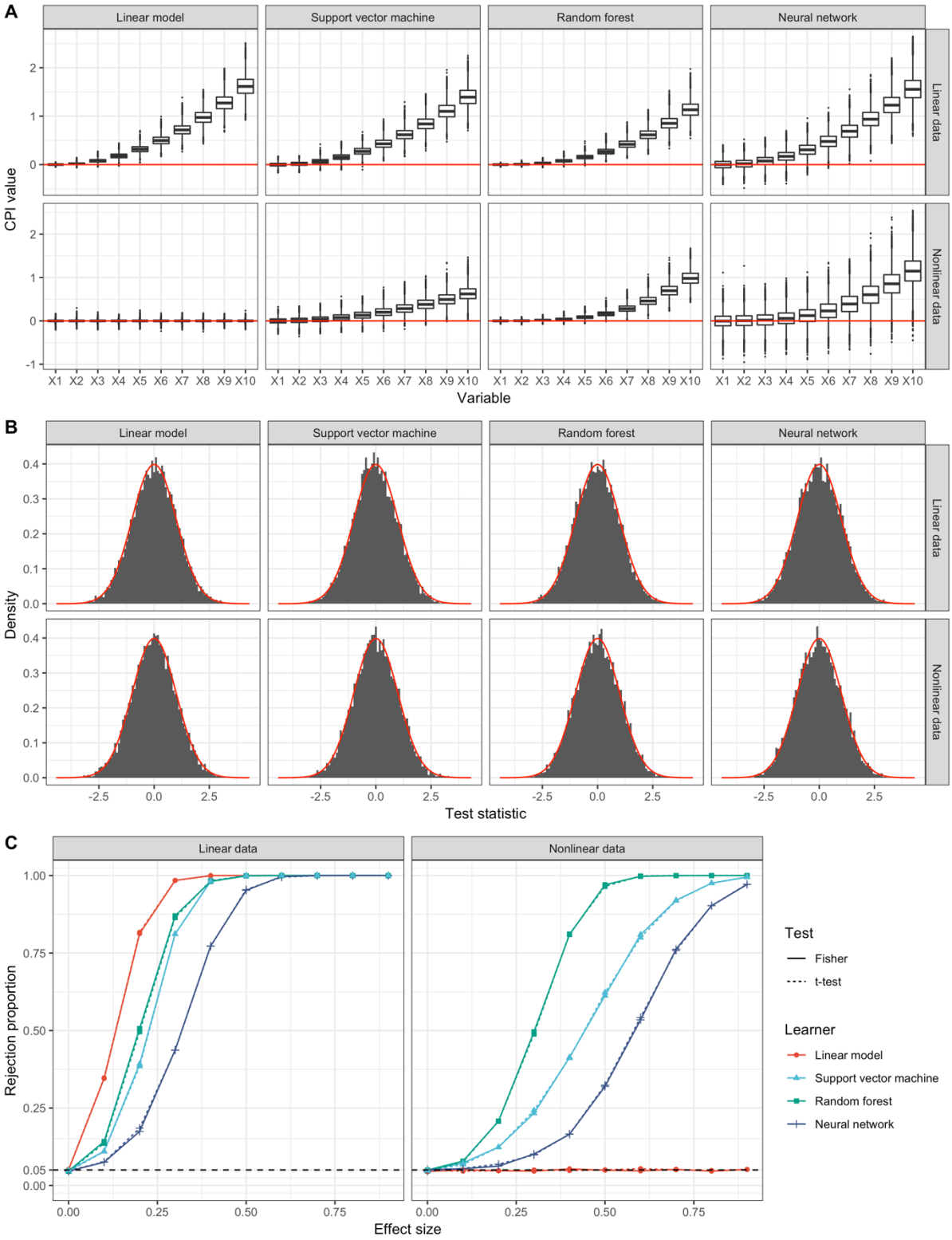

Figure A.2. Simulation results for continuous outcome with MSE loss and uncorrelated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

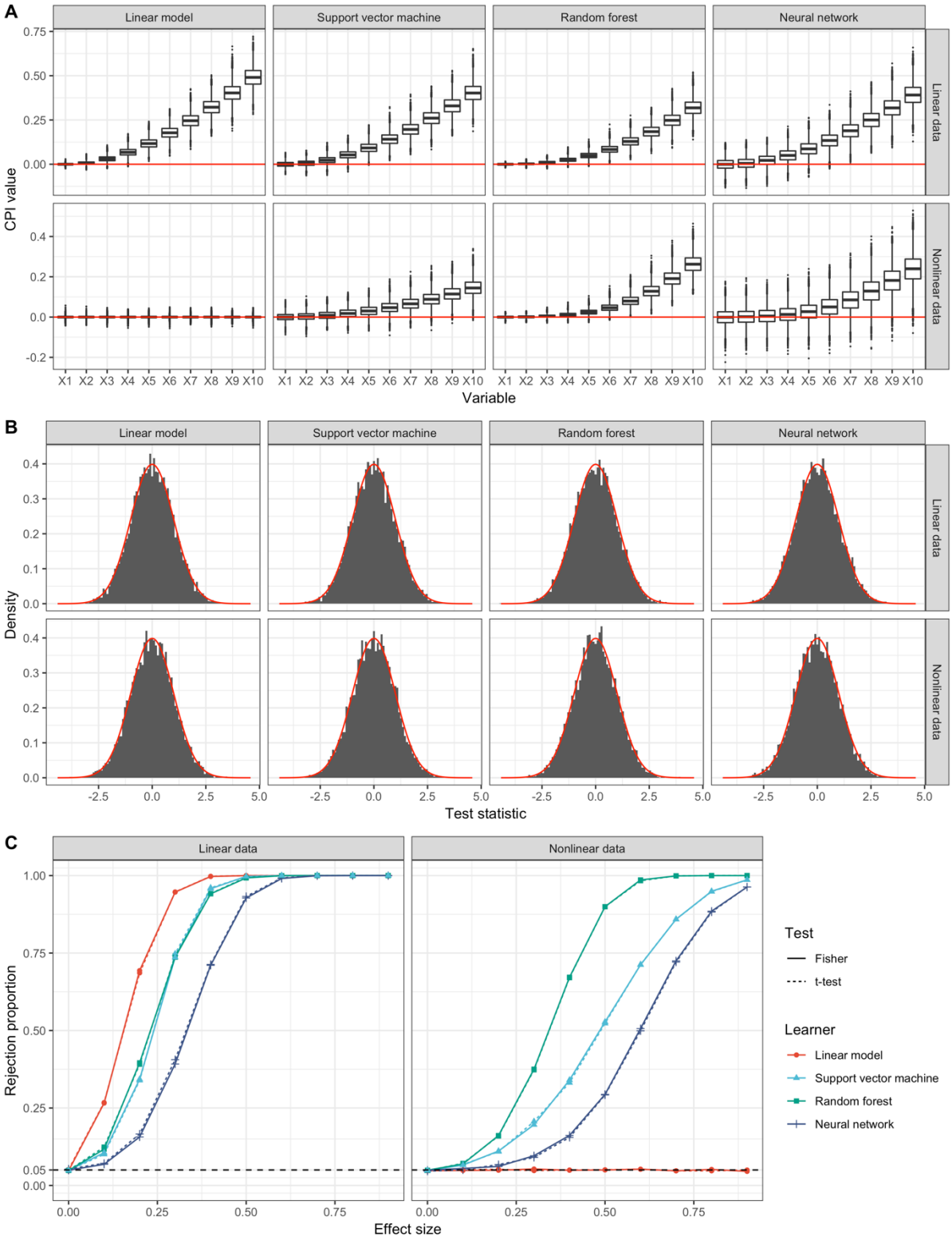

Figure A.3. Simulation results for continuous outcome with MAE loss and uncorrelated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

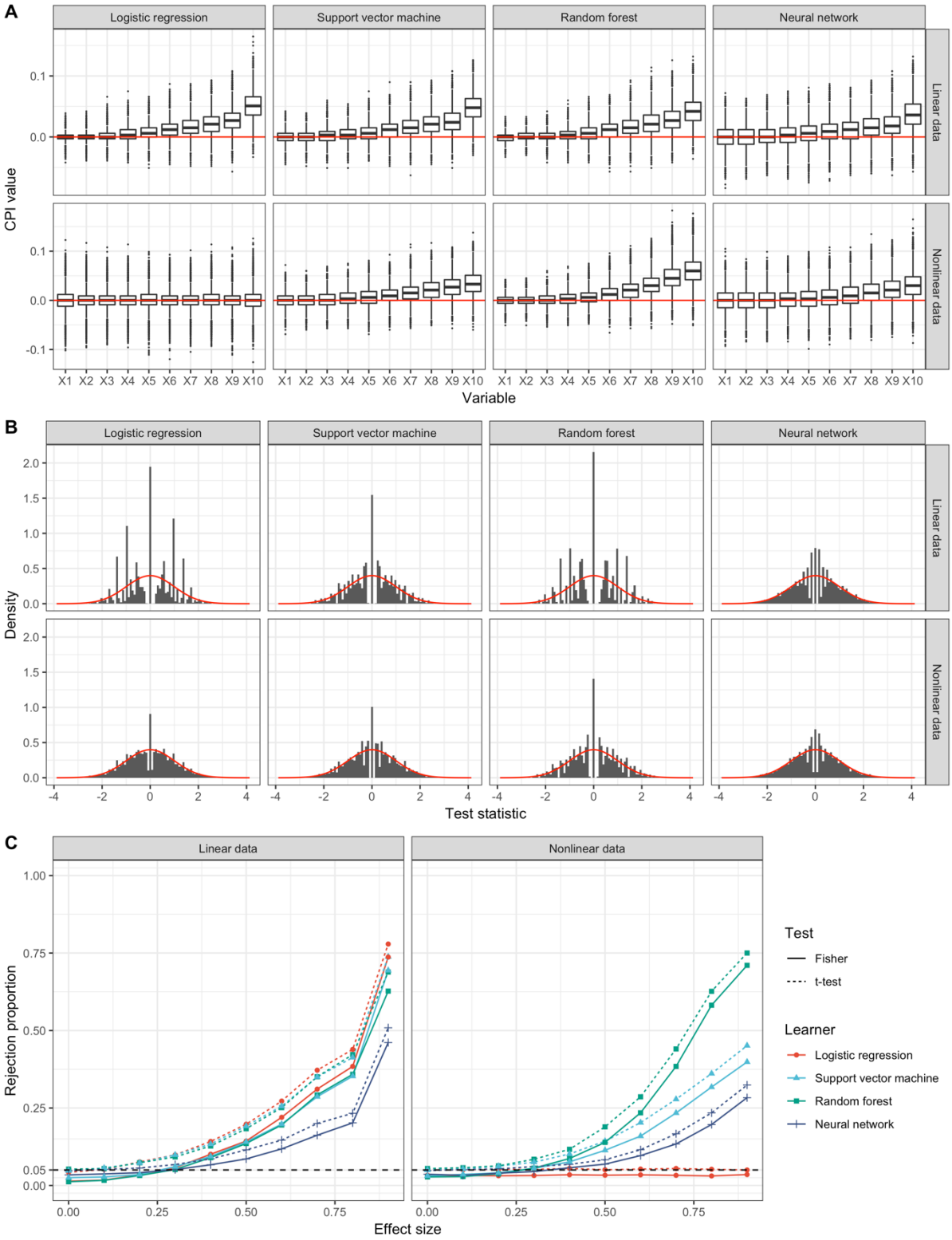

Figure A.4. Simulation results for binary outcome with MMCE loss and correlated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

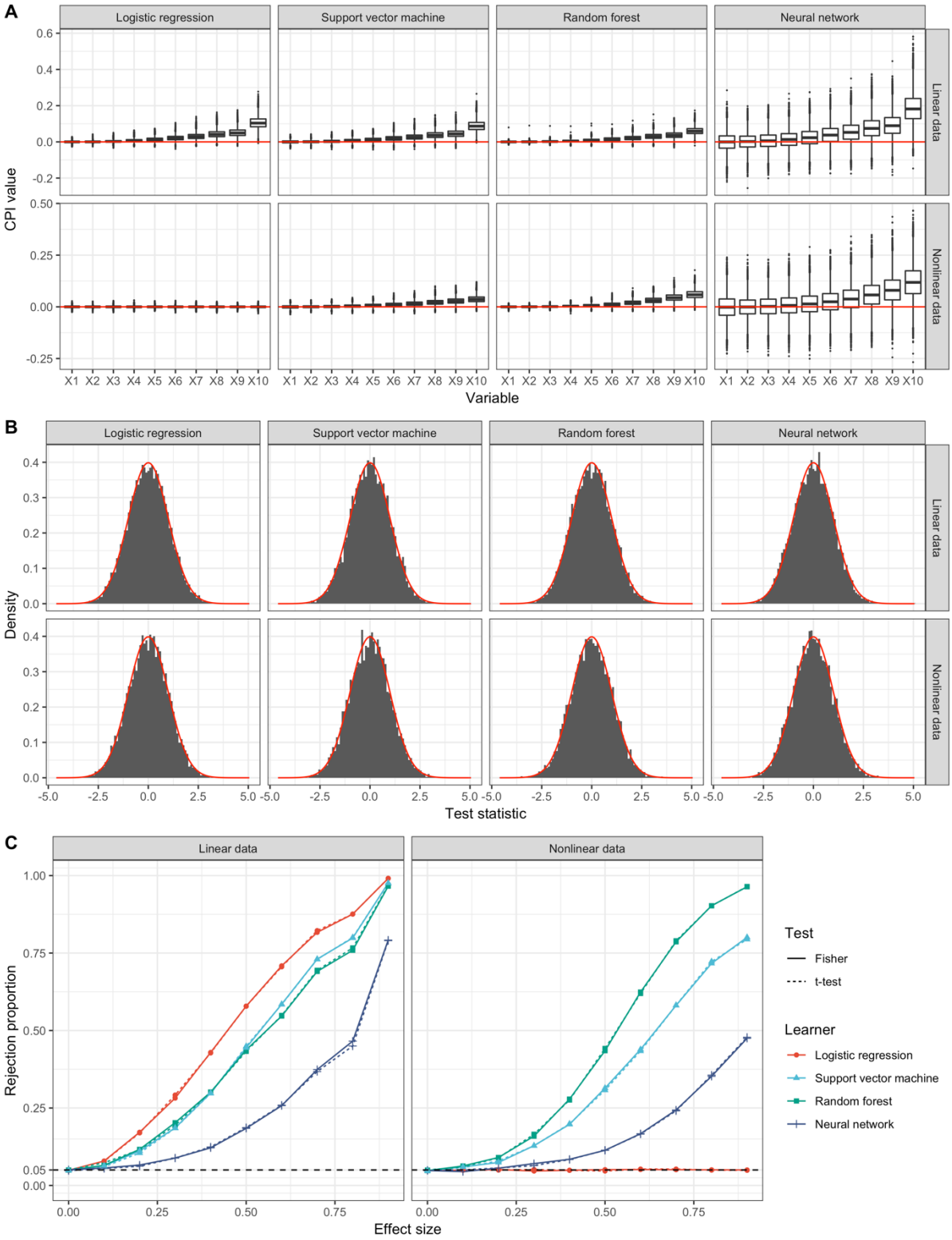

Figure A.5. Simulation results for binary outcome with CE loss and correlated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

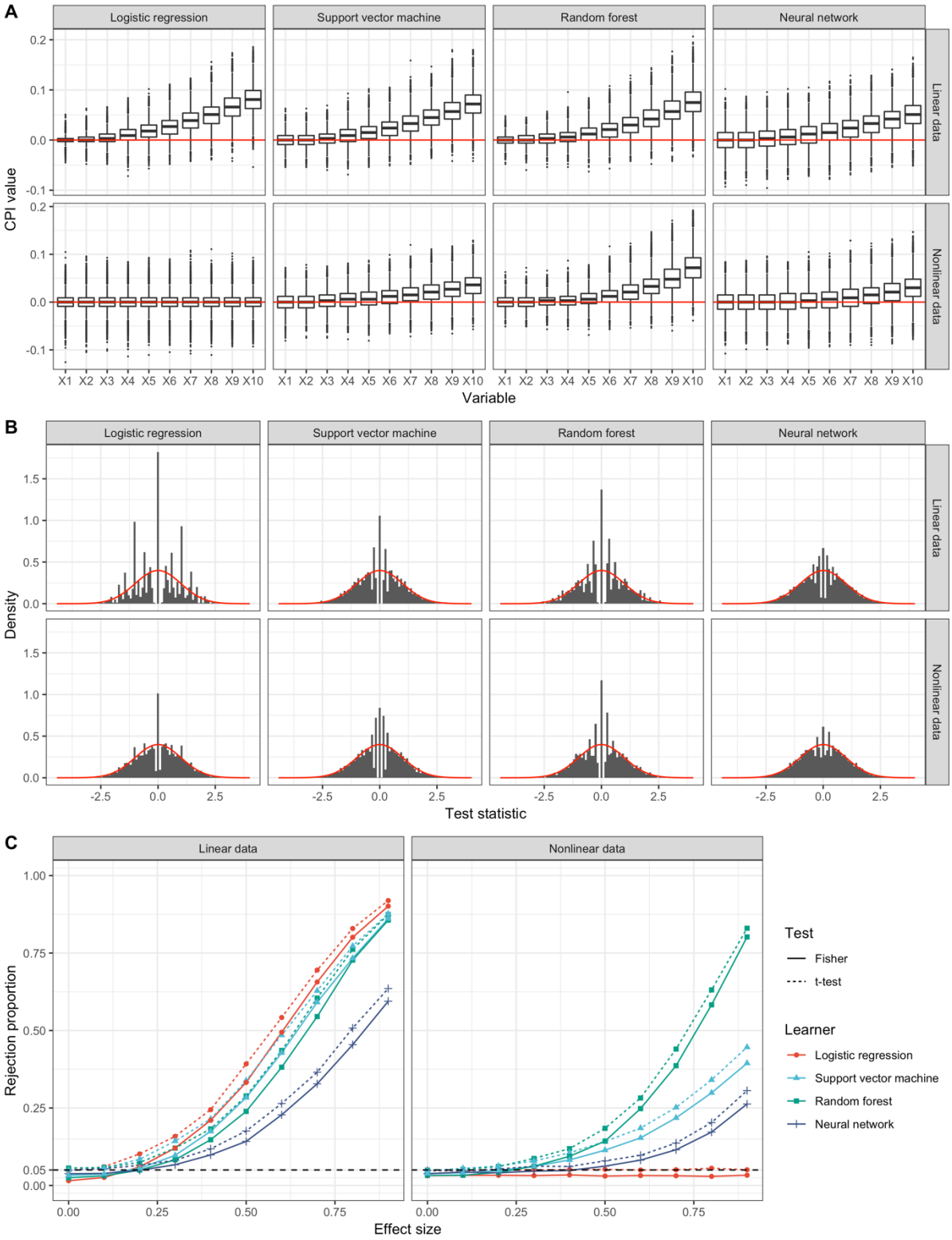

Figure A.6. Simulation results for binary outcome with MMCE loss and uncorrelated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

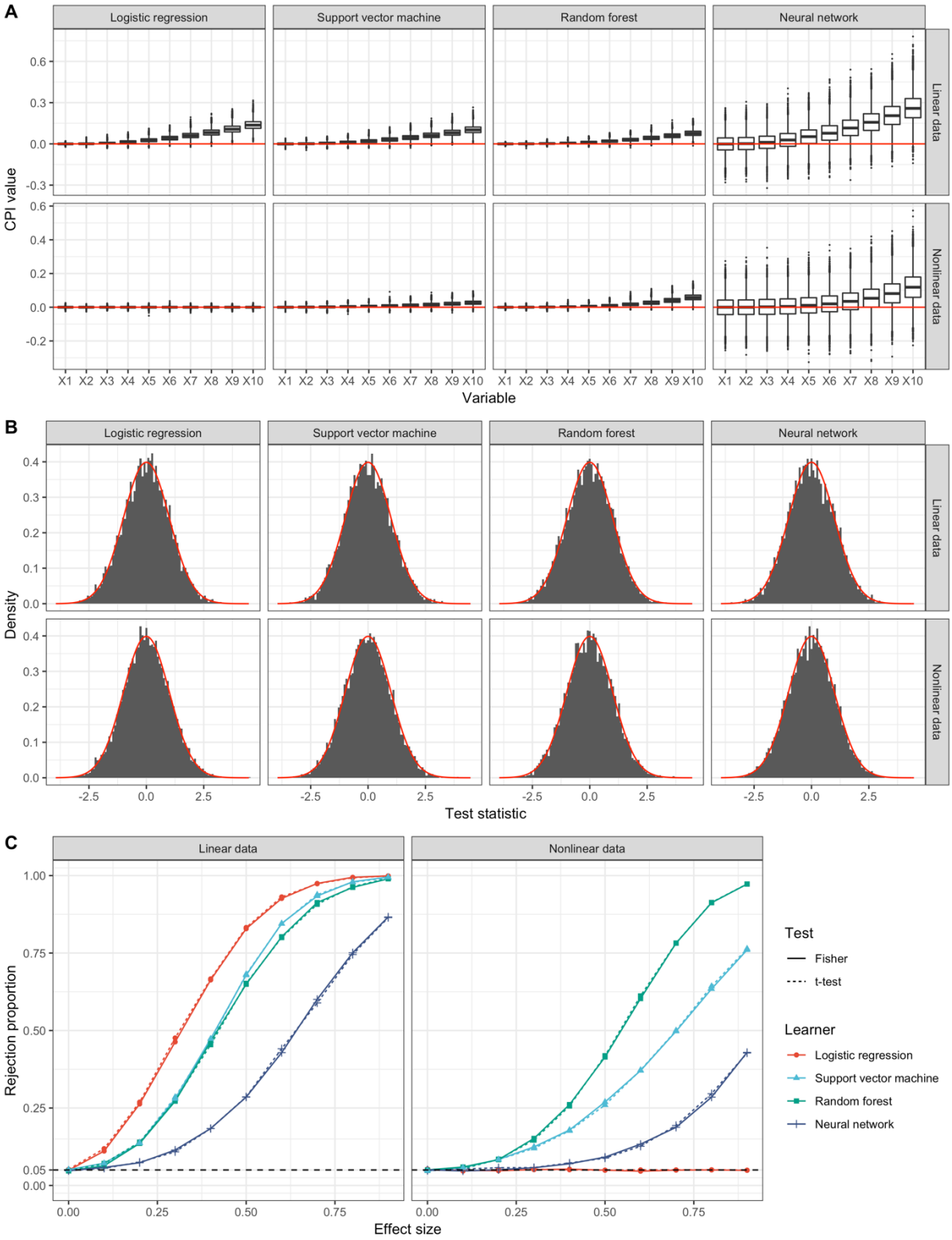

Figure A.7. Simulation results for binary outcome with CE loss and uncorrelated predictors. **A**: Boxplots of simulated CPI values of variables $X_1, \ldots, X_{10}$ with increasing effect size. The red line indicates a CPI value of 0, corresponding to a completely uninformative predictor. **B**: Histograms of simulation replications of t-statistics of variables with effect size 0. The distribution of the expected t-statistic under the null hypothesis is shown in red. **C**: Proportion of rejected hypotheses at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to the Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level, $\alpha = 0.05$.

§3 Comparative Performance

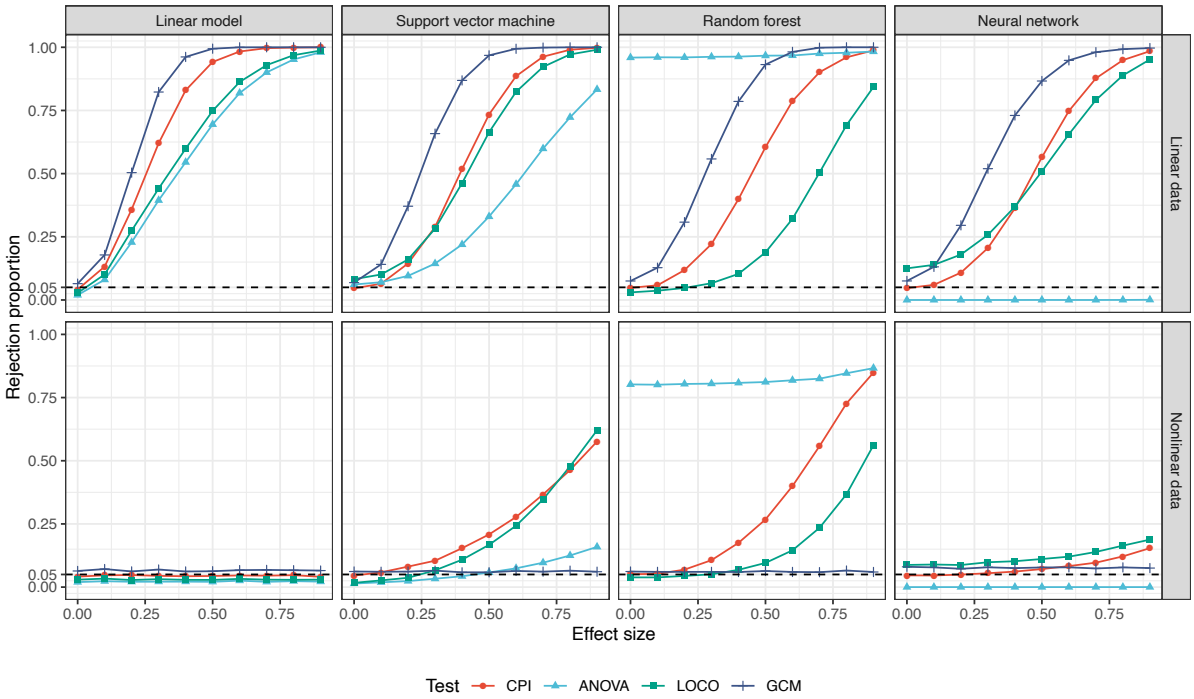

Figure A.8. Comparative performance of VI measures across different simulations and algorithms, computed with training and test samples of $n = 100, p = 10$ and correlated predictors. Plots depict the rejection rate at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level of $\alpha = 0.05$.

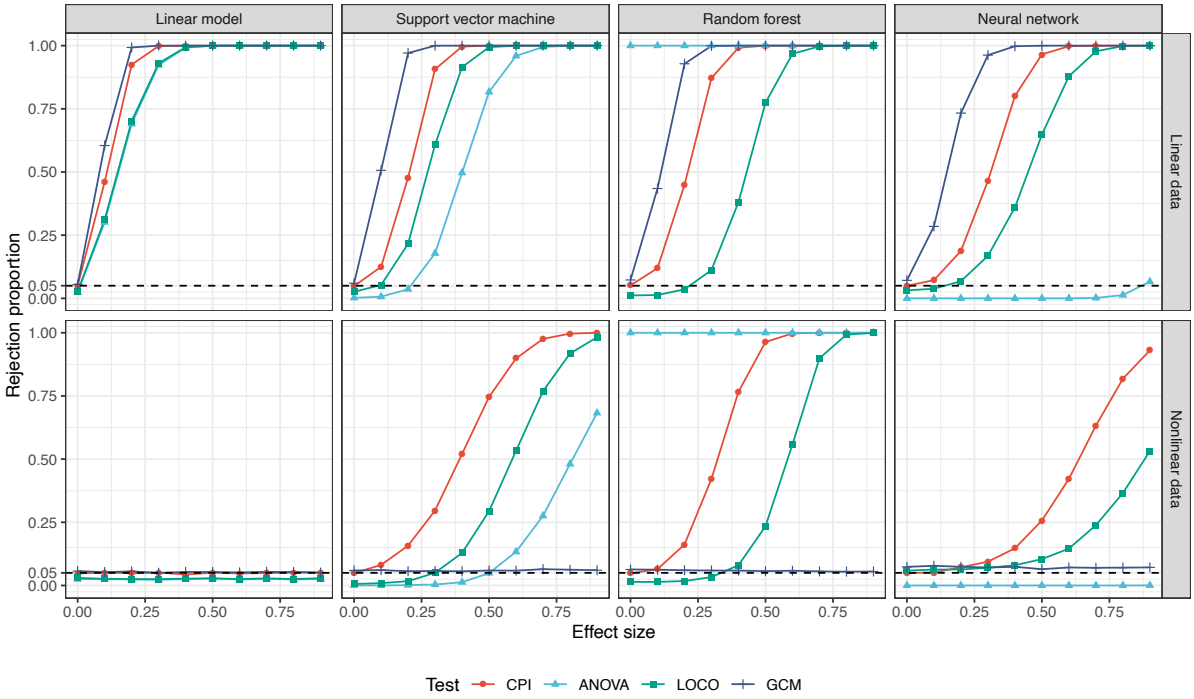

Figure A.9. Comparative performance of VI measures across different simulations and algorithms, computed with training and test samples of $n = 500, p = 10$ and correlated predictors. Plots depict the rejection rate at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level of $\alpha = 0.05$.

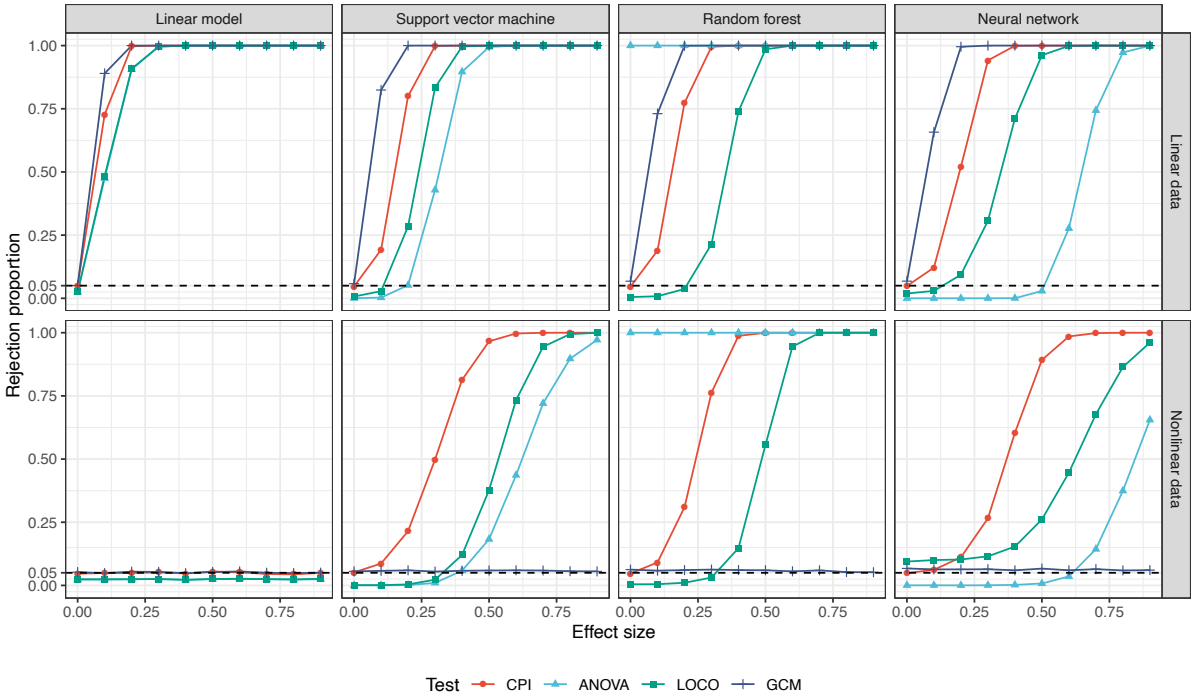

Figure A.10. Comparative performance of VI measures across different simulations and algorithms, computed with training and test samples of $n = 1000, p = 10$ and correlated predictors. Plots depict the rejection rate at $\alpha = 0.05$ as a function of effect size. Results at effect size 0 correspond to Type I error, at effect sizes > 0 to statistical power. The dashed line indicates the nominal level of $\alpha = 0.05$.

§4 Knockoff Sampler

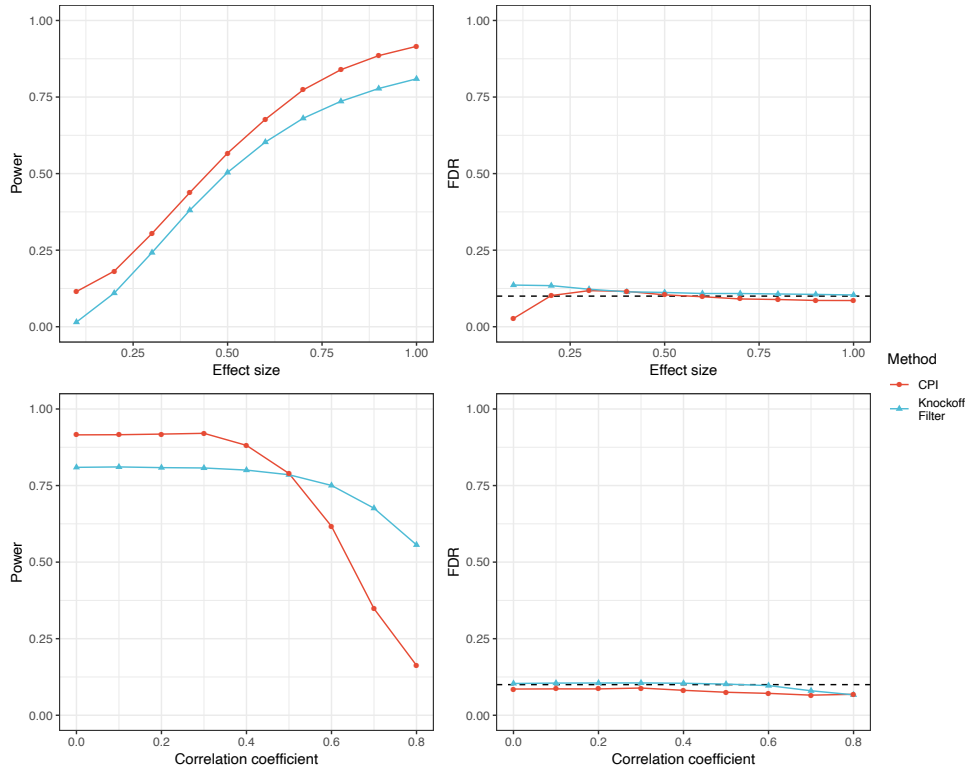

Figure A.11. Power and FDR as a function of effect size and autocorrelation for CPI and knockoff filter. Target FDR is 10%. Results are from a lasso regression with $n = 300, p = 2000$. Each point represents 10,000 replications.

Appendix B:
Chapter 7, Supplemental Materials

For completeness, this Appendix enumerates the axioms of Shapley values, utility theory, probability calculus, and *do*-calculus.

§1 Shapley Axioms

The axiomatic guarantees of Shapley values have been variously described (Sundararajan & Najmi, 2019). I list here four commonly cited properties of particular relevance to algorithmic explainability.

Efficiency. For all samples $i \in [n]$ and baselines ϕ_0,

$$\sum_{j=1}^{d} \phi_j = f(\boldsymbol{x}_i) - \phi_0.$$

This ensures that that attributions sum to the difference between observed and expected outputs for f.

Linearity. For all features $j \in [d]$ and models f_1, f_2, there exist some $a, b \in \mathbb{R}$ such that

$$\phi_j(af_1 + bf_2) = a\phi_j(f_1) + b\phi_j(f_2),$$

where $\phi_j(f)$ makes explicit the dependence of attributions on models via Eqs. 7.2 and 7.3. This ensures that feature attributions for a linear ensemble model are a linear combination of attributions for ensemble members.

Sensitivity. If for all pairs of values $x_{aj},\ x_{bj}$ of the variable X_j and all subvectors $\boldsymbol{x}^R$ such that $R = [d]\backslash\{j\}$,

$$f\left(x_{aj}, \boldsymbol{x}^R\right) = f\left(x_{bj}, \boldsymbol{x}^R\right) \rightarrow \phi_j = 0.$$

This ensures that irrelevant features receive zero attribution.

Symmetry. For all variable pairs $i, j \in [d]$ and subsets $S \subseteq [d]\backslash\{i, j\}$,

$$v(S \cup \{i\}) = v(S \cup \{j\}) \rightarrow \phi_i = \phi_j.$$

This ensures that features with the same expected impact on f receive equal attribution.

§2 Utility Axioms

The axioms of utility theory, as laid out by Von Neumann & Morgenstern (1944), are as follows. Assume a fixed agent with preferences over lotteries. If the agent strictly prefers A to B, write

$A \succ B$. If the agent is indifferent between A and B, write $A \sim B$. If the agent either prefers A to B or is indifferent between them, write $A \succcurlyeq B$.

Completeness. For any two lotteries A, B,

$$A \succcurlyeq B \vee B \succcurlyeq A.$$

Transitivity. For any three lotteries A, B, C,

$$A \succcurlyeq B \wedge B \succcurlyeq C \rightarrow A \succcurlyeq C.$$

Continuity. There exists some probability $p \in [0,1]$ such that

$$A \succcurlyeq B \succcurlyeq C \rightarrow pA + (1-p)B \sim C.$$

Independence. For any C and $p \in (0,1]$,

$$A \succcurlyeq B \leftrightarrow A + (1-p)C \succcurlyeq pB + (1-p)C.$$

§3 Probability Axioms

The axioms of the probability calculus, as laid out by Kolmogorov (1950), are as follows. Let (Ω, F, P) be a measure space with sample space Ω, event space F, and probability measure P.

Non-negativity. The probability of any elementary event $A \in F$ is a non-negative real number.

$$P(A) \in \mathbb{R}_{\geq 0}.$$

Unit measure. At least one elementary event in the sample space is guaranteed to occur.

$$P(\Omega) = 1.$$

σ-additivity. Any countable sequence of disjoint sets (i.e., mutually exclusive events) $A_1, A_2, \ldots$ satisfies

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

§4 *do*-calculus Axioms

The following axioms, first laid out by Pearl (2000), constitute a complete set of rules for reasoning about atomic interventions in SCMs (Huang & Valtorta, 2006). This exposition, including the numbering conventions, is taken from (Pearl, 2012). Let X, Y, Z, W denote arbitrary disjoint sets of nodes in the causal DAG $\mathcal{G}$. Let $\mathcal{G}_{\overline{X}}$ denote the graph obtained by deleting from $\mathcal{G}$ all arrows pointing to nodes in X. Let $\mathcal{G}_{\underline{X}}$ denote the graph obtained by deleting from $\mathcal{G}$ all arrows emerging from nodes in X. To represent deletion of both incoming and outgoing arrows, we write $\mathcal{G}_{\overline{X}\underline{Z}}$.

Rule 1: Insertion/deletion of observations. If $(Y \perp Z | X, W)_{\mathcal{G}_{\overline{X}}}$, then:

$$P(y|do(x), z, w) = P(y|do(x), w).$$

This ensures that conditional independencies in $\mathcal{G}$ unaffected by deleting edges into X are preserved after intervention.

Rule 2: Action/observation exchange. If $(Y \perp Z|X,W)_{\mathcal{G}_{\overline{X}\underline{Z}}}$, then:

$$P(y|do(x),do(z),w) = P(y|do(x),z,w).$$

This states the graphical conditions under which observations can be substituted for interventions and vice versa.

Rule 3: Insertion/deletion of actions. If $(Y \perp Z|X,W)_{\mathcal{G}_{\overline{XZ(W)}}}$, then:

$$P(y|do(x),do(z),w) = P(y|do(x),w),$$

where $Z(W)$ is the set of Z-nodes that are not ancestors of any W-node in $\mathcal{G}_{\overline{X}}$. This allows us to ignore interventions under certain graphical criteria.