Synthese (2022) 200:65
https://doi.org/10.1007/s11229-022-03485-5

ORIGINAL RESEARCH

Conceptual challenges for interpretable machine learning

David S. Watson[1]

Received: 6 August 2020 / Accepted: 2 November 2021

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 article, 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.

Keywords Artificial intelligence · Explainability · Scientific explanation · Causality · Severe testing

1 Introduction

Machine learning (ML) is ubiquitous in modern society. Complex learning algorithms are widely deployed in private industries like finance (Heaton et al., 2017) and insurance (Lin et al., 2017), as well as public services such as healthcare (Topol, 2019) and education (Peters, 2018). Their prevalence is largely driven by results. ML models

✉ David S. Watson
david.watson@ucl.ac.uk

[1] University College London, London, UK

Published online: 01 March 2022

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 antibiotic discovery (Stokes et al., 2020) and predicting protein structure (Jumper et al., 2021).

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, but epistemologists are well aware that a good track record is no guarantee of future success. Just as inductive inferences can lead us astray when presumptions of uniformity fail, so models can err when deployed in new contexts. This can lead to discriminatory predictions with potentially disastrous consequences 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, 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). I employ the former term because it emphasizes the subjective goal of interpretation over the (ostensibly) objective goal of explanation, while simultaneously specifying the focus on ML as opposed to more generic artificial intelligence tasks. IML comprises a diverse collection of technical approaches intended to render statistical predictions more intelligible to humans.[1] My focus in this article 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 article, 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.

[1] For good overviews of the current state of the art, see (Barredo Arrieta et al., 2020; Das & Rad, 2020; Molnar, 2019; Vilone & Longo, 2020).

A number of other conceptual challenges surrounding IML have already garnered much attention in the literature, especially those pertaining to subtle distinctions between explanations, interpretations, and understanding (Krishnan, 2020; Páez, 2019; Zednik, 2019); the purported trade-off between model accuracy and intelligibility (Rudin, 2019; Zerilli et al., 2019); as well as typologies and genealogies of algorithmic opacity (Burrell, 2016; Creel, 2020). I have little to add to those debates here, which I believe have been well argued by numerous authors. The challenges I highlight in this article, 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 to unpack points (1)-(3) and demonstrate their relevance for IML through a number of hypothetical and real-world examples. While each challenge is unique, together they point toward a singular conclusion—that despite undeniable technical advances, the conceptual foundations of IML remain underdeveloped. 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 article is structured as follows. I review relevant background material in Sect. 2, framing IML as a demand for causal explanations. In Sect. 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 Sect. 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 Sect. 5, arguing that satisfactory solutions must include some degree of user interaction and feedback. I conclude in Sect. 6 with a review of my findings and some reflections on the role and limits of philosophy as a theoretical guide in critiquing and designing algorithmic explanations.

2 Background

In this section, I provide necessary background on IML methods, as well as formal details on empirical risk minimization and structural causal models. Building on Woodward (2003)'s minimal theory of explanation, I frame the IML project as a certain sort of causal inquiry. This perspective elucidates the conceptual challenges that follow, as causal reasoning helps to disambiguate targets (Sect. 3), identify proper estimands for inference (Sect. 4), and ensure fruitful explanatory dialogue (Sect. 5).

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. Jack's 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 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 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 which, when set to certain values, are sufficient to bring about the outcome in question; and, when set to alternative values, are sufficient to alter the outcome in some prespecified way. Woodward (2003, p. 203) formalizes these criteria, stating that model $\mathcal{M}$ provides a causal explanation for outcome Y if and only if:

(i) The generalizations 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.

The full details of Woodward's program are beyond the scope of this article. However, his minimal account of explanation is a valuable starting point for analysis. In Jack's case, we may satisfy these criteria 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 $\boldsymbol{x}'$ identical to Jack's input $\boldsymbol{x}$ 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 explanation constitutes at least one viable explanans for the target explanandum.

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 instances 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 inputs that best explains the decision boundary or regression surface near a point of interest (see Fig. 1). Counterfactual explanations (Wachter et al., 2018), which account for predictions via synthetic matching techniques like those described above, 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

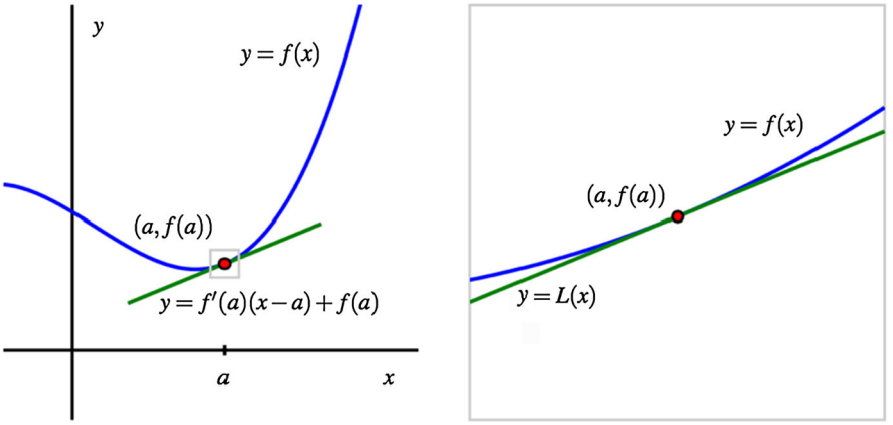

Fig. 1 A nonlinear function $f(x)$ (blue curve) is approximated by a linear function $L(x)$ (green curve) at the point $x = a$. Since L is simpler than f, it may help users better understand the model's predictive behavior near the input. Computing such tangents is the basic idea behind local linear approximators like LIME and SHAP

as Google,[2] Microsoft,[3] and IBM.[4] 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. At either resolution, successful answers must satisfy Woodward's three criteria. Those that fail to do so are unfaithful to their target (i), or else do not provide necessary (iii) or sufficient (ii) conditions for the explanandum.[5] This is perhaps most obviously true in the case of rule lists (see, e.g., Ribeiro et al., 2018), 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. Though there are important differences between this account and the interventionist theory I endorse here, the citation

[2] See https://pair-code.github.io/what-if-tool/.

[3] See https://github.com/interpretml/interpret.

[4] See http://aix360.mybluemix.net/.

[5] This insight has recently led several groups of IML researchers to design tools that directly optimize for these desiderata. See (Galhotra, Pradhan, & Salimi, 2021; Mothilal et al., 2021; Watson et al., 2021).

only serves to underscore the reliance of IML on causal frameworks—as well as the ambiguity this reliance can engender.

If the causal foundations 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 prioritized prediction over explanation (Breiman, 2001; Shmueli, 2010). I will briefly formalize the distinction between supervised learning and causal modelling to pre-empt any potential confusion and ground the following discussion in established theory.

2.2 Empirical risk minimization and structural causal models

A supervised learning algorithm is a method for predicting outcomes $Y \in \mathbb{R}^k$ based on inputs $\boldsymbol{X} \in \mathbb{R}^d$ with minimal error.[6] 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 unknown distribution $\mathbb{P}(\boldsymbol{Z})$. An algorithm is associated with a function space $\mathcal{F}$, and the goal is to find the model $f \in \mathcal{F}$ that minimizes some predetermined loss function $L(f, \boldsymbol{Z})$, which quantifies the distance between model outputs $f(\boldsymbol{X}) = \hat{Y}$ and true 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 minimization (ERM) is the learning strategy whereby we select whichever model attains the minimal loss within a given function class $\mathcal{F}$. ERM is provably consistent (i.e., guaranteed to converge uniformly upon the best model in $\mathcal{F}$) under two key assumptions (Vapnik & Chervonenkis, 1971): (1) samples are independently and identically distributed (i.i.d.); and (2) $\mathcal{F}$ is of bounded complexity.[7]

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).[8] As noted in Sect. 1, these algorithms have proven incredibly effective at predicting outcomes for complex tasks like image classification and natural language processing. 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

[6] In the classification setting, we typically 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 paper applies equally to regression and classification problems, as well as cases where $k > 1$.

[7] Exact proposals for bounding the complexity of $\mathcal{F}$ vary. In this article, 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 (Shalev-Shwartz & Ben-David, 2014).

[8] Note that though ERM may motivate these algorithms, they do not all enjoy uniform convergence guarantees. For instance, radial basis function kernels are known to be of infinite VC dimension, as are over-parametrized mixed layer networks. However, both methods work well in many settings. In this section, I identify ERM with the goal of minimizing empirical risk rather than the guarantee of uniform convergence.

assigned low probability of death to pneumonia patients with a history of asthma, a well-known risk factor for emergency room patients under acute pulmonary distress. The unexpected association was no simple 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 learning on its own.

Examples like this highlight the importance of interpretable explanations for high-stakes ML predictions such as 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. The external validity of a given model depends on structural facts about training and test environments (Pearl & Bareinboim, 2014), e.g. the assignment mechanism that dictates which patients are sent to the ICU. If we were to deploy the pneumonia triage algorithm in a new hospital where doctors are not already predisposed to provide extra care for asthma patients—perhaps a hospital where doctors rely exclusively on a high-performance ML model to prioritize treatment—then empirical risk may substantially underestimate the true generalization error. In light of these considerations, a number of prominent authors have advocated for an explicitly causal approach to statistical learning (Pearl, 2000; Peters et al., 2017; Spirtes et al., 2000; van der Laan & Rose, 2011). The basic strategy can be elucidated through the formalism of structural causal models (SCMs). A probabilistic SCM $\mathcal{M}$ is a tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathcal{F}, \mathbb{P}(\boldsymbol{u})\rangle$, 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; $\mathcal{F}$ is a set of deterministic functions mapping causes to direct effects; and $\mathbb{P}(\boldsymbol{u})$ is a probability distribution over $\boldsymbol{U}$. An SCM induces an associated graph, where nodes are variables and directed edges denote causal relationships (see Fig. 2). A fully specified $\mathcal{M}$ provides a map from background conditions to a joint distribution over observables, $\mathcal{M} : \boldsymbol{U} \rightarrow \mathbb{P}(\boldsymbol{v})$.

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 the submodel $\mathcal{M}_x$. Interventions are formally expressed by Pearl's (2000) *do*-operator. The interventional distribution $\mathbb{P}(Y|do(X = 1))$ may deviate considerably from the observational distribution $\mathbb{P}(Y|X = 1)$ within a given $\mathcal{M}$. 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

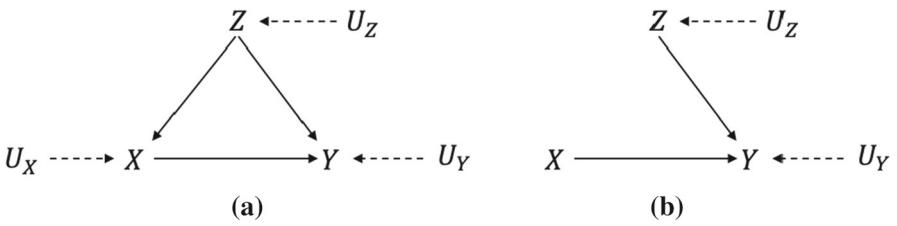

Fig. 2 Simple examples of causal graphs. Solid edges denote observed causal relationships, dashed edges unobserved. (a) A model with confounding between variables X and Y. (b) The same model after intervening on X, thereby eliminating all incoming causal effects and removing the confounding signal from Z

$\mathbb{P}(Y|X = 1) = \mathbb{P}(Y|Z = 1)$. However, if we randomly assign treatment to patients independent of their sex, then we may get a very different value for $\mathbb{P}(Y|do(X = 1))$, 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 randomized control trials (RCTs), which are widely used by scientists and regulatory agencies to establish treatment efficacy.[9] The *do*-calculus provides a provably complete set of rules for reasoning about interventions (Shpitser & Pearl, 2008), including criteria for deciding whether and how causal effects can be estimated from observational data.

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. Sullivan (2020) argues that algorithmic opacity derives not from any inherent complexity in models or systems per se, but rather from the "link uncertainty" that results when there is little empirical evidence connecting the two levels. Even when such links are well-established, however, it is not always clear which level is the intended target of explanation. Causal reasoning, as formalized by SCMs, can help diagnose and resolve issues of link uncertainty by making the assumptions of any given IML tool more explicit.

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 underspecified. 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 recognize 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.

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.

[9] The supremacy of RCTs for causal inference has not gone unchallenged. See (Deaton & Cartwright, 2018; Kaptchuk, 2001; Pearl, 2018; Worrall, 2007).

The problem of induction. Although commonly associated with Hume (1739, 1748) in the anglophone tradition, inductive skepticism goes back at least as far as Sextus Empiricus. The basic idea, familiar to philosophy undergraduates the world over, is that inference from particular observations to universal generalizations relies on some assumption 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). Skeptics argue that such a premise cannot be justified by reason (because it is conceivably false) or experience (as this would be circular). This poses major challenges for any account of causality that seeks to go beyond mere correlations, since, according to the inductive skeptic, 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. 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 a clinical study if it turns out that treatment and control groups differ substantially along relevant variables such as age or sex, as per the example above. 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 the Duhem-Quine thesis (Duhem, 1954; Quine, 1960), which states 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).

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. 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. This inspired Rubin's (1974) potential outcomes framework, in which causal inference is treated as a missing data problem. Lewis (1973) elevates this challenge into a unifying principle, reducing all causality to relations of counterfactual dependence. Among analytic philosophers, Quine (1960, 1980) is perhaps the most forceful in his opposition to this view, 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 unnecessary given careful Bayesian decision analysis.

Upon considering these issues, 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 behavior. 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.[10] 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? Causal inference is trivial at the model-LoA, where many IML tools implicitly operate, and notoriously challenging at the system-LoA, where many practitioners expect and require explanations.

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 $\boldsymbol{x}'$ achieves the desired outcome y', but there are legitimate concerns about how informative this is when $\boldsymbol{x}'$ 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.

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 et al., 2016), although user-friendly Python implementations have made the aforementioned LIME (Ribeiro et al., 2016) and SHAP

[10] For more on these points, see (Watson & Floridi, 2020, Sect. 4).

(Lundberg & Lee, 2017) algorithms dominant in recent years. Methodological differences notwithstanding, each of these algorithms attempts to answer Q by means of a linear combination of input features optimized to hold around the point $\boldsymbol{x}_i$. Crucially, these methods all assume that predictors are mutually independent, i.e. for all $j, k \in [d]$ such that $j \neq k$, $X_j \perp X_k$. This enables something like the naïve approach described above, in which predictors may be dialed 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 mutual 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 criticized the original SHAP algorithm for failing to model covariate dependencies and proposed various "improvements" that incorporate conditional information (Aas et al., 2021; Frye et al., 2020; Kumar et al., 2020). 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 should be interpreted. In statistical terms, we may formalize 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.

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 inflation is interpreted as strong evidence that f relies

on X_j to estimate outcomes Y. A more general "reliance" statistic is introduced by Fisher et al. (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 plot, originally proposed by Friedman (2001), is another popular PaP method. These graphs visualize the change in expected value of a function f as we marginalize over the empirical distribution of a feature subset while holding values for the complementary subset constant.

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 numerous authors have proposed alternatives designed to overcome this perceived shortcoming of Breiman's permutation importance (Altmann et al., 2010; Gregorutti et al., 2015; Mentch & Hooker, 2016; Nembrini et al., 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 et al., 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 and 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 to see f err 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, Sect. 8) demonstrate that under some common assumptions,[11] their model reliance measure may be decomposed into a sum of familiar terms from causal inference. Zhao and Hastie

[11] 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).

(2019) observe that Friedman's partial dependence function is formally identical to Pearl's (2000) backdoor adjustment when conditioning variables meet certain conditions.[12] In other words, the partial dependence of an outcome on a given feature subset in a purely predictive model may have a natural 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) defense of the original SHAP method. Both groups argue that marginalizing 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.

3.3 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 Sect. 3.1, lies in the realization 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 formalization that requires relatively little background, I turn to Floridi's (2011) correctness theory of truth (CTT). The complete details of Floridi's epistemology of information are beyond the scope of this article; for our purposes, I shall focus merely on how it clarifies the essential semantic work done (often implicitly) by pragmatic auxiliaries. According to the CTT, information can always be polarized 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:

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

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" (2011, 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.

[12] 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).

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 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 interests and motivations. 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 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 Sect. 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 substantially 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.

The choice of whether to assume the mutual independence of predictive features cannot be determined by mathematical considerations alone, for it depends crucially on the inquiring agent's CLP parameters. 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 troubleshooting 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 and planning, 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 behavior 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., 2021; Frye et al., 2020; Kumar et al., 2020), 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 a 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.

4 Error rates and severe testing

In Sect. 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 rigor in this young and fast-evolving field of research.

4.1 Severity Criteria

In a pair of influential monographs, Mayo (1996, 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 just of the hypothesis itself or the data it purportedly explains, but, crucially, of the tests it has passed. When a test of claim C given data x is sufficiently sensitive (i.e., likely to affirm true C) and specific (i.e., likely to reject false C), then we say it is *severe*.

Mayo works in the falsificationist tradition of Popper (1959). 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 (defeasible, statistical) evidence in favor of particular hypotheses. Unlike Bayesian epistemologists, who typically interpret probabilities as degrees of belief computed by combining subjective priors[13] with evidential likelihoods, Mayo places her approach in the frequentist tradition, emphasizing the importance of hypothesis testing with bounded 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 and 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.

[13] 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.

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.[14] 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).[15] 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.

Mayo grounds her severity approach in the Neyman-Pearson (NP) testing framework (Lehmann & Romano, 2005). To make matters concrete, I will briefly explicate her severity criteria using the simplest and most common sort of hypothesis in statistical inference, namely one positing some value or range of values for a single parameter.[16] Let T be a test that decides between, say, $H_0 : \mu \leq 0$ and $H_1 : \mu > 0$. 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 recognize 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. 3). 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

[14] See (Romeijn, 2017) for a good introduction. For a more comprehensive compendium, see (Bandyopadhyay & Forster, 2011).

[15] Somewhat disconcertingly, it can also be shown that, for any body of evidence, we may construct priors such that corresponding posteriors differ by an arbitrarily large amount. See (Kyburg, 1992).

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

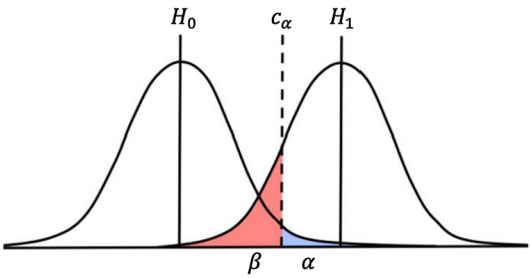

Fig. 3 Null and alternative distributions for a given hypothesis test. The critical value is denoted by the dashed line. Type I error is represented by the blue integral; type II error β is depicted by the red integral

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.

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 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 skeptical objections such as the aforementioned underdetermination of theory by evidence. The best antidote is generally a combination of statistical assumptions, epistemological theory, and rigorous misspecification tests; see (Mayo & Spanos, 2004) for an overview. No matter the details of particular testing methods, the point I want to stress is that minimizing 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.[17]

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 be only of the "accept" or "reject" variety? Surely, we can entertain intermediate degrees of belief. 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 hypotheses are accepted or rejected. A similar procedure occurs in Bayesian analysis, where decisions typically turn on Bayes factors or credible intervals. Blind insistence on particular 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

[17] Other sorts of statistical errors have also been studied, 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.

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 Sect. 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.

4.2 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 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.

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). 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.

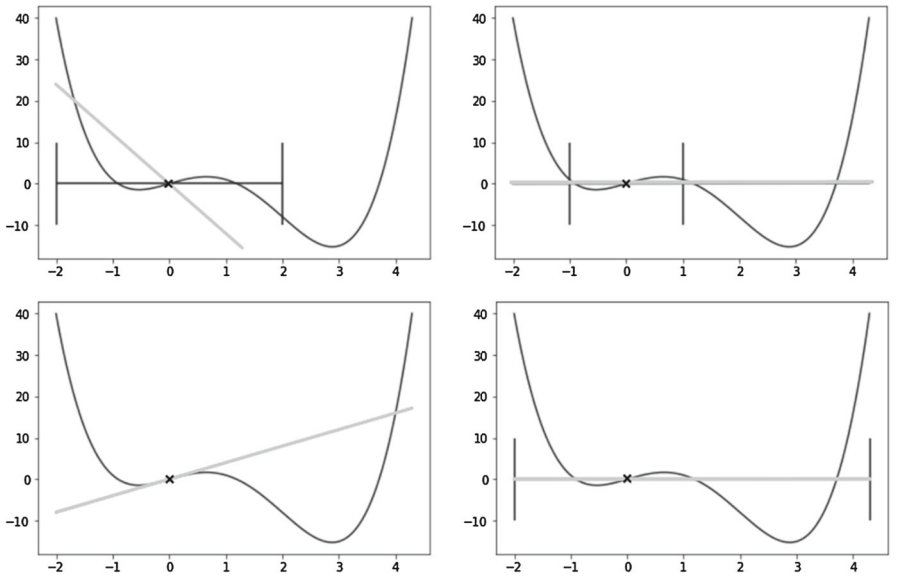

Fig. 4 Unstable linear approximations. The grey line in each panel shows a local approximation of the same function centered at the same location. The varying range is indicated by the black bars, leading to vastly different linear explanations. From (Wachter et al., 2018, p. 885)

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 regularization 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.[18] 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

[18] Assume here, for simplicity, that all predictors have been scaled to unit variance and our distance metric is L_2. Of course, both assumptions are often violated in practice, but these complications do nothing to mitigate the problem; on the contrary, they only make matters worse.

linear approximation itself. Extreme sensitivity to ε on the part of VI scale and location parameters indicates a highly nonlinear neighborhood 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 features grows.

4.2.2 Correlated predictors

Another challenging scenario for IML tools is when predictors are strongly correlated. For instance, 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 and 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 a criminal recidivism dataset that passes fairness audits according to both IML methods.

Severe testing cannot, on its own, prevent bad actors from engaging in discriminatory behavior. 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 et al., 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 and Peters (2020)'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.[19] Severity therefore has a central role to play in holding people and institutions accountable for their algorithmically mediated decisions.

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 harbor 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 sets of Boolean conditions that hold at the target point, selected to ensure some minimal level of precision (i.e., guaranteed with some fixed probability near the input) and optimized for coverage (i.e., designed to apply across a maximally large region of the feature space). Other methods for testing local explanations include the localized knockoff procedure (Gimenez & Zou, 2019), as well as leave-one-covariate-out (LOCO) inference techniques (Lei et al., 2018; Rinaldo et al., 2019). Some recent NeurIPS papers suggest a growing interest in the problem (Schwab & Karlen, 2019; Slack et al., 2021).

[19] For a good review of such methods, see (Heinze-Deml et al., 2018).

These methods are not without their difficulties. They typically require onerous pre- or post-processing, and what few formal guarantees they provide often rely on heuristic reasoning or convenient assumptions. However, they represent a notable advance over the previous state of the art, in that they explicitly try to quantify and optimize 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 rigor required for widespread adoption and user trust.

5 Process versus 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.

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 (Hempel, 1965), the statistical relevance model (Salmon, 1971), the causal mechanical model (Dowe, 2000; Salmon, 1984), and the unificationist model (Kitcher, 1989).[20] 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, 2002) and early moderns (e.g., Berkeley, 1979). 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 Sect. 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 predictions of a target algorithm. Yet even if we allow

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

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 proposed alternative accounts (Achinstein, 1983; Bromberger, 1966; Scriven, 1962), but perhaps no one crystallizes 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 explanations, no single set of parameters to optimize. 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 behavioral 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, Sect. 7), during which agents jointly explore a target system's behavior and anomalies. Building on Moulin et al.'s (2002) tripartite distinction between trace explanations, strategic explanations, and deep explanations

in AI, 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 them as unclear or anomalous through a series of speech acts that gradually bring them closer to understanding the original prediction. Such a personalized process is simply impossible with popular IML algorithms like LIME and SHAP.

5.2 Advantages for IML

There are at least two clear advantages to interactive explanations for IML. First, such approaches are inherently customizable, 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. 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 (Eberhardt, 2010). A preliminary question about some particular observation (e.g., "Why do finch beaks vary so widely across the Galápagos 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 Sect. 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 behavior under real and hypothetical interventions. Just as a knowledgeable scientist should be able to answer a student's questions about a target system, a successful IML algorithm should promote learning and trust among users.

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 et al., 2019; Murdoch et al., 2019). A small group of intrepid computer scientists is actively working to fill the lacuna. The project has come farthest in algorithmic recourse, the IML subdiscipline devoted to advising agents on how to change unfavorable outcomes (e.g., unsuccessful loan applications). Aware that not every feature is within an agent's power to alter, authors have devised various methods for computing counterfactuals that are "actionable" or "feasible" based on user-provided criteria (Karimi et al., 2020; Poyiadzi et al., 2020; Ustun et al., 2019). Despite their promise, these algorithms have not generally been implemented via graphical user interfaces—with the notable exception of Google's simplified method based on empirical sampling (Wexler et al., 2020)—which means that widespread adoption by non-data scientists remains aspirational.

Lakkaraju et al. (2019), specifically motivated by clinical applications for ML, developed a customizable decision set algorithm that allows users to specify features of interest. The explanations provided by Model Understanding through Subspace Explanations (MUSE) 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 and Flach (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 (e.g., the triangle inequality). 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.

6 Conclusion

Feyerabend (1975) famously argues that the ideal structure of scientific discovery is neither a logical sequence of conjectures and refutations (Popper, 1959) 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 represents a scientific ideal, 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 henceforth be conducted. A degree of pluralism is welcome and fruitful in young, dynamic subdisciplines such as this, and indeed, the methodological imperatives enumerated above may occasionally be incompatible. 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, pp. 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 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 that it does not take much to break these methods. Some simple confounding between predictors is typically sufficient. These worries are especially urgent as algorithms expand into ever more sensitive and high-risk areas of public and private life. Data regulation policy notwithstanding, the so-called "right to explanation" will remain not just unrealized 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 relative merits of potentially incompatible explanatory desiderata must be carefully weighed. The resulting algorithms may be 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 favors 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.

Acknowledgements This work was partially funded by ONR Grant N62909-19-1-2096.

References

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

Achinstein, P. (1983). *The nature of explanation*. Oxford University Press.

Akula, R. A., Todorovic, S., Chai, Y. J., & Zhu, S.-C. (2019). Natural language interaction with explainable ai models. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR) workshops*.

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*(10), 1340–1347.

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.

Austin, J. L. (1961). *Philosophical papers* (J. O. Urmson & G. J. Warnock, Eds.). Oxford: Clarendon Press.

Bandyopadhyay, P. S., & Forster, M. R. (Eds.). (2011). *Philosophy of statistics*. Elsevier.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning*. fairmlbook.org.

BarredoArrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1*(3), 385–402.

Berkeley, G. (1979). *Three dialogues between hylas and philonous* (R. M. Adams, Ed.). Indianapolis: Hackett.

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Eckersley, P. (2020). Explainable machine learning in deployment. In: *Proceedings of the conference on fairness, accountability, and transparency* (pp. 648–657).

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Bromberger, S. (1966). Why questions. In R. Colodny (Ed.), *Mind and cosmos: Essays in contemporary science and philosophy.* Pittsburgh: University of Pittsburgh Press.

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12.

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., & Smith, A. (2019). From soft classifiers to hard decisions: How fair can we be? In: *Proceedings of the conference on fairness, accountability, and transparency* (pp. 309–318).

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*, (pp, 1721–1730).

Caruana, R. (1997). Multitask learning. *Machine Learning, 28*(1), 41–75.

Creel, K. A. (2020). Transparency in complex computational systems. *Phil. Sci., 87*(4), 568–589.

Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv* preprint, 2006.11371.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In: *Proceedings of the IEEE symposium on security and privacy* (pp. 598–617).

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). Proxy non-discrimination in data-driven systems. *arXiv* preprint, 1707.08120.
Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. Cambridge University Press.
Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of American Statistical Association, 95*(450), 407–424.
Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science and Medicine, 210*, 2–21.
Doran, G., Muandet, K., Zhang, K., & Schölkopf, B. (2014). A Permutation-based kernel conditional independence test. In: *Proceedings of the 13th conference on uncertainty in artificial intelligence*, (pp. 132–141).
Dowe, P. (2000). *Physical causation*. Cambridge University Press.
Duhem, P. (1954). *The aim and structure of physical theory* (P. W. Wiener, Ed.). Princeton, NJ: Princeton University Press.
Earman, J. (1992). *Bayes or Bust? A critical examination of bayesian confirmation theory*. The MIT Press.
Eberhardt, F. (2010). Causal discovery as a game. In: *Proceedings of NIPS workshop on causality* (pp. 87–96).
Efron, B. (2010). *Large-scale inference: Empirical bayes methods for estimation, testing, and prediction*. Cambridge University Press.
Feyerabend, P. (1975). *Against Method*. New Left Books.
Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*(177), 1–81.
Fisher, R. A. (1925). *Statistical methods for research workers*. Oliver & Boyd.
Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.
Floridi, L. (2011). Semantic information and the correctness theory of truth. *Erkenntnis, 74*(2), 147–175.
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.
Frye, C., Feige, I., & Rowat, C. (2020). Asymmetric Shapley values: Incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, pp. 1229–1239.
Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. *Advances in Neural Information Processing Systems*, pp. 489–496.
Galhotra, S., Pradhan, R., & Salimi, B. (2021). Explaining black-box algorithms using probabilistic contrastive counterfactuals. In: *Proceedings of the international conference on management of data* (pp. 577–590).
Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641–651.
Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology, 66*(1), 8–38.
Gimenez, J. R., & Zou, J. (2019). Discovering conditionally salient features with statistical guarantees. In: *Proceedings of the 36th international conference on machine learning* (pp. 2290–2298).
Goldman, A. (1979). What is justified belief? In G. S. Pappas (Ed.), *Justification and Knowledge* (pp. 1–25). Reidel.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their sesolution with s-values. *American Statistician, 73*(sup1), 106–114.
Gregorutti, B., Michel, B., & Saint-Pierre, P. (2015). Grouped variable importance with random forests and application to multiple functional data analysis. *Computational Statistics & Data Analysis, 90*, 15–35.
Grice, P. (1989). *Studies in the way of words*. Harvard University Press.
Guedj, B. (2019). A Primer on PAC-Bayesian Learning. *arXiv* preprint, 1901.05353.
Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3–12.
Heinze-Deml, C., Peters, J., & Meinshausen, N. (2018). Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, *6*(2), 20170016.
Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. Free Press.
Hintikka, J. (1999). *Inquiry as inquiry: A logic of scientific discovery*. Springer.

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.
Holland, P. W. (1986). Statistics and causal inference. *Journal of American Statistical Association, 81*(396), 945–960.
Hooker, G., & Mentch, L. (2019). Please stop permuting features: An explanation and alternatives. *arXiv* preprint, 1905.03151.
Hume, D. (1739). *A treatise of human nature* (L. A. Selby-Bigge & P. H. Nidditch, Eds.). Oxford: Clarendon Press.
Hume, D. (1748). *An enquiry concerning human understanding*. Oxford: Oxford University Press.
Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124.
Janzing, D., Minorics, L., & Bloebaum, P. (2020). Feature relevance quantification in explainable AI: A causal problem. In: *Proceedings of the 23rd international conference on artificial intelligence and statistics* (pp. 2907–2916).
Jaynes, E. T. (2003). *Probability theory: The logic of science* (G. L. Bretthorst, Ed.). Cambridge: Cambridge University Press.
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589.
Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., & Wu, Z. S. (2019). Eliciting and enforcing subjective individual fairness. *arXiv* preprint, 1905.10660.
Kaptchuk, T. J. (2001). The double-blind, randomized, placebo-controlled trial: Gold standard or golden calf? *Journal of Clinical Epidemiology, 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.
Keiff, L. (2011). Dialogical logic. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 201). Metaphysics Research Lab, Stanford University.
Kitcher, P. (1989). Explanatory unification and the causal structure of the world. In P. Kitcher & W. Salmon (Eds.), *Scientific explanation* (pp. 410–505). University of Minnesota Press.
Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In: *8th Innovations in theoretical computer science conference (ITCS 2017)* (pp. 43.1–43.23).
Mothilal, R. K., Mahajan, D., Tan, C., & Sharma, A. (2021). Towards unifying feature attribution and counterfactual explanations: Different means to the same end. In *Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society* (pp. 652–663).
Krishnan, M. (2020). Against interpretability: A critical examination of the interpretability problem in machine learning. *Philosophy & Technology, 33*(3), 487–502.
Kuhn, T. S. (1970). *The structure of scientific revolutions*. University of Chicago Press.
Kumar, I., Venkatasubramanian, S., Scheidegger, C., & Friedler, S. (2020). Problems with Shapley-value-based explanations as feature importance measures. In: *Proceedings of the 37th international conference on machine learning* (pp. 1–10).
Kyburg, H. (1992). The scope of Bayesian reasoning. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, 2*, 139–152.
Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. In: *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society*, pp. 131–138.
Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?": Manipulating user trust via misleading black box explanations. In:*Proceedings of the AAAI/ACM conference on ai, ethics, and society*, pp. 79–85.
Lehmann, E. L., & Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd ed.). Springer.
Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of American Statistical Association, 113*(523), 1094–1111.
Lewis, D. (1973). *Counterfactuals*. Blackwell.
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.
Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, pp. 4765–4774.

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. *Nature Machine Intelligence, 2*(1), 56–67.

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy in Science, 71*(5), 1007–1025.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science, 57*(2), 323–357.

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

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Mittelstadt, B., Russel, C., & Wachter, S. (2019). Explaining explanations in AI. In: *Proceedings of the conference on fairness, accountability, and transparency*.

Molnar, C. (2019). *Interpretable machine learning: A guide for making black box models interpretable*. Münich: Christoph Molnar.

Moulin, B., Irandoust, H., Bélanger, M., & Desbordes, G. (2002). Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review, 17*(3), 169–222.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071–22080.

Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? *Bioinformatics, 34*(21), 3711–3718.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society 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.

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.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*(3), 441–459.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2018). Challenging the hegemony of randomized controlled trials: A commentary on Deaton and Cartwright. *Social Science and Medicine, 210*, 60–62.

Pearl, J., & Bareinboim, E. (2014). External validity: From do-calculus to transportability across populations. *Statistical Science, 29*(4), 579–595.

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

Peters, M. A. (2018). Deep learning, education and the final stage of automation. *Educational Philosophy and Theory, 50*(6–7), 549–553.

Popper, K. (1959). *The logic of scientific discovery*. Routledge.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020). FACE: Feasible and actionable counterfactual explanations. In: *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 344–350).

Pruthi, D., Gupta, M., Dhingra, B., Neubig, G., & Lipton, Z. C. (2020). Learning to deceive with attention-based explanations. In: *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 4782–4793).

Quine, WvanO. (1960). *Word and Object*. The MIT Press.

Quine, WvanO. (1980). *Methods of logic* (4th ed.). Harvard University Press.

Reichenbach, H. (1956). *The direction of time*. University of California Press.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. *AAAI*, pp. 1527–1535.
Rinaldo, A., Wasserman, L., & G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics, 47*(6), 3438–3469.
Romeijn, J.-W. (2017). Philosophy of statistics. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 201). Metaphysics Research Lab, Stanford University.
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66*(5), 688–701.
Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.
Salmon, W. (1971). Statistical explanation. In W. Salmon (Ed.), *Statistical explanation and statistical relevance* (pp. 29–87). University of Pittsburgh Press.
Salmon, W. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.
Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and algorithms*. MIT Press.
Schölkopf, B., & Smola, A. (2017). *Learning with kernels: Support vector machines, regularization, optimization, and beyond* (2nd ed.). The MIT Press.
Schwab, P., & Karlen, W. (2019). CXPlain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems 32* (pp. 10220–10230).
Scriven, M. (1962). Explanations, predictions, and laws. In H. Feigl & G. Maxwell (Eds.), *Scientific explanation, space, and time* (pp. 170–230). University of Minnesota Press.
Selbst, A., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233–242.
Shah, R., & Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *Annals of Statistics, 48*(3), 1514–1538.
Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.
Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.
Shpitser, I., & Pearl, J. (2008). Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research, 9*, 1941–1979.
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), 1140–1144.
Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods. In: *Proceedings of the AAAI/ACM conference on ai, ethics, and society*, pp. 180–186.
Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021). Reliable post-hoc explanations: Modeling uncertainty in explainability. *Advances in Neural Information Processing Systems, 34*.
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.
Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search* (2nd ed.). 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.
Strawson, P. F. (1964). Intention and convention in speech acts. *Philosophical Review, 73*(4), 439–460.
Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics, 9*(1), 307.
Sullivan, E. (2020). Understanding from machine learning models. *The British Journal for the Philosophy of Science*
Talbott, W. (2016). Bayesian epistemology. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.
Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56.
Ustun, B., Spangher, A., & Liu, Y. (2019). Actionable recourse in linear classification. In: *Proceedings of the conference on fairness, accountability, and transparency* (pp. 10–19).
Valiant, L. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.

van der Laan, M. J., & Rose, S. (Eds.). (2011). *Targeted learning: Causal inference for observational and experimental data*. Springer.
van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.
Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies to their probabilities. *Theory Probab. Appl., 16*(2), 264–280.
Vilone, G., & Longo, L. (2020). Explainable artificial intelligence: A systematic review. *arXiv* preprint, 2006.00093.
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.
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. *International Data Privacy 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 Journal of Law & Technology, 31*(2), 841–887.
Walton, D. (2004). A new dialectical theory of explanation. *Philosophical Explorations, 7*(1), 71–89.
Walton, D. (2006). Examination dialogue: An argumentation framework for critically questioning an expert opinion. *Journal of Pragmatics, 38*(5), 745–777.
Walton, D. (2011). A dialogue system specification for explanation. *Synthese, 182*(3), 349–374.
Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *American Statistician, 70*(2), 129–133.
Watson, D., Gultchin, L., Taly, A., & Floridi, L. (2021). Local explanations via necessity and sufficiency: Unifying theory and practice. In:*Proceedings of the 37th conference on uncertainty in artificial intelligence*.
Watson, D. S., & Floridi, L. (2020). The explanation game: a formal framework for interpretable machine learning. *Synthese, 198*(10), 9211–9242.
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. *BMJ, 364*, 446–448.
Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics, 26*(1), 56–65.
Wittgenstein, L. (1953). *Philosophical investigations* (R. Rhees & G. E. M. Anscombe, Eds.; G. E. M. Anscombe, Trans.). Oxford: Blackwell.
Woodward, J. (2019). Scientific explanation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.
Worrall, J. (2007). Why there's no cause to randomize. *The British Journal for the Philosophy of Science, 58*(3), 451–488.
Zednik, C. (2019). Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology, 34*(1), 265–288.
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. In: *Proceedings of the 27th conference on uncertainty in artificial intelligence*, pp. 804–813.
Zhao, Q., & Hastie, T. (2019). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 272–281.
Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.