The Explanation Game:
A Formal Framework for Interpretable Machine Learning

David S. Watson[1], Luciano Floridi[1,2]

[1] Oxford Internet Institute, University of Oxford, 1 St. Giles', Oxford OX1 3LW, UK
[2] The Alan Turing Institute, 96 Euston Road, Kings Cross, London NW1 2DB, UK
Email for correspondence: david.watson@oii.ox.ac.uk

§0. Abstract

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

Keywords

Algorithmic explainability; Explanation game; Interpretable machine learning; Pareto frontier; Relevance.

§1. Introduction

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

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

Explainability is fast becoming a top priority in statistical research, where it is often abbreviated as xAI (explainable Artificial Intelligence) or iML (interpretable Machine Learning). We adopt the latter initialism here to emphasise our focus on supervised learning algorithms (formally defined in §3.1) as opposed to other, more generic artificial intelligence applications.

Several commentators have argued that the central aim of iML is underspecified (Doshi-Velez & Kim, 2017; Lipton, 2016). They raise concerns about the irreducible subjectivity of explanatory success, a concept that they argue is poorly defined and difficult or impossible to measure. In this article, we tackle this problem head on. We provide a formal framework for conceptualising the goals and constraints of iML systems by designing an idealised *explanation game*. Our model clarifies the trade-offs inherent in any iML solution, and characterises the conditions under which epistemic agents are almost surely guaranteed to converge on an optimal set of explanations in polynomial time. The game serves a descriptive and a normative function, establishing a conceptual space in which to analyse and compare existing proposals, as well as design new and improved solutions.

The remainder of this paper is structured as follows. In §2, we identify three distinct goals of iML. In §3, we review relevant background material. We clarify the scope of our proposal in §4. In §5, we articulate the rules of the explanation game and outline the procedure in pseudocode. A discussion follows in §6. We consider four objections in §7, before concluding in §8.

§2. Why Explain Algorithms?

We highlight three goals that guide those working in iML: to *audit*, to *validate*, and to *discover*. These objectives help motivate and focus the discussion, providing an intuitive typology for the sorts of explanations we are likely to seek and value in this context. Counterarguments to the project of iML are delayed until §7.

§2.1 Justice as (Algorithmic) Fairness

Perhaps the most popular reason to explain algorithms is their large and growing social impact. ML has been used to help evaluate loan applications (Khandani, Kim, & Lo, 2010) and student admissions (Kotsiantis, 2012), predict criminal recidivism (Dressel & Farid, 2018), and identify military targets (Nasrabadi, 2014), to name just a few controversial examples. Failure to properly screen training datasets for biased inputs threatens to automate injustices already present in society (Mittelstadt et al., 2016). For instance, studies have indicated that algorithmic profiling consistently shows online advertisements for higher paying jobs to men over women (Datta et al., 2015); that facial recognition software is often trained on predominantly white subjects, making them inaccurate classifiers for black and brown faces (Buolamwini & Gebru, 2018); and that predatory lenders use financial data to disproportionately target poor communities (Eubanks, 2018). Critics point to these failures and argue that there is a dearth of fairness, accountability, and transparency in ML – collectively acronymised as FAT ML, an annual conference on the subject that began meeting in 2014.

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

§2.2 The Context of (Algorithmic) Justification

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

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

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

Advances in iML represent a statistical answer to the reliabilist challenge, enabling sceptics to analyse the internal behaviour of a model when deliberating on particular predictions. This is the goal, for instance, of all local linear approximation techniques, including popular iML algorithms like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), which assign weights to input variables so users can verify that the model has not improperly focused on uninformative features like the aforementioned watermark. These methods will be examined more closely in §6.

§2.3 The Context of (Algorithmic) Discovery

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

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

The case is not purely hypothetical. A wave of research in the early 2000s established a connection between transcriptomic signatures and clinical outcomes for breast cancer patients (e.g., Sørlie et al., 2001; van 't Veer et al., 2002; van de Vijver et al., 2002). The studies employed a number of sophisticated statistical techniques, including unsupervised clustering and survival analysis. Researchers found, among other things, a strong association between BRCA1 mutations and basal-like breast cancer, an especially aggressive form of the disease. Genomic analysis remains one of the most active and promising areas of research in the natural sciences, and whole new subfields of ML have emerged to tackle the unique challenges presented by these high-dimensional datasets (Bühlmann et al., 2016; Bühlmann & van de Geer, 2011). Successful iML strategies will be crucial to realising the promise of high-throughput sciences.

§3. Formal Background

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

§3.1 Supervised Learning

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

Each observation in a training dataset consists of a pair $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$, where $\boldsymbol{x}_i$ denotes a point in p-dimensional space, $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$, and y_i represents the corresponding outcome. We assume that samples are independently and identically distributed according to some fixed but unknown joint probability distribution $\mathbb{P}(\mathbf{Z}) = \mathbb{P}(\mathbf{X}, Y)$. Using n observations, an algorithm maps a dataset to a function, a: $\mathbf{Z} \rightarrow f$; the function in turn maps features to outcomes, f: $\mathbf{X} \rightarrow$

Y. We consider both cases where Y is categorical (in which case f is a classifier) and where Y is continuous (in which case f is a regressor). We make no additional assumptions about the structure or properties of f.

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

$$R(f, \mathbf{Z}) = \mathbb{E}[L(f, \mathbf{Z})] \quad (1)$$

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

$$R_{\text{emp}}(f, \mathbf{Z}) = \frac{1}{n} \sum_i L(f, \mathbf{z}_i) \quad (2)$$

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

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

§3.2 Causal Interventionism

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

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

Under certain common assumptions,[2] Pearl's *do*-calculus provides a complete set of formal tools for reasoning about causal interventions (Huang & Valtorta, 2006). A key element of Pearl's notation system is the *do* operator, which allows us to denote, for example, the probability of Y, conditional on an intervention that sets variable X to value x, with the concise formula $\mathbb{P}(Y|do(X = x))$. A structural model M is a tuple $\langle U, V, F\rangle$ consisting of exogenous variables U, endogenous variables V, and a set of functions F that map each V_j's causal antecedents to its observed values. M may be visually depicted as a graph with nodes corresponding to variables and directed edges denoting causal relations between endogenous features (see Fig. 1). We restrict our attention here to directed acyclic graphs (DAGs), which are the focus of most work in causal interventionism.

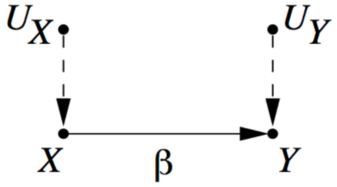

Figure 1. A simple example of a causal graph with two exogenous variables, U_X and U_Y, with unobserved causal effects on two endogenous variables, X and Y, respectively. The causal influence of X on Y is given by a coefficient β.

If the model M contains no exogenous variables, then M is said to be *Markovian*. In this case, factorisation of a graph's joint distribution is straightforward and causal effects can be computed directly from the data. However, when unobserved variables influence observed variables, as in Fig. 1, then we say that M is *semi-Markovian*, and more elaborate methods are needed to estimate causal effects. Specifically, some sort of adjustment must be made by conditioning on an appropriate set of covariates. While several overlapping formulations have been proposed for such adjustments (Galles & Pearl, 1995; Pearl, 1995; Robins, 1997), we follow Tian & Pearl (2002), who provide a provably sound and complete set of causal identifiability conditions for semi-Markovian models (Huang & Valtorta, 2008).

Their criteria are as follows. The causal effect of the endogenous variable X_j on all remaining covariates is identifiable if and only if there is no consecutive sequence of confounding edges between X_j and X_j's immediate successors in the graph. Weaker conditions are sufficient when we focus on a proper subset S of V. In this case, $\mathbb{P}(S|do(X_j = x_{ij}))$ is identifiable so long as there is no consecutive sequence of confounding edges between X_j and X_j's children in the subgraph composed of the ancestors of S.

We take it that the goal in most iML applications is to provide a causal explanation for one or more algorithmic predictions. Identifiability is therefore a central concern, and another

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

key component to defining convergence conditions in §5.3. Fortunately, as we argue in §4.1, many cases of interest in this setting involve Markovian graphs, and therefore need no covariate adjustments. Semi-Markovian alternatives are considered in §5.2.2, although guarantees cannot generally be provided in such instances without additional assumptions.

If successful, a causal explanation for some algorithmic prediction(s) will accurately answer a range of what Woodward calls "what-if-things-had-been-different questions" (henceforth w-questions). For instance, we may want to know what feature(s) about an individual caused her loan application to be denied. What if she had been wealthier? Or older? Would a hypothetical applicant identical to the original except along the axis of wealth or age have had more luck? Several authors in the iML literature explicitly endorse such a counterfactual strategy (Kusner et al., 2017; Wachter et al., 2018). We revisit these methods in §6.

§3.3 Decision Theory

Decision theory provides formal tools for reasoning about choices under uncertainty. These will prove useful when attempting to quantify explanatory relevance in §5.2.3. We assume the typical setup, in which an individual considers a finite set of actions A and a finite set of outcomes C. According to expected utility theory,[3] an agent's rational preferences may be expressed as a utility function u that maps the Cartesian product of A and C to the real numbers, $u: A \times C \rightarrow \mathbb{R}$. For instance, Jones may be unsure whether to pack his umbrella today. He could do so (a_1), but it would add considerable bulk and weight to his bag; or he could leave it at home (a_2) and risk getting wet. The resulting utility matrix is depicted in Table 1.

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

Table 1. Utility matrix for John when deciding whether or not to pack his umbrella.

The rational choice for Jones depends not just on his utility function u but also on his beliefs about whether or not it will rain. These are formally expressed by a (subjective) probability distribution over C, $\mathbb{P}(C)$. We compute each action's expected utility by taking a weighted average over outcomes:

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

$$\mathbb{E}[u(a_i, C)|E] = \sum_j \mathbb{P}(c_j|E)u(a_i, c_j) \quad (3)$$

where the set of evidence E is either empty (in which case Eq. 3 denotes a prior expectation) or contains some relevant evidence (in which case Eq. 3 represents a posterior expectation). Posterior probabilities are calculated in accordance with Bayes's theorem:

$$\mathbb{P}(c_i|E) = \frac{\mathbb{P}(E|c_i)\mathbb{P}(c_i)}{\mathbb{P}(E)} \quad (4)$$

which follows directly from the Kolmogorov axioms for the probability calculus (Kolmogorov, 1950). By solving Eq. 3 for each element in A, we identify at least one utility-maximising action:

$$a^* = \underset{a_i \in A}{\operatorname{argmax}} \mathbb{E}[u(a_i, C)|E] \quad (5)$$

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

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

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

and

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

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

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

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

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

Of course, humans can be notoriously irrational. Experiments in psychology and behavioural economics have shown time and again that people rely on heuristics and cognitive biases instead of consistently applying the axioms of decision theory or probability calculus (Kahneman, 2011). Thus, the concepts and principles we outline here are primarily normative. They prescribe an optimal course of behaviour, a sort of Kantian regulative ideal when utilities and probabilities are precise, and posterior distributions are properly calculated. For the practical purposes of iML, these values may be estimated via a hybrid system in which software aids an inquisitive individual with bounded rationality. We revisit these issues in §7.1.

§4. Scope

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

Strevens (2013) differentiates between three modes of understanding: *that*, *why*, and *with*.[4] Understanding *that* some proposition *p* is true is simply to be aware that *p*. Understanding *why p* is true requires some causal explanation for *p*. Strevens's third kind of understanding, however, applies only to theories or models. Understanding *with* a model amounts to knowing how to apply it in order to predict or explain real or potential phenomena. For instance, a physicist who uses Newtonian mechanics to explain the motion of billiard balls thereby demonstrates her ability to understand *with* the theory. Since this model is strictly speaking false, it would be incorrect to say that her explanation provides a true understanding of *why* the billiard balls move as they do. (Of course, she could be forgiven for sparing her poolhall companions the relativistic details of metric tensors and spacetime curvature in this case.) Yet our physicist has clearly understood something – namely the Newtonian theory itself – even if the classical account she offers is inaccurate or incomplete. Similarly, the goal in iML is to help epistemic agents understand *with* the target model *f*, independent of whatever realities *f* was intended to capture. The situation is slightly more complicated in the case of discovery (§2.3). The strategy here is to use understanding *with* as an indirect path to understanding *why*, on the assumption that if model *f* performs well then it has probably learned some valuable information about the target system.

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

§4.1 Complete

Model *f* is complete with respect to the input features $\mathbf{X}$ in the sense that exogenous variables have no influence whatsoever on predicted outcomes $\hat{Y}$. Whereas nature is full of unobserved confounders that may complicate or undermine even a well-designed study, fitted models are self-contained systems impervious to external variation. They therefore instantiate Markovian, rather than semi-Markovian graphs. This is true even if dependencies between predictors are not explicitly modelled, in which case we may depict *f* as a simple DAG with directed edges from each feature $X_1, \ldots, X_p$ to $\hat{Y}$.

In what follows, we presume that the agents in question know which variables were used to train *f*. This may not always be the case in practice, and without such knowledge it

[4] In what follows, we take it more or less for granted that explanations promote understanding and that understanding requires explanations. Both claims have been disputed. For a discussion, see (de Regt et al., 2009; Grimm, 2006; Khalifa, 2012). We revisit the relationship between these concepts in §7.2.

becomes considerably more difficult to explain algorithmic predictions. Whatever the epistemic status of the inquiring agent(s), however, the underlying model itself remains complete.

Issues arise when endogenous variables serve as proxies for exogenous variables. For instance, a model may not explicitly include a protected attribute such as race, but instead use a seemingly innocuous covariate like zip code, which is often a strong predictor of race (Datta et al., 2017). In this case, an intervention that changes a subject's race will have no impact on model *f*'s predictions unless we take the additional step of embedding *f* in a larger causal structure *M* that includes a directed edge from race to zip code. We consider possible strategies for resolving problems of this nature in §5.2.2.

§4.2 Precise

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

Woodward (2003; 2010) emphasises the role of "stability" in causal generalisations, a concept that resembles what we call precision. The difference is that stability in Woodward's sense can only be applied to a proper subset of the edges (usually just a single edge) in a causal graph. The generalisation that "variable X causes variable Y" is *stable* to the extent that it persists across a wide range of background conditions, i.e. alternative states of the model M. Precision in our sense requires completeness, because it applies only to the causal relationship between the set of all predictors $\mathbf{X}$ and the outcome Y, which is strictly deterministic at the token level.

§4.3 Forthcoming

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

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

§5. The Explanation Game

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

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

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

- Alice is a rational agent. Her preferences over alternatives are complete and transitive, she integrates new evidence through Bayesian updating, and she does her best to maximise expected utility subject to constraints on her cognitive resources.
- Bob is Alice's accomplice. He has data on the features $\mathbf{V} = (X_1, \ldots, X_p)$ that are endogenous to *f*, as well a (possibly empty) set of exogenous variables $\mathbf{U} = (X_{p+1}, \ldots,$

X_{p+m}) that are of potential interest to Alice. He may query f with any well-formed input at little or no cost.

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

§5.1 Three Desiderata

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

(i) The generalisations described by M are accurate, or at least approximately so, as are the observations $Y = y_i$ and $X = x_j$.

(ii) According to M, $Y = y_i$ under an intervention that sets X to x_j.

(iii) There exists some possible intervention that sets X to x_k (where $x_j \neq x_k$), with M correctly describing the value y_l (where $y_i \neq y_l$) that Y would assume under the intervention.

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

§5.1.1 Accuracy

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

Problem solved? Not quite. Bob's causal graph M is every bit as opaque as the underlying model f. In fact, the two are identical. So while this explanation may be maximally accurate, it is far too complex to be of any use to Alice. The result is not unlike the map of

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

Borges's famous short story (1946), in which imperial cartographers aspire to such exactitude that they draw their territory on a 1:1 scale. Black box explanations of this sort create a kind of Chinese room (Searle, 1980), in which the inquiring agent is expected to manually perform the algorithm's computations in order to trace the path from input to output. Just as the protagonist of Searle's thought experiment has no understanding of the Chinese characters he successfully manipulates, so Alice gains no explanatory knowledge about f by instantiating the model herself. Unless she is comfortable computing high-dimensional tensor products on the fly, Alice cannot use M to build a mental model of the target system f or its behaviour near $\boldsymbol{x}_i$. She cannot answer relevant w-questions without consulting the program, which will merely provide her with new labels that are as unexplained as the original.

§5.1.2 Simplicity

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

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

Rather than adjudicate between the various competing notions of simplicity that abound in the literature, we opt for a purely relational approach upon which simplicity is just equated with *intelligibility for Alice*. We are unconvinced that there is any sense to be made of an absolute, mind-independent notion of simplicity. Yet even if there is, it would be of little use to Alice if we insist that explanation g_1 is simpler than g_2 on our preferred definition of the term, despite the empirical evidence that she understands the implications of the latter better than the former. What is simple for some agents may be complex for others, depending on background knowledge and contextual factors. In §5.2, we operationalise this observation by measuring simplicity in explicitly agentive terms.

§5.1.3 Relevance

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

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

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

§5.2 Rules of the Game

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

```
Inputs:
Environment: supervised learner $f$, endogenous variables $\mathbf{V} = (X_1, \ldots, X_p)$, data $D \sim \mathbb{P}(M)$ possibly including exogenous covariates $\mathbf{U} = (X_{p+1}, \ldots, X_{p+m})$
Alice: explanandum $f(\boldsymbol{x}_i) = \hat{y}_i$, contrastive outcome $f(\boldsymbol{x}_i) \neq \tilde{y}_i$, level of abstraction LoA, choice set $A$, causal hypotheses $C$, utility function $u$, prior distribution over causal hypotheses $\mathbb{P}(C)$, function space $\mathcal{H}$
Bob: function space $\mathcal{G}$, kernel $k$, loss function $L$
for each round:
    (1) Bob creates a mapping $\psi: \mathcal{Z}_f \rightarrow \mathcal{Z}_g$ from the original $f$-space to an explanatory $g$-space designed to (a) shift the input distribution to Alice's desired LoA and (b) help provide evidence for or against at least one hypothesis in $C$. Whereas $\mathbf{Z}_f = (\mathbf{X}, \hat{Y})$, $\mathbf{Z}_g = (\mathbf{X}', Y')$.
    if $\mathbf{X}'$ includes variables $\mathbf{U}$ that are exogenous to $f$:
        (2) Bob trains an empirical risk minimising (ERM) model $g' \in \mathcal{G}$ such that $g': \mathbf{V} \rightarrow \mathbf{U}$, optionally fit using kernel $k$.
    end if
    (3) Bob creates a training dataset by repeatedly querying $f$ (and $g'$, if exogenous variables are required). The resulting data are mapped to $g$-space via $\psi$.
    for each map $\psi$:
        (4) Bob trains an ERM model $g \in \mathcal{G}$ such that $g: \mathbf{X}' \rightarrow Y'$, optionally fit using kernel $k$. Performance is evaluated with the loss function $L$, optionally weighted by $k$, estimated in $f$-space via the inverse mapping $\psi^{-1}$.
        for each model $g$:
            (5) Alice asks $w$-questions of the form $\mathbb{E}[Y'|do(X_j' = x_{ij}')] = ?$ In other words, she queries the effects of interventions in $g$-space. Bob reports both the predicted outcome and the empirical risk.
            (6) By repeating step (5) some large number of times $q$, Alice builds an ERM model $h \in \mathcal{H}$ such that $h: \mathbf{X}' \rightarrow Y'$. Performance is evaluated with the loss function $L$, optionally weighted by $k$, estimated in $g$-space.
            (7) The information Alice learns from and about $g$ and $h$ constitutes a body of evidence $E$, which she uses to update her beliefs regarding $C$.
            (8) Alice calculates the posterior expected utility of each action in $A$, producing at least one optimal choice $a^*$.
        end for
    end for
end for
Outputs: $R_{\text{emp}}(g, \mathbf{Z}_f)$, $R_{\text{emp}}(h, \mathbf{Z}_g)$, $\mathbb{E}[u(a^*, C)|E]$
```

This game has a lot of moving parts, but at its core the process is quite straightforward. Essentially, Bob does his best to proffer an accurate explanation in terms that Alice can understand. She learns by asking w-questions until she feels confident enough to answer such questions herself. The result is scored by three measures: accuracy (error of Bob's model), simplicity (error of Alice's model), and relevance (expected utility for Alice). All inputs and steps are discussed in greater detail below.

§5.2.1 Inputs

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

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

Pragmatism demands that Alice have some notion why she is playing this game. Her choices A, preferences u, and beliefs $\mathbb{P}(C)$ will guide Bob in his effort to supply a satisfactory explanation and constrain the set of possible solutions. The MNIST example is a case of iML for validation (§2.2), in which Alice's choice set may include the option to deploy or not deploy the model f. Her degrees of belief with respect to various causal hypotheses are determined by her expertise in the data and model. Perhaps it is well known that algorithms struggle to

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

differentiate between “7” and “1” when the former appears without a horizontal line through the digit. The cost of such a mistake is factored into her utility function.

Bob, for his part, enters into the game with two key components: a kernel function k and a loss function L. The kernel tunes the locality of the explanation, weighting observations by their distance from the original input $\boldsymbol{x}_i$, as measured by some appropriate metric. Whether the kernel is used to train the model g or simply evaluate g’s empirical risk is left up to Bob. Abandoning the kernel altogether results in a global explanation, with no particular emphasis on the neighbourhood of $\boldsymbol{x}_i$. The choice of loss function is likewise left unspecified.

§5.2.2 Mapping the Space

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

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

In any event, the goal at this stage is to make the input features sufficiently intelligible to Alice that they can accommodate her likely w-questions and inform her beliefs about causal hypotheses C. Some general purpose methods for causal feature learning have been proposed (Chalupka, Eberhardt, & Perona, 2017), however, critics have persuasively argued that such procedures cannot be implemented in a context-independent manner (Kinney, 2018). Some areas of research, such as bioinformatics and computer vision, have well-established conventions on how to coarsen high-dimensional feature spaces. Other domains may prove more challenging. Accessibility to external data on exogenous variables of interest will likewise vary from case to case. Even when such datasets are readily available, there is no guarantee that the functional relationships sought can be estimated with high accuracy or precision. As in any other explanatory context, Alice and Bob must do the best they can with their available resources and knowledge.

§5.2.3 Building Models, Scoring Explanations

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

Bob begins by selecting what he judges to be an appropriate algorithm for the task at hand, choosing from the preselected function space $\mathcal{G}$. Popular options include linear models and rule lists. Alice explores g-space by asking a number of w-questions that posit relevant interventions. For instance, she may want to know if the presence of a horizontal line through the middle of a numeral determines whether f predicts a "7". If so, then this will be a hypothesis in C and we should find a corresponding variable in $\mathbf{X}'$. Because we leave open the possibility that the target model f and/or Bob's explanation g may involve implicit or explicit structural equations, we use the *do*-calculus to formalise such interventions.

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

We may now locate explanations generated by this game in three-dimensional space, with axes corresponding to accuracy, simplicity, and relevance. An explanation is deemed *satisfactory* if it does not exceed preselected values of $\varepsilon_1, \varepsilon_2$, and ε_3. These parameters can be interpreted as budgetary constraints on Alice and Bob. How much inaccuracy, complexity, and irrelevance can they afford? We assign equal weight to all three criteria here, but relative costs could easily be quantified through a differential weighting scheme. Together, these points define the extremum of a cuboid, whose opposite diagonal is the origin (see Fig. 2). Any point falling within this cuboid is $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$-satisfactory.

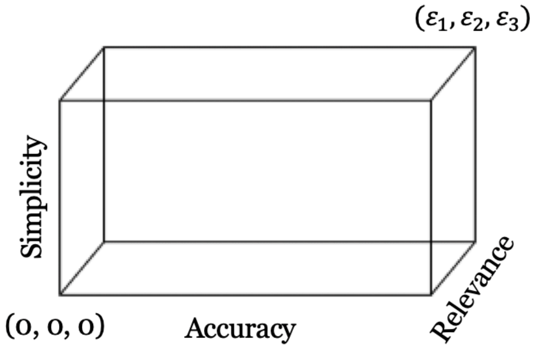

Figure 2. The space of satisfactory explanations is delimited by upper bounds on the error (ε_1), complexity (ε_2), and irrelevance (ε_3) of explanations Alice is willing to accept.

§5.3 Consistency and Convergence

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

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

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

- *Condition 1.* The function spaces $\mathcal{G}$ and $\mathcal{H}$ are of finite VC dimension.
- *Condition 2.* The causal effect of any g-space intervention is uniquely identifiable.
- *Condition 3.* Alice is a rational agent and consistent Bayesian updater.

Condition (1) entails the statistical consistency of Bob's model g, which ensures that accuracy is reliably measured as sample size grows. Conditions (1) and (2) entail the consistency of Alice's model h, which ensures that simplicity is reliably measured as sample size grows. Condition (3) entails the existence of at least one utility-maximising action $a^* \in A$ with well-defined posterior expectation $\mathbb{E}[u(a^*, C)|E]$. If her probabilities are well-calibrated, then Alice will tend to pick the "right" action, or at least an action with no superior alternative in A. With these conditions in place, each round of the game will result in an explanation that cannot be improved upon without altering the input parameters.

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

§6. Discussion

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

The most popular iML methods in use today are local linear approximators like LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017). These methods explain predictions by randomly sampling around the point of interest. Observations are weighted by their distance from the target point, and a regularised linear model is fit to the simulated data by weighted least squares. The final result is a (possibly sparse) set of coefficients indicating the positive or negative association between input features and the response, at least near $\boldsymbol{x}_i$ and conditional on the covariates.

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

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

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

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

Rule lists and GANs have some clear advantages over linear approximators like LIME and SHAP. However, no method in use today explicitly accounts for user interests, an omission that may lead to problematic outcomes. In short, they do not pass the eclipsing test. Recall the case of the (bad) bank in §5.1.3. Suppose that Alice's choice set contains just two options, A = {Sue, Don't sue}, and she considers two causal hypotheses as potential explanations for her denied loan, C = {Wealth, Race}. Alice's utility matrix is given in Table 2.

	c_1: Wealth	c_2: Race
a_1: Sue	−1	5
a_2: Don't sue	0	0

Table 2. Utility matrix for Alice in the (bad) bank scenario.

Alice assigns a uniform prior over C to begin with, such that $\mathbb{P}(c_1) = \mathbb{P}(c_2) = 0.5$. She receives two explanations from Bob: g_1, according to which Alice's application was denied due to her wealth; and g_2, according to which Alice's application was denied due to her race. Using misclassification rate as our loss function and assuming a uniform density over the dichotomous features Wealth ∈ {Rich, Poor} and Race ∈ {White, Black}, we find that both explanations are equally accurate:

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

and equally simple:

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

However, they induce decidedly different posteriors over C:

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

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

The posterior expected utility of a_1 under g_1 is therefore

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

whereas under g_2 the expectation is

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

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

Methods that fail to pass the eclipsing test pose problems for all three iML goals outlined in §2. Irrelevant explanations can undermine tests of validity or quests of discovery by failing to recognise the epistemological purpose that motivated the question in the first place. When those explanations are accurate and simple, Alice can easily be fooled into thinking she has learned some valuable information. In fact, Bob has merely overfit the data. Matters are even worse when we seek to audit algorithms. In this case, eclipsing explanations may actually offer loopholes to bad actors wishing to avoid controversy over questionable decisions. For instance, a myopic focus on accuracy and simplicity would allow (bad) banks to get away with racist loan policies so long as black applicants are found wanting along some other axis of variation.

§7. Objections

In this section, we consider four objections of increasing generality. The first two are levelled against our proposed game, the latter two against the entire iML project.

§7.1 Too Highly Idealised

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

Our response is twofold. First, we remind the sceptical reader that idealisations are a popular and fruitful tool in conceptual analysis. There are no frictionless planes or infinite populations, but such assumptions have contributed to successful theories in physics and genetics. Moreover, it is not unreasonable to assume that an agent would seek an algorithmic explanation with a counterfactual outcome and choice set to hand. (Alice may enter into the game expressly because she suspects her loan application was denied due to her race.) Values for LoA, C, $\mathbb{P}(C)$, and u will likely be less specific, but even vague or incomplete starting points are enough to begin the game. Alice and Bob can iteratively refine their inputs as the rounds pass. Something like this process is how a great deal of research is in fact conducted.

Second, and perhaps more importantly, we stress that Alice and Bob are generalised agents that can and often will be implemented by hybrid systems involving numerous humans and machines working in concert. There is no reason to artificially restrict the cognitive resources of either to that of any specific individual. The problems iML is designed to tackle are beyond the remit of any single person, especially one operating without the assistance of statistical software. When we broaden the cognitive scope of Alice and Bob, the idealisations demanded of them become decidedly more plausible. The explanation game is an exercise in sociotechnical epistemology, where knowledge emerges from the continuous interaction of individuals, groups, and technology (Watson & Floridi, 2018). The essential point is whether the explanation game we have designed is possible and fruitful, not whether a specific Alice and a specific Bob can actually play it according to their idiosyncratic abilities.

§7.2 Pragmatism + Pluralism = Relativist Anarchy?

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

The short answer is no. The objection assumes that for any given fact or event there exists some uniquely satisfactory, mind- and context-independent explanation, presumably in terms of fundamental physical units and laws. Call this view explanatory monism. It amounts to a metaphysical doctrine whose merits or shortcomings are frankly beside the point. For even if the "true" explanation were always available, it would not in general be of much use. The goal of the explanation game is to promote greater *understanding for Alice*. This may come in many forms. For instance, the predictions of image classifiers are often explained by heatmaps highlighting the pixels that most contributed to the given output. The fact that complex mathematical formulae could in this case provide a maximally deep and stable explanation is irrelevant (see §5.1.1). Pragmatic goals require pragmatic strategies. Because iML is fundamentally about getting humans to understand the behaviour of machines, there is a growing call for personalised solutions (Páez, 2019). We take this pragmatic turn seriously and propose formal methods to implement it.

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

§7.3 No Trade-Off

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

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

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

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

§7.4 Double Standards

Zerilli et al. (2018) argue that proponents of iML place an unreasonable burden on algorithms by demanding that they not only perform better and faster than humans, but explain why they do so as well. They point out that human decision-making is far from transparent, and that people are notoriously bad at justifying their actions. Why the double standard? We already have systems in place for accrediting human decision-makers in positions of authority (e.g., judges and doctors) based on their demonstrated track record of performance. Why should we expect anything more from automated decision-makers? The authors conclude that requiring intelligibility of high-performing algorithms is not just unreasonable but potentially harmful if it hinders the implementation of models that could improve services for end users.

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

We argue that the potential benefits of successful iML strategies are more varied and numerous than Kleinberg et al. acknowledge. To reiterate the motivations listed in §2, we see three areas of particular promise. In the case of algorithmic auditing, iML can help ensure the fair, accountable, and transparent application of complex statistical models in high-stakes applications like criminal justice and healthcare. In the case of validation, iML can be used to test algorithms before and during deployment to ensure that models are performing properly and not overfitting to uninformative patterns in the training data. In the case of discovery, iML can reveal heretofore unknown mechanisms in complex target systems, suggesting new theories and hypotheses for testing. Of course, there is no guarantee that such methods will work in every instance – iML is no panacea – but it would foolish not to try. The double standard that Zerilli et al. caution against is in fact a welcome opportunity.

§8. Conclusion

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

There is less reason to believe that algorithms will become any more transparent or intelligible, at least not without the explicit and sustained effort of dedicated researchers in the burgeoning field of iML. We have argued that there are good reasons to value algorithmic interpretability on ethical, epistemological, and scientific grounds. We have outlined a formal framework in which agents can collaborate to explain the outputs of any supervised learner.

The explanation game serves both a descriptive function – providing a common language in which to compare iML proposals – and a normative function – highlighting aspects that are underexplored in the current literature and pointing the way to new and improved solutions.

Future work will relax the assumptions upon which this beta version of the game is based. Of special interest are adversarial alternatives in which Carol and Bob compete to find superior explanations that serve opposing interests, from which Alice may choose. Other promising directions include implementing semi-automated explanation games with greedy algorithms that take turns maximising one explanatory desideratum at a time until convergence. Similar proposals have already been implemented for optimising mixed objectives in algorithmic fairness (Kearns et al., 2018), but we are unaware of any similar work in explainability. Finally, we intend to expand our scope to unsupervised learning algorithms, which pose a number of altogether different explanatory challenges.

The problems we highlight here will be with us for the foreseeable future. We hope that explanation games like the one proposed in this article will help users maximise the benefits and minimise the risks of machine learning algorithms.

References

Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., & Rudin, C. (2018). Learning Certifiably Optimal Rule Lists for Categorical Data. *Journal of Machine Learning Research*, *18*(234), 1–78.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*.

Baker, A. (2016). Simplicity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 201). Metaphysics Research Lab, Stanford University.

Barocas, S., & Selbst, A. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(1), 671–729.

Bell, R. M., & Koren, Y. (2007). Lessons from the Netflix Prize Challenge. *SIGKDD Explor. Newsl.*, *9*(2), 75–79.

Boca, P. P., Bowen, J. P., & Siddiqi, J. I. (2010). *Formal methods: State of the art and new directions*. London: Springer.

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. New York: Oxford University Press.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *Statist. Sci.*, *16*(3), 199–231.

Bühlmann, P., Drineas, P., Kane, M., & van der Laan, M. (Eds.). (2016). *Handbook of Big Data*. Boca Raton, FL: Chapman and Hall/CRC.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer-Verlag.

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91).

Cartwright, N. (2002). Against Modularity, the Causal Markov Condition, and Any Link between the Two: Comments on Hausman and Woodward. *The British Journal for the Philosophy of Science*, *53*(3), 411–453.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge: Cambridge University Press.

Chalupka, K., Eberhardt, F., & Perona, P. (2017). Causal feature learning: an overview. *Behaviormetrika*, *44*(1), 137–164.

Corfield, D., Schölkopf, B., & Vapnik, V. (2009). Falsificationism and statistical learning theory: Comparing the Popper and Vapnik-Chervonenkis dimensions. *Journal for General Philosophy of Science*, *40*(1), 51–58.

Datta, Amit, Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, (1), 92–112.

Datta, Anupam, Fredrikson, M., Ko, G., Mardziel, P., & Sen, S. (2017). *Proxy Non-Discrimination in Data-Driven Systems*.

de Regt, H. W., Leonelli, S., & Eigner, K. (Eds.). (2009). *Scientific understanding: Philosophical perspectives*. Pittsburgh: University of Pittsburgh Press.

Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv* preprint, 1702.08608.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580.

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a "right to explanation" is probably not the remedy you are looking for. *Duke Law and Technology Review*, *16*(1), 18–84.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*, 115.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's Press.

Fisher, A., Rudin, C., & Dominici, F. (2018). All Models are Wrong but many are Useful: Variable Importance for Black-Box, Proprietary, or Misspecified Prediction Models, using Model Class Reliance. *arXiv* preprint, 1801.01489.

Floridi, L. (2004). On the Logical Unsolvability of the Gettier Problem. *Synthese*, *142*(1), 61–79.

Floridi, L. (2017). The logic of design as a conceptual logic of information. *Minds and Machines*, *27*(3), 495–519.

Floridi, Luciano. (2008a). The method of levels of abstraction. *Minds and Machines*, *18*(3).

Floridi, Luciano. (2008b). Understanding Epistemic Relevance. *Erkenntnis*, *69*(1), 69–92.

Floridi, Luciano. (2011). *The Philosophy of Information*. Oxford: Oxford University Press.

Floridi, Luciano. (2012). Semantic information and the network theory of account. *Synthese*, *184*(3), 431–454.

Floridi, Luciano, & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*.

Floridi, Luciano, Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707.

Franklin-Hall, L. R. (2014). High-Level Explanation and the Interventionist's 'Variables Problem".' *The British Journal for the Philosophy of Science*, *67*(2), 553–577.

Galles, D., & Pearl, J. (1995). Testing Identifiability of Causal Effects. *Proceedings of the*

Eleventh Conference on Uncertainty in Artificial Intelligence, 185–195.

Gettier, E. L. (1963). Is Justified True Belief Knowledge? *Analysis*, *23*(6), 121–123.

Goldman, A. (1979). What is Justified Belief? In G. S. Pappas (Ed.), *Justification and Knowledge* (pp. 1–25). Dordrecht: Reidel.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative Adversarial Nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2672–2680).

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine*, *38*(3), 76–99.

Grimm, S. R. (2006). Is Understanding a Species of Knowledge? *The British Journal for the Philosophy of Science*, *57*(3), 515–535.

Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). *Local Rule-Based Explanations of Black Box Decision Systems.*

Gunning, D. (2017). *Explainable Artificial Intelligence (XAI).*

Halpern, J. Y. (2016). *Actual Causality*. Cambridge, MA: MIT Press.

Harman, G., & Kulkarni, S. (2007). *Reliable Reasoning: Induction and Statistical Learning Theory*. Cambridge: MIT Press.

Hausman, D. M., & Woodward, J. (2004). Modularity and the Causal Markov Condition: A Restatement. *The British Journal for the Philosophy of Science*, *55*(1), 147–161.

Hitchcock, C. (1999). Contrastive explanation and the demons of determinism. *British Journal for the Philosophy of Science*, *50*(4).

HLEGAI. (2019). *Ethics guidelines for trustworthy AI.*

Huang, Y., & Valtorta, M. (2006). Pearl's Calculus of Intervention is Complete. *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 217–224.

Huang, Y., & Valtorta, M. (2008). On the completeness of an identifiability algorithm for semi-Markovian models. *Annals of Mathematics and Artificial Intelligence*, *54*(4), 363–408.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Penguin.

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (pp. 2564–2572).

Kelly, K. T., Genin, K., & Lin, H. (2016). Realism, rhetoric, and reliability. *Synthese*, *193*(4), 1191–1223.

Khalifa, K. (2012). Inaugurating Understanding or Repackaging Explanation? *Philosophy of Science*, *79*(1), 15–37.

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, *34*(11), 2767–2787.

Kinney, D. (2018). On the Explanatory Depth and Pragmatic Value of Coarse-Grained, Probabilistic, Causal Explanations. *Philosophy of Science*, *86*(1), 145–167.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in the Age of Algorithms. *Journal of Legal Analysis*.

Kolmogorov, A. N. (1950). *Foundations of the Theory of Probability* (N. Morrison, Ed. & Trans.). New York: Chelsea Publishing Company.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artificial Intelligence Review*, *37*(4), 331–344.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual Fairness. In I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4066–4076).

Lage, I., Chen, E., He, J., Narayanan, M., Gershman, S., Kim, B., & Doshi-Velez, F. (2018). An Evaluation of the Human-Interpretability of Explanation. *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Correcting and Critiquing Trends in Machine Learning*.

Lapuschkin, S., Binder, A., Montavon, G., Müller, K. R., & Samek, W. (2016). Analyzing Classifiers: Fisher Vectors and Deep Neural Networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2912–2920.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). *How we analyzed the COMPAS recidivism algorithm*.

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *arXiv* preprint, 1606.03490.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774).

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

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*.

Mittelstadt, B., Russel, C., & Wachter, S. (2019). Explaining explanations in AI. *Proceedings of FAT* '19: Conference on Fairness, Accountability, and Transparency.*

Nasrabadi, N. (2014). Hyperspectral Target Detection: An Overview of Current and Future Challenges. *IEEE Signal Processing Magazine*, *31*(1), 34–44.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence*.

Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and*

Machines. *29*(3), 441-459.

Pasquale, F. (2015). *The Black Box Society*. Cambridge: Harvard University Press.

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, *82*(4), 669–688.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations*. Washington, D.C.: RAND Corporation.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Routledge.

Potochnik, A. (2015). Causal patterns and adequate explanations. *Philosophical Studies*, *172*(5), 1163–1182.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-Precision Model-Agnostic Explanations. *AAAI*.

Robins, J. M. (1997). Causal Inference from Complex Longitudinal Data. In M. Berkane (Ed.), *Latent Variable Modeling and Applications to Causality* (pp. 69–117). New York: Springer.

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

Rudin, C., Wang, C., & Coker, B. (2018). The age of secrecy and unfairness in recidivism prediction. *arXiv* preprint, 1811.00731.

Sanguinetti, G., & Huynh-Thu, V. A. (2018). *Gene Regulatory Networks: Methods and Protocols*. New York: Springer.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*(7698), 604–610.

Selbst, A., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law*, *7*(4), 233–242.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., … Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140 LP – 1144.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., … Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(19), 10869–10874.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, Prediction, and Search* (2nd Editio).

Strevens, M. (2010). *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Strevens, M. (2013). No understanding without explanation. *Studies in History and Philosophy of Science Part A*, *44*(3), 510–515.

Stutz, D., Hermans, A., & Leibe, B. (2018). Superpixels: An evaluation of the state-of-the-art. *Computer Vision and Image Understanding*, *166*, 1–27.

Tian, J., & Pearl, J. (2002). A General Identification Condition for Causal Effects. *Eighteenth National Conference on Artificial Intelligence*, 567–573. Menlo Park, CA, USA: American Association for Artificial Intelligence.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., … Friend, S. H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*, 530.

van de Vijver, M. J., He, Y. D., van 't Veer, L. J., Dai, H., Hart, A. A. M., Voskuil, D. W., … Bernards, R. (2002). A Gene-Expression Signature as a Predictor of Survival in Breast Cancer. *New England Journal of Medicine*, *347*(25), 1999–2009.

van Fraassen, B. C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Vapnik, V. (1998). *Statistical Learning Theory*. New York: John Wiley & Sons.

Vapnik, V., & Chervonenkis, A. (1971). On the uniform convergence of relative frequencies to their probabilities. *Theory Probab. Appl.*, *16*(2), 264–280.

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

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 and Technology*, *31*(2), 841–887.

Watson, D., & Floridi, L. (2018). Crowdsourced science: sociotechnical epistemology in the e-research paradigm. *Synthese*, *195*(2), 741–764.

Watson, D. S., 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*.

Weinberger, N. (2018). Faithfulness, Coordination and Causal Coincidences. *Erkenntnis*, *83*(2), 113–133.

Weslake, B. (2010). Explanatory Depth. *Philosophy of Science*, *77*(2), 273–294.

Williams, M. (2016). Internalism, Reliabilism, and Deontology. In B. McLaughlin & H. Kornblith (Eds.), *Goldman and His Critics* (pp. 1–21). Oxford: John Wiley & Sons.

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, J. (2008). Cause and Explanation in Psychiatry: An Interventionist Perspective. In K. Kendler & J. Parnas (Eds.), *Philosophical Issues in Psychiatry* (pp. 287–318). Baltimore: Johns Hopkins University Press.

Woodward, J. (2010). Causation in biology: stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318.

Woodward, J. (2015). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, *91*(2), 303–347.

Woodward, J., & Hitchcock, C. (2003). Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs*, *37*(1), 1–24.

Yang, H., Rudin, C., & Seltzer, M. (2017). Scalable Bayesian Rule Lists. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 3921–3930.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., & Telenti, A. (2019). A primer on deep learning in genomics. *Nature Genetics*, *51*(1), 12–18.