Explaining The Behavior Of Black-Box Prediction Algorithms With Causal Learning

Numair Sani[*]
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
snumair1@jhu.edu

Daniel Malinsky[*]
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
malinsky@jhu.edu

Ilya Shpitser
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
ilyas@cs.jhu.edu

Abstract

We propose to explain the behavior of black-box prediction methods (e.g., deep neural networks trained on image pixel data) using causal graphical models. Specifically, we explore learning the structure of a causal graph where the nodes represent prediction outcomes along with a set of macro-level "interpretable" features, while allowing for arbitrary unmeasured confounding among these variables. The resulting graph may indicate which of the interpretable features, if any, are possible causes of the prediction outcome and which may be merely associated with prediction outcomes due to confounding. The approach is motivated by a counterfactual theory of causal explanation wherein good explanations point to factors which are "difference-makers" in an interventionist sense. The resulting analysis may be useful in algorithm auditing and evaluation, by identifying features which make a causal difference to the algorithm's output.

1 Introduction

In recent years, black-box machine learning prediction methods have exhibited impressive performance in a wide range of prediction tasks. In particular, methods based on deep neural networks (DNNs) have been successfully used to analyze image data in settings such as healthcare and social data analytics [15, 16]. An important obstacle to widespread adoption of such methods, particularly in socially-impactful settings, is their black-box nature: it is not obvious, in many cases, how to explain the predictions produced by such algorithms when they succeed (or fail), given that they find imperceptible patterns among high-dimensional sets of features. Moreover, the relevant "explanatory" or "interpretable" units may not coincide nicely with the set of raw features used by the prediction method (e.g., image pixels). Here we present an approch to post-hoc explanation of algorithm behavior which builds on ideas from causality and graphical models. We propose that to explain *post hoc* the output of a black-box method is to understand which variables, from among a set of interpretable features, make a causal difference to the output. That is, we ask which potential targets of manipulation may have non-zero or strong intervention effects on the prediction outcome.

[*] Authors contributed equally.

Preprint. Under review.

There have been numerous approaches to explainability of machine learning algorithms [29, 26, 43, 44, 20, 1, 13, 23, 39]. Many have focused on (what may be broadly called) "feature importance" measures. Feature importance measures typically approach explanation from a purely associational standpoint: features ranked as highly "important" are typically inputs that are highly correlated with the predicted outcome (or prediction error) in the sense of having a large regression coefficient or perturbation gradient, perhaps in the context of a local and simple approximating model class (e.g., linear regression, decision trees, or rule lists). However, the purely associational "importance" standpoint has at least two shortcomings. First, the inputs to a DNN (e.g., individual pixels) are often at the wrong level of description to capture a useful or actionable explanation. For example, an individual pixel may contribute very little to the output of a prediction method but contribute a lot in aggregate – higher-level features comprised of many pixels or patterns across individual inputs (e.g., differences between or variances among collections of lower-level attributes) may be the appropriate ingredients of a more useful explanation. Second, features (at whatever level of description) may be highly associated with outcomes without causing them. Two variables may be highly associated because they are both determined by a common cause that is not among the set of potential candidate features. That is, if the black-box algorithm is in fact tracking some omitted variable that is highly correlated with some input feature, the input feature may be labelled "important" in a way that does not support generalization or guide action. We propose to use causal discovery methods (a.k.a. causal structure learning) to determine which interpretable features, from among a pre-selected set of candidates, may plausably be causal deteriminants of the outcome behavior, and distinguish these causal features from variables that are associated with the behavior due to confounding.

We begin by providing some background on causal explanation and the formalism of causal inference, including causal discovery. We then describe our proposal for explaining the behaviors of black-box prediction algorithms and present a simulation study that illustrates our ideas. We also apply a version of our proposal to annotated image data for bird classification. Finally, we discuss some applications, limitations, and future directions of this work.

2 Causal Explanation

2.1 Explaining Algorithm Behaviors

There is a long history of debate in science and philosophy over what properly constitues an explanation of some phenomenon. (In our case, the relevant phenomenon will be the output of a prediction algorithm.) A connection between explanation and "investigating causes" has been influential, in Western philosophy, at least since Aristotle [2]. More recently, scholarship on *causal explanation* [4, 30, 42, 21] has highlighted various benefits to pursuing understanding of complex systems via causal or counterfactual knowledge, which may be of particular utility to the machine learning community. We focus here primarily on some relevant ideas discussed by Woodward [42] to motivate our perspective in this paper, though similar issues are raised elsewhere in the literature.

In some 20th-century philosophical proposals, explanation was construed via applications of deductive logical reasoning (i.e., showing how observations could be derived from physical laws and background conditions) or simple probabilistic reasoning [18]. One shortcoming of all such proposals is that explanation is intuitively asymmetric: the height of a flagpole explains the length of its shadow (given the sun's position in the sky) but not vice versa; the length of a classic mechanical pendulum explains the device's period of motion, but not vice versa. Logical and associational relationships do not exhibit such asymmetries. Moreover, some true facts or strong associations seem explanatorily irrelevant to a given phenomenon, as when the fact that somebody forgot to water the office rock "explains" why it is not living. (An analogous fact may have been more relevant for an office plant.) Woodward argues that "explanatory relevance" is best understood via counterfactual contrasts and that the asymmetry of explanation reflects the role of causality.

On Woodward's counterfactual theory of causal explanation, explanations answer *what-would-have-been-different* questions. Specifically, the relevant counterfactuals describe the outcomes of interventions or manipulations. X helps explain Y if, under suitable background conditions, some intervention on X produces a change in the distribution of Y. (Here we presume the object of explanation to be the random variable Y, not a specific value. That is, we focus on *type-level* explanation rather *token-level* explanations of particular events.) This perspective has strong connections to the literature on causal models in artificial intelligence [34, 24, 25]. A causal model for outcome

Y precisely stipulates how Y would change under various interventions. So, to explain black-box algorithms we endeavour to build causal models for their behaviors. We propose that such causal explanations can be useful for algorithm evaluation and informing decision-making. In contrast, purely associational measures will be symmetric, include potentially irrelevant information, and fail to support (interventionist) counterfactual reasoning.[1]

Despite a paucity of causal approaches to explainability in the ML literature (with some exceptions, discussed later), survey research suggests that causal explanations are of particular interest to industry practitioners; [3] quote one chief scientist as saying "Figuring out causal factors is the holy grail of explainability," and report similar sentiments expressed by many organizations.

2.2 Causal Modeling

Next we provide some background to make our proposal more precise. Throughout, we use uppercase letters (e.g., X, Y) to denote random variables or vectors and lowercase (x, y) to denote fixed values.

We use *causal graphs* to represent causal relations among random variables [34, 24]. In a causal directed acyclic graph (DAG) $\mathcal{G} = (V, E)$, a directed edge between variables $X \rightarrow Y$ $(X, Y \in V)$ denotes that X is a direct cause of Y, relative to the variables on the graph. Direct causation may be explicated via a system of nonparametric structural equations (NPSEMs) a.k.a. a structural causal model (SCM). The distribution of Y given an intervention that sets X to x is denoted $p(y \mid \text{do}(x))$ by Pearl [24].[2]. Causal effects are often defined as interventional constrasts, e.g., the *average causal effect* (ACE): $\mathbb{E}[Y \mid \text{do}(x)] - \mathbb{E}[Y \mid \text{do}(x')]$ for values x, x'. If an interventional distribution can be written as a function of the observed data distribution under assumptions encoded by a given model, it is said to be *identified*; see [33] for an overview of identification theory.

Given some collection of variables V and observational data on V, one may hope to learn the causal structure, i.e., to select a causal graph supported by the data. We focus on learning causal relations from purely observational (non-experimental) data here, though in some ML settings there exists the capacity to "simulate" interventions directly, which may be even more informative. There exists a significant literature on selecting causal graphs from a mix of observational and interventional data, e.g. [36, 40], and though we do not make use of such methods here, the approach we propose could be applied in those mixed settings as well.

There are a variety of algorithms for causal structure learning, but what most approaches share is that they exploit patterns of statistical constraints implied by distinct causal models to distinguish among candidate graphs. One paradigm is constraint-based learning, which will be our focus here. In constraint-based learning, the aim is to select a causal graph or set of causal graphs consistent with observed data by directly testing a sequence of conditional independence hypotheses – distinct models will imply distinct patterns of conditional independence, and so by rejecting (or failing to reject) a collection of independence hypotheses, one may narrow down the set of models consistent with the data. For example, a paradigmatic constraint-based method is the PC algorithm [34], which aims to learn an equivalence class of DAGs by starting from a complete model (implying no independencies) and removing edges when conditional independence constraints are discovered. Since multiple DAGs may imply the same set of conditional independence constraints, PC estimates a CPDAG (completed partial DAG), a mixed graph with directed and undirected edges that represents a Markov equivalence class of DAGs. (Two graphs are called Markov equivalent if they imply the same conditional independence constraints.) Variations on the PC algorithm and related approaches to selecting CPDAGs have been thoroughly studied in the literature [10, 12]. In settings with unmeasured (latent) confounding variables, it is typical to study graphs with bidirected edges to represent dependence due

[1] Some approaches to explainability focus on a different counterfactual notion: roughly, they aim to identify values in the input space for which a prediction decision changes, assuming all variables are independent of each other [37, 32]. In most settings of interest, the relevant features are not independent of each other, as will become clear in our examples below. Some promising recent work has combined causal knowledge with counterfactual explanations of this sort [22], focusing on counterfactual input values that are consistent with background causal relationships. While interesting, such work is orthogonal to our proposal here, which focuses on type-level rather than token-level explanation, operates on a different set of features than the ones used to generate the prediction, and does not presume that causal relationships among variables are known a priori.

[2] An alternative formalism would denote the post-intervention distribution by $p(Y(x))$, where $Y(x)$ is a *potential outcome* (or *counterfactual*) random variable: the value of Y under the assignment x to X. Differences between these formalisms are irrelevant for our purposes here, c.f. [27].

to confounding. For example, a partial ancestral graph (PAG) [45] is a graphical representation which includes directed edges ($X \rightarrow Y$ means X is a causal ancestor of Y), bidirected edges ($X \leftrightarrow Y$ means X and Y are both caused by some unmeasured common factor(s), e.g., $X \leftarrow U \rightarrow Y$), and partially directed edges ($X \circ\!\!\rightarrow Y$ or $X \circ\!\!-\!\!\circ Y$) where the circle marks indicate ambiguity about whether the endpoints are arrows or tails. (Generally, PAGs can also include undirected or partially undirected edges to represent selection bias, but this is irrelevant for our purposes here.) PAGs inherit their causal interpretation by encoding the commonalities among a set of underlying causal DAGs with latent variables. A bit more formally, a PAG represents an equivalence class of maximal ancestral graphs (MAGs), which encode the independence relations among observed variables when some variables are unobserved [28, 45]. The FCI algorithm [34, 46] is a well-known constraint-based method (similar to PC), which uses sequential tests of conditional independence to select a PAG from data. Variations on the FCI algorithm have also been studied [9, 11].

3 Explaining Black-Box Predictions

Consider a supervised learning setting with a high-dimensional set of "low-level" features (e.g., pixels in an image) $X = (X_1, ..., X_q)$ taking values in an input space $\mathcal{X}^q$ and outcome Y taking values in $\mathcal{Y}$. A prediction or classification algorithm (e.g., DNN) learns a map $f : \mathcal{X}^q \mapsto \mathcal{Y}$. Predicted values from this function are denoted $\widehat{Y}$. To explain the predictions $\widehat{Y}$, we focus on a smaller set of "high-level" features $Z = (Z_1, ..., Z_p)$ $(p \ll q)$ which are "interpretable" in the sense that they correspond to human-understandable and potentially manipulable elements of the application domain of substantive interest, e.g., "presence of tumor in the upper lobe," "diffuse alveolar damage," and so on in lung imaging or "has red colored wing feathers," "has a curved beak," and so on in bird classification. For now, we assume that the interpretable feature set is predetermined and available in the form of an annotated data set, i.e., that each image is labeled with interpretable features. Later, we discuss the possibility of automatically extracting a set of possibly interpretable features from data which lacks annotations. We also assume that Z contains no variables which are deterministically or logically related (e.g., where $Z_i = z_i$ always implies $Z_j = z_j$ for some i, j), though we also return to this issue later. An example from our bird classification dataset (described below) is seen in Figure 1: X consists of the raw pixels and Z includes the interpretable annotations identified in the figure.

Our proposal is to learn a causal PAG $\widehat{\mathcal{G}}$ among the variable set $V = (Z, \widehat{Y})$, with the minimal background knowledge imposed that $\widehat{Y}$ is a non-ancestor of Z (there is no directed path from $\widehat{Y}$ to any element of Z). Additional background knowledge may also be imposed if it is known, for example, that none of the elements of Z may cause each other (they are only dependent due to latent common causes) or there are groups of variables which precede others in a causal ordering. If in $\widehat{\mathcal{G}}$, there is a directed edge $Z_i \rightarrow \widehat{Y}$, then Z_i is a definite cause of the prediction, if instead bidirected edge $Z_i \leftrightarrow \widehat{Y}$ then Z_i is not a cause of the prediction (only dependent due to common latent factors), and if $Z_i \circ\!\!\rightarrow \widehat{Y}$ then Z_i is a possible cause of the prediction. The reason it is important to search for a PAG and not a DAG (or equivalence class of DAGs) is that Z will in general *not* include all possibly relevant variables. Even if only interpretable features strongly associated with prediction outcomes are pre-selected, observed associations may be attributed in part or in whole to latent common causes.

We emphasize that the graphical representation $\widehat{\mathcal{G}}$ is an intentionally crude approximation to the prediction process. By assumption, $\widehat{Y}$ is truly generated from $X_1, ..., X_q$, not $Z_1, ..., Z_p$. However, we view $\widehat{\mathcal{G}}$ as a useful approximation insofar as it captures some of the salient "inner workings" of the opaque and complicated prediction algorithm. If in fact the DNN's internal representation is entirely disjoint from functions of $Z_1, ..., Z_p$, we expect to find that these features are not causes; if some subset of $Z_1, ..., Z_p$ nearly correspond to important features in the internal representation we hope to find that those will be picked out as causes, and have corresponding strong estimated causal effects.

Two causality-inspired approaches to explanation are worth mentioning here. [8] propose a causal attribution method for neural networks. They estimate the ACE of each input neuron on each output neuron under a model assumption motivated by the network structure: they assume if input neurons are dependent, it is only because of latent common causes. The "contribution" of each input neuron is then quantified by the magnitude of this estimated causal effect. This is similar in spirit to our approach, but we do not take the input neurons as the relevant units of analysis (instead focusing on substantively interpretable "macro" features which may be complex aggregates), nor do we assume

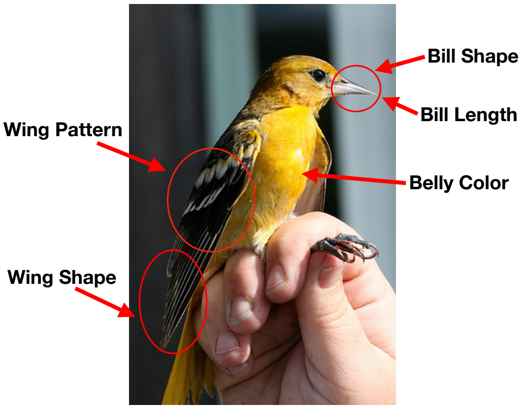


Figure 1: An image of a Baltimore Oriole annotated with interpretable features.

the causal structure is fixed a priori. Our proposal is also not limited to prediction models based on neural networks or any particular model architecture. [31] introduce an approach (CXPlain) which is model-agnostic but similar to [8] in taking the "low-level" input features X as the relevant units of analysis (in contrast with the "high-level" features Z). CXPlain is based on Granger causality and their measure effectively quantifies the change in prediction error when an input feature is left out. Unlike our proposal, this measure does not have an interventionist interpretation except under the restrictive assumption that all relevant variables have been measured, i.e., no latent confounding.

Next we illustrate a basic version of our procedure with a simulation study.

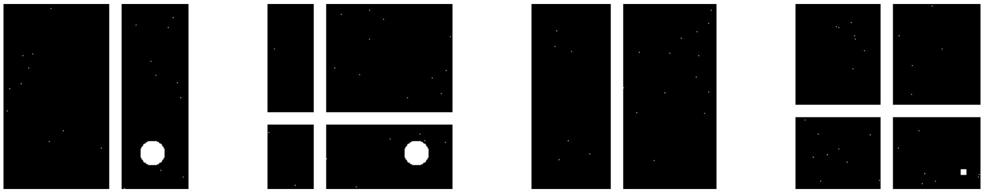
4 A Simulation Study

Consider the following experiment, inspired by and modified from a study reported in [7]. A research subject is presented with a sequence of 2D black and white images, and responds ($Y = 1$) or does not respond ($Y = 0$) with some target behavior for each image presented. The raw features X are then the $d \times d$ image pixels. The images may contain several shapes (alone or in combination) – a horizontal bar (H), vertical bar (V), circle (C), or rectangle (R) – in addition to random pixel noise; see Fig. 2. The target behavior Y is caused only by the presence of verticle bars and circles, in the sense that manipulating the image pixel arrangement to contain a verticle bar or a circle (or both) makes $Y = 1$ much more likely, whereas manipulating the presence of the other shapes does not change the distribution of Y at all. In our simulations, this is accomplished by sampling the target behavior $Y = 1$ with probability depending monotonically only on V and C. However, the various shapes are not independent. Circles and horizontal bars also cause rectangles to be more likely. R would thus be associated with the outcome Y, though conditionally independent given C. H would be marginally independent of Y but dependent given R. The details of the simulation are given below, as well as summarized by the DAG in Fig. 3(a).

$$\begin{aligned}
U_1 &\sim \text{Uniform}(0,1) & V &\sim \text{Bernoulli}(1 - U_1)\\
U_2 &\sim \text{Uniform}(0,1) & C &\sim \text{Bernoulli}(U_2)\\
H &\sim \text{Bernoulli}(U_1) & R &\sim \text{Bernoulli}(\text{expit}(0.75H + 0.5C))\\
Y &\sim \text{Bernoulli}(\text{expit}(-0.5 + 2.5V + 1.75C))
\end{aligned}$$

Using the above process, 5000 images are generated. Next, we train a deep convolutional neural network (CNN) on the raw image pixels X according to the standard supervised paradigm. We use Resnet18 [17] and initialize the weights with values estimated by pre-training the network on ImageNet [14]. This was done using the PyTorch framework.[3] The model performs quite well – 81% accuracy on a held out test set of 2000 images – so we may reasonably hope that these predictions track the underlying mechanism by which the data was generated. Our interpretable features Z are indicators of the presence of the various shapes in the image. Since the true underlying behavior

[3] The software is freely available at: https://pytorch.org



Labeled Y = 1 Labeled Y = 1 Labeled Y = 0 Labeled Y = 0

Figure 2: Simulated image examples with horizontal bars, vertical bars, circles, and rectangles.

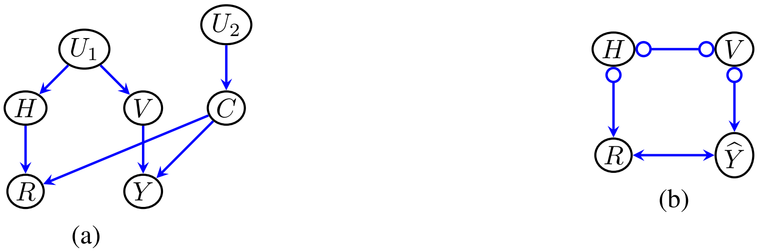


Figure 3: (a) A causal diagram representing the true data generating process. (b) The PAG learned using FCI.

is causally determined by V and C, we expect V and C to be "important" for the predictions $\widehat{Y}$, but due to the mutual dependence the other shapes are also highly correlated with $\widehat{Y}$. Moreover, we mimic a setting where C is (wrongfully) omitted from the set of candidate interpretable features; in practice, the feature set proposed by domain experts or available in the annotated data will typically exclude various relevant determinants of the underlying outcome. In that case, C is an unmeasured confounder. Applying the PAG-estimation algorithm FCI to variable set $(H, V, R, \widehat{Y})$ we learn the structure in Fig. 3(b), which indicates correctly that V is a (possible) cause, but that R is not: R is associated with the prediction outcomes only due to an unmeasured common cause.[4] This simple example illustrates how the estimated PAG using (incomplete) interpretable features can be useful: we disentangle mere statistical associations from (potentially) causal relationships, in this case indicating correctly that interventions on V may make a difference to the prediction algorithm's behavior but interventions on R and H would not, and moreover that there are potentially important causes of the output (C) that have been excluded from our set of candidate features.

5 Application: Bird Classification

We demonstrate the utility of our approach by explaining a bird classification neural network, trained on the Caltech-UCSD 200-2011 image dataset [38]. It consists of 200 categories of birds, with 11,788 images in total. Each image comes annotated with 312 binary attributes describing interpretable bird characteristics like eye color, size, wing shape, etc. We build a black-box prediction model using raw pixel features to predict the class of the bird and then use FCI to explain the output of the model.

5.1 Data Preprocessing & Model Training

Since many of the species classifications have few associated images, we first broadly group species into 9 coarser groups. For example we group the Baird Sparrow and Black Throated Sparrow into one Sparrow class. Details on the grouping scheme can be found in the supplementary material. This leads to 9 possible outcome labels and 5514 images (instances that do not fit into one of these categories are excluded from analysis). The number of images across each class is not evenly balanced. So, we

[4] Under a somewhat different data-generating process, FCI could learn a directed edge $V \rightarrow \widehat{Y}$ instead of a partially directed edge, thus more definitively ruling out unmeasured confounding between V and $\widehat{Y}$. This depends on the observed patterns of association and independence; see [45] on so-called "visible" and "invisible" edges in a PAG.

Table 1: Possible Causes of Bird Category with Relative Frequencies

Size: 82.5%	Wing Shape: 52.5%	Belly Pattern: 52.5%	Bill Color: 52.5%
Bill Length: 62.5%	Back Pattern: 52.5%	Tail Pattern: 52.5%	Primary Color: 52.5%
Bill Shape: 62.5	Upper Tail color: 52.5%	Wing Color: 52.5%	Leg Color: 52.5%
Eye Color: 57.5%	Breast Pattern: 52.5%	Nape Color: 52.5%	Belly Color: 52.5%
Forehead Color: 52.5%	Under Tail Color: 52.5%	Wing Pattern: 52.5%	Breast Color: 52.5%
Upperparts Color: 52.5%	Tail Shape: 52.5%	Throat Color: 52.5%	
Underparts Color: 52.5%	Crown Color: 52.5%	Back Color: 52.5%	

subsample overrepresented classes in order to get roughly equal number of images per class. This yields a dataset of 3538 images, which is then partitioned into a training, validation, and testing datasets of 2849, 520, and 529 images respectively. We train the ResNet18 architecture (pre-trained on ImageNet) on our dataset and achieve an accuracy of 86.57% on the testing set. We consolidate the 312 available binary attributes into ordinal attributes. For example, four binary attributes describing back pattern, namely `has_back_pattern::solid, has_back_pattern::spotted, has_back_pattern::striped, has_back_pattern::multi-colored` are consolidated into one attribute `back pattern` taking on five possible values: the 4 designations above and an additional category if the attribute is missing. Even once this consolidation is performed, there still exist many attributes with a large number of possible values, so we group together similar values. For example we group dark colors including gray, black, and brown into one category and warm colors including red, yellow, and orange into another. Other attributes are consolidated similarly. The full grouping scheme is described in the supplementary material. After the above preprocessing, for each image we have a predicted label from the CNN and 26 ordinal attributes.

5.2 Structure Learning

Structure learning methods such as FCI produce a single estimated PAG as output. In applications, it is common to repeat the graph estimation on multiple bootstraps or subsamples of the original data, in order to control false discovery rates or to mitigate the cumulative effect of statistical errors in the independence tests [35]. We create 40 bootstrapped replicates of the dataset and run FCI on each with tuning parameter $\alpha = 10^{-5}$, with the additional background knowledge imposed that $\widehat{Y}$ cannot cause any of the interpretable features.[5] Here FCI is used with the χ^2 independence test, and we limit the maximum conditioning set size to 4 for computational tractability.

5.3 Results

We compute the relative frequency over bootstrap trials of $Z_i \circ\!\!\rightarrow \widehat{Y}$ edges from all attributes. This represents the frequency with which an attribute is determined to be a possible cause of the predicted label and constitutes a rough measure of "confidence" in that attribute's causal status. The computed relative frequencies are presented in the Table 1. We find that the most likely candidate causes include `size, bill length, bill shape,` and `eye color`, which are intuitively salient features for distinguishing among bird categories. We have lower confidence that the other features are possible causes of $\widehat{Y}$.

6 Discussion

We have presented a tool to support explaining the behavior of black-box prediction algorithms. Below we discuss some potential uses and limitations.

6.1 Algorithm Auditing and Evaluation

One important goal related to building explainable AI systems is the auditing and evaluation of algorithms post-hoc. If a prediction algorithm appears to perform well, it is important to understand why it performs well before deploying the algorithm. Users will want to know that the algorithm is "paying attention to" the right aspects of the input, and not tracking spurious artifacts [3]. This is important both from the perspective of generalization to new domains as well as from the perspective

[5] We use the command-line interface to the TETRAD freeware: `https://github.com/cmu-phil/tetrad`

of fairness. To illustrate the former, consider instances of "dataset bias" or "data leakage" wherein an irrelevant artifact of the data collection process proves very important to the performance of the prediction method. This may impede generalization to other data sets where the artifact is absent or somehow the data collection process is different. For example, [41] study the role of violet image highlighting on dermoscopic images in a skin cancer detection task. They find that this image highlighting significantly affects the likelihood that a skin lesion is classified as cancerous by a commerical CNN tool. (The researchers are able to diagnose the problem because they have access to the same images pre- and post-highlighting: effectively, they are able to simulate an intervention on the highlighting.)

To illustrate the fairness perspective, consider a recent episode of alleged racial bias in Google's Vision Cloud API, a tool which automatically labels images into various categories [19]. Users found an image of a dark-skinned hand holding a non-contact digital thermometer that was labelled "gun" by the API, while a similar image with a light-skinned individual was labelled "electronic device." More tellingly, when the image with dark skin was crudely altered to contain light beige-colored skin (an intervention on "skin color"), the same object was labelled "monocular." This simple experiment was suggestive of the idea that skin color was inappropriately a cause of the object label and tracking biased or stereotyped associations. Google apologized and revised their algorithm, though denied any "evidence of systemic bias related to skin tone."

Auditing algorithms to determine whether inappropriate features have a causal impact on the output can be an import part of the bias-checking pipeline. Moreover, a benefit of our proposed approach is that the black-box model may be audited without access to the model itself, only the predicted values. This may be desirable in some settings where the model itself is proprietary.

6.2 Informativeness and Background Knowledge

It is important to emphasize that a PAG-based causal discovery analysis is informative to a degree that depends on the data (the patterns of association) and the strength of imposed background knowledge. Here we only imposed the minimal knowledge that $\widehat{Y}$ is not a cause of any of the image features and we allowed for arbitrary causal relationships and latent structure otherwise. Being entirely agnostic about the possibility of unmeasured confounding may lead, depending on the observed patterns of dependence and independence in the data, to only weakly informative results if the patterns of association cannot rule out possible confounding anywhere. If the data fails to rule out confounders and fails to identify definite causes of $\widehat{Y}$, this does not indicate that the analysis has failed but just that only negative conclusions are supported – e.g., the chosen set of interpretable features and background assumptions are not sufficiently rich to identify the causes of the output. It is standard in the causal discovery literature to acknowledge that the strength of supported causal conclusions depends on the strength of input causal background assumptions [34]. In some cases, domain knowledge may support restricting the set of possible causal structures, e.g., when it is believed the some relationships must be unconfounded or some correlations among Z may only be due to latent variables (because some elements of Z cannot cause each other).

6.3 Selecting or Constructing Interpretable Features

In our experiments we use hand-crafted interpretable features that are available in the form of annotations with the data. Annotations are not always available in applications. In such settings, one approach would be to manually annotate the raw data with expert evaluations, e.g., when clinical experts annotate medical images with labels of important features (e.g. "tumor in the upper lobe," "diffuse aveolar damage," etc). Alternatively, one may endeavour to extract interpretable features automatically from the raw data. Unsupervised learning techniques may be used in some contexts to construct features, though in general there is no guarantee that these will be substantively (e.g., clinically) meaningful or correspond to manipulable elements of the domain. However, a series of papers [7, 5, 6] has introduced techniques for extracting *causal features* from "low-level" data, i.e., features that have a causal effect on the target outcome. We leave to future work the possibility of combining this causally-motivated feature extraction technique with the present proposal. A related issue is that "high-level" features associated with data sets (whether they are extracted by human judgement or automatically) may not always be interpretable in the sense intended here: they may not correspond to manipulable elements of the domain. Moreover, some features may be deterministically related (which would pose a problem for most structure learning algorithms), and so some feature pre-selection may be necessary. Thus, human judgement may be indispensible at the feature-selection stage of the process.

7 Conclusion

Causal structure learning algorithms – specifically PAG learning algorithms such as FCI and its variants – may be a valuable tool for explaining black-box prediction methods. We have demonstrated the utility of using FCI in both simulated and real data experiments, where we are able to distinguish between possible causes of the prediction outcome and features that are associated due to unmeasured confounding. We hope the analysis presented here stimulates further cross-pollination between research communities focusing on causal discovery and explainable AI.

References

[1] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

[2] Aristotle. *Posterior Analytics*.

[3] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

[4] Nancy Cartwright. Causal laws and effective strategies. *Noûs*, pages 419–437, 1979.

[5] Krzysztof Chalupka, Tobias Bischoff, Pietro Perona, and Frederick Eberhardt. Unsupervised discovery of el nino using causal feature learning on microlevel climate data. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence*, pages 72–81, 2016.

[6] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164, 2017.

[7] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence*, pages 181–190, 2015.

[8] Aditya Chattopadhyay, Piyushi Manupriya, Anirban Sarkar, and Vineeth N Balasubramanian. Neural network attributions: A causal perspective. In *Proceedings of the 36th International Conference on Machine Learning*, pages 981–990, 2019.

[9] Tom Claassen and Tom Heskes. A Bayesian approach to constraint based causal inference. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 207–216, 2012.

[10] Diego Colombo and Marloes H. Maathuis. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1):3741–3782, 2014.

[11] Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Annals of Statistics*, pages 294–321, 2012.

[12] Ruifei Cui, Perry Groot, and Tom Heskes. Copula PC algorithm for causal discovery from mixed data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 377–392, 2016.

[13] Sanjeeb Dash, Oktay Gunluk, and Dennis Wei. Boolean decision rules via column generation. In *Advances in Neural Information Processing Systems*, pages 4655–4665, 2018.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009.

[15] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

[16] Rita Georgina Guimaraes, Renata L. Rosa, Denise De Gaetano, Demostenes Z Rodriguez, and Graca Bressan. Age groups classification in social network using deep learning. *IEEE Access*, 5:10805–10816, 2017.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[18] Carl G. Hempel. *Aspects of scientific explanation*. Free Press, 1965.

[19] Nicolas Kayser-Bril. Google apologizes after its Vision AI produced racist results. *AlgorithmWatch*, 2020. `https://algorithmwatch.org/en/story/google-vision-racism/`.

[20] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.

[21] Tania Lombrozo and Nadya Vasilyeva. Causal explanation. In *Oxford Handbook of Causal Reasoning*, pages 415–432. Oxford University Press, 2017.

[22] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving causal constraints in counterfactual explanations for machine learning classifiers. *arXiv preprint arXiv:1912.03277*, 2019.

[23] Christoph Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[24] Judea Pearl. *Causality*. Cambridge University Press, 2009.

[25] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[27] Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Working Paper 128*, pages 1–146, 2013.

[28] Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030, 2002.

[29] Peter R Rijnbeek and Jan A. Kors. Finding a short and accurate decision rule in disjunctive normal form by exhaustive search. *Machine Learning*, 80(1):33–62, 2010.

[30] Wesley C. Salmon. *Scientific explanation and the causal structure of the world*. Princeton University Press, 1984.

[31] Patrick Schwab and Walter Karlen. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems*, pages 10220–10230, 2019.

[32] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models. *arXiv preprint arXiv:1905.07857*, 2019.

[33] Ilya Shpitser. Identification in graphical causal models. In *Handbook of Graphical Models*, pages 381–403. CRC Press, 2018.

[34] Peter L. Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

[35] Daniel J. Stekhoven, Izabel Moraes, Gardar Sveinbjörnsson, Lars Hennig, Marloes H. Maathuis, and Peter Bühlmann. Causal stability ranking. *Bioinformatics*, 28(21):2819–2823, 2012.

[36] Sofia Triantafillou and Ioannis Tsamardinos. Constraint-based causal discovery from multiple interventions over overlapping variable sets. *Journal of Machine Learning Research*, 16:2147–2205, 2015.

[37] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841, 2017.

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[39] Tong Wang. Gaining free or low-cost transparency with interpretable partial substitute. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6505–6514, 2019.

[40] Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems*, pages 5822–5831, 2017.

[41] Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatology*, 155(10):1135–1141, 2019.

[42] James Woodward. *Making things happen: A theory of causal explanation*. Oxford University Press, 2005.

[43] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. Scalable Bayesian rule lists. In *Proceedings of the 34th International Conference on Machine Learning*, pages 3921–3930, 2017.

[44] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.

[45] Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9:1437–1474, 2008.

[46] Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896, 2008.

Supplement: Preprocessing for Birds Dataset

Bird Classification Grouping

The Caltech-UCSD 200-2011 Birds dataset comes with 200 categories of birds. We coarsen them using the group scheme below:

- Flycatcher
  - 041.Scissor_Tailed_Flycatcher
  - 040.Olive_Sided_Flycatcher
  - 043.Yellow_Bellied_Flycatcher
  - 038.Great_Crested_Flycatcher
  - 042.Vermilion_Flycatcher
  - 037.Acadian_Flycatcher
  - 039.Least_Flycatcher
- Gull
  - 059.California_Gull
  - 065.Slaty_Backed_Gull
  - 063.Ivory_Gull
  - 060.Glaucous_Winged_Gull
  - 066.Western_Gull
  - 062.Herring_Gull
  - 064.Ring_Billed_Gull
  - 061.Heermann_Gull
- Kingfisher
  - 081.Pied_Kingfisher
  - 082.Ringed_Kingfisher
  - 080.Green_Kingfisher
  - 079.Belted_Kingfisher
  - 083.White_Breasted_Kingfisher
- Sparrow
  - 127.Savannah_Sparrow
  - 126.Nelson_Sharp_Tailed_Sparrow
  - 116.Chipping_Sparrow
  - 114.Black_Throated_Sparrow
  - 121.Grasshopper_Sparrow
  - 119.Field_Sparrow
  - 122.Harris_Sparrow
  - 130.Tree_Sparrow
  - 128.Seaside_Sparrow
  - 118.House_Sparrow
  - 133.White_Throated_Sparrow
  - 115.Brewer_Sparrow
  - 117.Clay_Colored_Sparrow
  - 131.Vesper_Sparrow
  - 123.Henslow_Sparrow
  - 120.Fox_Sparrow
  - 129.Song_Sparrow
  - 125.Lincoln_Sparrow

  - 132.White_Crowned_Sparrow
  - 124.Le_Conte_Sparrow
  - 113.Baird_Sparrow
- Tern
  - 141.Artic_Tern
  - 144.Common_Tern
  - 146.Forsters_Tern
  - 143.Caspian_Tern
  - 145.Elegant_Tern
  - 147.Least_Tern
  - 142.Black_Tern
- Vireo
  - 156.White_Eyed_Vireo
  - 154.Red_Eyed_Vireo
  - 152.Blue_Headed_Vireo
  - 157.Yellow_Throated_Vireo
  - 153.Philadelphia_Vireo
  - 155.Warbling_Vireo
  - 151.Black_Capped_Vireo
- Warbler
  - 168.Kentucky_Warbler
  - 163.Cape_May_Warbler
  - 158.Bay_Breasted_Warbler
  - 162.Canada_Warbler
  - 182.Yellow_Warbler
  - 161.Blue_Winged_Warbler
  - 167.Hooded_Warbler
  - 180.Wilson_Warbler
  - 171.Myrtle_Warbler
  - 166.Golden_Winged_Warbler
  - 173.Orange_Crowned_Warbler
  - 177.Prothonotary_Warbler
  - 178.Swainson_Warbler
  - 169.Magnolia_Warbler
  - 159.Black_and_White_Warbler
  - 165.Chestnut_Sided_Warbler
  - 164.Cerulean_Warbler
  - 174.Palm_Warbler
  - 176.Prairie_Warbler
  - 160.Black_Throated_Blue_Warbler
  - 179.Tennessee_Warbler
  - 170.Mourning_Warbler
  - 172.Nashville_Warbler
  - 181.Worm_Eating_Warbler
  - 175.Pine_Warbler
- Woodpecker
  - 188.Pileated_Woodpecker
  - 189.Red_Bellied_Woodpecker
  - 190.Red_Cockaded_Woodpecker
  - 187.American_Three_Toed_Woodpecker
  - 191.Red_headed_Woodpecker,
  - 192.Downy_Woodpecker
- Wren

  - 195.Carolina_Wren
  - 197.Marsh_Wren
  - 196.House_Wren
  - 193.Bewick_Wren
  - 198.Rock_Wren
  - 199.Winter_Wren
  - 194.Cactus_Wren

Any other bird labels that did fit into these categories were excluded from our analysis.

Attribute Groupings

The attributes were grouped as follows:

- bill_shape:
  - 0 - Missing
  - 1 - curved_(up_or_down), hooked, hooked_seabird
  - 2 - dagger, needle, cone
  - 3 - specialized, all-purpose
  - 4 - spatulate
- wing_color:
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- upperparts_color:
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- underparts_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- breast_pattern
  - 0 - Missing
  - 1 - solid
  - 2 - spotted
  - 3 - striped
  - 4 - multi-colored
- back_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- tail_shape
  - 0 - Missing
  - 1- forked_tail
  - 2 - rounded_tail

  - 3 - notched_tail
  - 4 - fan-shaped_tail
  - 5 - pointed_tail
  - 6 - squared_tail
- upper_tail_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- breast_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- throat_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- eye_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- bill_length
  - 0 - Missing
  - 1 - about_the_same_as_head
  - 2 - longer_than_head
  - 3 - shorter_than_head
- forehead_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- under_tail_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- nape_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- belly_color

  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white

- wing_shape
  - 0 - Missing
  - 1 - rounded-wings
  - 2 - pointed-wings
  - 3 - broad-wings
  - 4 - tapered-wings
  - 5 - long-wings
- size
  - 0 - Missing
  - 1 - large_(16-32in)
  - 2 - small_(5-9in)
  - 3 - very_large_(32-72in)
  - 4 - medium_(9-16in)
  - 5 - very_small_(3-5in)
- back_pattern
  - 0 - Missing
  - 1 - solid
  - 2 - spotted
  - 3 - striped
  - 4 - multi-colored
- tail_pattern
  - 0 - Missing
  - 1 - solid
  - 2 - spotted
  - 3 - striped
  - 4 - multi-colored
- belly_pattern
  - 0 - Missing
  - 1 - solid
  - 2 - spotted
  - 3 - striped
  - 4 - multi-colored
- primary_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- leg_color
  - 0 - Missing
  - 1 - blue, yellow, red
  - 2 - green, purple, orange, pink, buff, iridescent
  - 3 - rufous, grey, black, brown
  - 4 - white
- bill_color
  - 0 - Missing
  - 1 - blue, yellow, red

    - 2 - green, purple, orange, pink, buff, iridescent
    - 3 - rufous, grey, black, brown
    - 4 - white

- crown_color
    - 0 - Missing
    - 1 - blue, yellow, red
    - 2 - green, purple, orange, pink, buff, iridescent
    - 3 - rufous, grey, black, brown
    - 4 - white
- wing_pattern
    - 0 - Missing
    - 1 - solid
    - 2 - spotted
    - 3 - striped
    - 4 - multi-colored

Supplement: Deep Network Hyperparameters

The ResNet18 architecture utilized used pre-trained weights learned by training on the ImageNet dataset. We then fine-tuned ResNet on our birds dataset. We minimized cross-entropy loss using PyTorch's built-in stochastic gradient descent algorithm. We specified a learning rate of 0.001 along with momentum equal to 0.9, and the learning rate was decayed by a factor of 0.1 every 7 epochs.