Explaining Model Behavior with Global Causal Analysis

Marcel Robeer[1,2(✉)], Floris Bex[2,3], Ad Feelders[2], and Henry Prakken[2,4]

[1] Netherlands Police Lab AI, Netherlands Police, The Hague, The Netherlands
[2] Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
m.j.robeer@uu.nl
[3] Tilburg Institute for Law, Technology, and Society, Tilburg University, Tilburg, The Netherlands
[4] Faculty of Law, University of Groningen, Groningen, The Netherlands

Abstract. We present Global Causal Analysis (GCA) for text classification. GCA is a technique for global model-agnostic explainability drawing from well-established observational causal structure learning algorithms. GCA generates an explanatory graph from high-level human-interpretable features, revealing how these features affect each other and the black-box output. We show how these high-level features do not always have to be human-annotated, but can also be computationally inferred. Moreover, we discuss how the explanatory graph can be used for global model analysis in natural language processing (NLP): the graph shows the effect of different types of features on model behavior, whether these effects are causal effects or mere (spurious) correlations, and if and how different features interact. We then propose a three-step method for (semi-)automatically evaluating the quality, fidelity and stability of the GCA explanatory graph without requiring a ground truth. Finally, we provide a detailed GCA of a state-of-the-art NLP model, showing how setting a global one-versus-rest contrast can improve explanatory relevance, and demonstrating the utility of our three-step evaluation method.

Keywords: Explainable Machine Learning (XML) · Causal explanation · Model-agnostic explanation · Natural Language Processing (NLP)

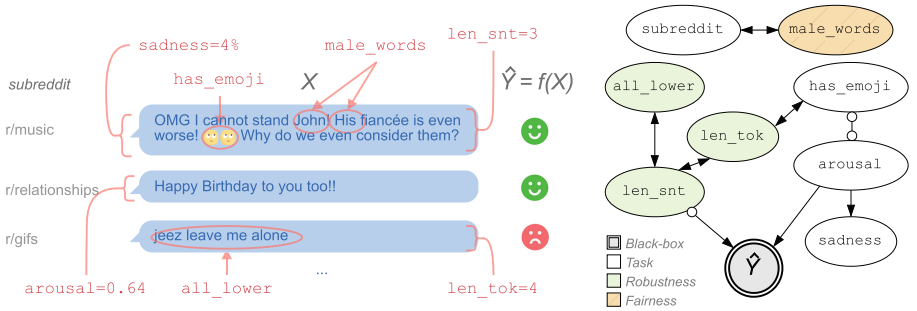
1 Introduction

Explaining the global behavior of a machine learning (ML) model remains a difficult and laborious task. It is hard to distinguish features with directed influences from ones related through (spurious) correlations. Causal explanations could help in this regard, providing explanations discerning causal effects from correlational ones [9]. Even when these can be distinguished, then generalizing—in a human-understandable way—if and how features relate to the model behavior

L. Longo (Ed.): xAI 2023, CCIS 1901, pp. 299–323, 2023.
https://doi.org/10.1007/978-3-031-44064-9_17

over a large input space remains challenging. Providing a human-understandable explanation inevitably requires selection (e.g. limiting the features under consideration by setting a contrast between outputs) and an appropriate level of explanation (e.g. abstracting detailed behavior into high-level tasks) [36].

Sani, Malinsky and Shpitser [48] proposed a method to explain the global behavior of black-box prediction methods using well-established causal graphical model learning techniques. Their method summarizes the behavior of a black-box model (e.g. a convolutional neural network) using low-level features (e.g. pixels) to predict a label (e.g. bird species), with a graph of high-level 'human-interpretable' features (e.g. the belly color, and wing pattern and shape). The generated global (causal) graph shows the (in)dependence relations amongst the high-level features themselves and with the predicted label, and how these are affected by unobserved confounders. Sani et al. illustrate the utility of their approach on image classification tasks with human-annotated high-level features: a simulated dataset, bird classification and pneumonia detection from X-rays. However, they were unable to (a) infer/select features with computational approaches for the image modality—thus always requiring expensive and time-consuming human annotation—and (b) assess the causal graph quality and faithfulness to the model—providing no guarantee that the explanatory model generalizes well over the data and is actually telling of model behavior.



(a) Example data with predictions and inferred features (b) GCA explanatory graph

Fig. 1. Explaining the global behavior of a black-box predictor of emotions on Reddit comment data, with a (causal) explanatory graph for multi-aspect analysis of high-level features related to the task, robustness and fairness.

In this paper, we show how to computationally infer high-level features and how to use them to study multiple aspects of model behavior. We introduce a method to evaluate the quality and faithfulness of the explanatory model without requiring a ground-truth reference. Additionally, we enhance explanatory relevance through class-wise contrastive explanations. Global Causal Analysis (GCA)[1] summarizes black-box model behavior in a single graphical overview,

[1] https://github.com/MarcelRobeer/GlobalCausalAnalysis.

showing (directed) feature interactions, and if and how they influence the black-box decision function. Key to GCA is our proposed (semi-)automatic evaluation method, which supports in estimating (a) how telling the features are of model behavior and (b) the quality of the explanatory graph.

Figure 1 illustrates the example case in this paper, explaining the behavior of a black-box predictor $\hat{Y} = f(X)$ for sentiment analysis of Reddit comments X over a dataset (e.g. the test set). The inferred high-level features (e.g. the presence of `male_words`, `arousal` score and number of tokens [`len_tok`]) and the human-annotated ones (e.g. `subreddit`) are combined in a global explanatory (causal) graph, showing features with a direct influence on prediction $\hat{Y}$ ($\cdot \rightarrow \hat{Y}$), indirectly related features (with a directed path to $\hat{Y}$), correlated features ($\leftrightarrow$ indicates a confounder $\cdot \leftarrow U \rightarrow \cdot$) and uncertain directed relations ($\circ$ indicates an end can be $<$ or $-$). In summary, we make the following contributions:

1. We introduce the idea to use GCAs to describe model-agnostic black-box behavior to the area of **natural language processing** (NLP)—which has a well-established body of work on linguistic phenomena and methods for inferring them [4]—;
2. we extend the human-labelled features with **inferred high-level features**—considering model behavior with features related to multiple aspects, such as features related to the task at hand, robustness (generalizability) and fairness (protected attributes)—;
3. we propose a three-step method to (semi-)automatically **evaluate the quality, fidelity, and stability** of the explanatory graph—which does not rely on a ground truth as these are unavailable for a black box [22]—, and;
4. we study improving the relevance of high-level features by applying concepts from **global (type-level) contrastive explanation** [35,37,60].

The remainder of this paper is structured as follows. Section 2 discusses the background, techniques & evaluations of model-agnostic global explanation and causal models, and details the technique in [48]. Section 3 describes our extension for the NLP domain, and the experimental set-up. Section 4 discusses the results of the experiment, and illustrates GCA with three detailed analyses. Finally, Section 5 summarizes our findings and provides avenues for future research.

2 Background: Model-Agnostic Global Explanation and Causal Models

We describe the background on model-agnostic global explanation and causal models, and provide a detailed description of how causal models can be applied for model-agnostic global explanation. *Global explanation* (sometimes referred to as *model explanation*) aims to provide insights into the entire machine learning (ML) model it aims to explain [5]. It is distinguished in scope from *local explanation* (*instance explanation*; with well-known techniques such as LIME [44] and SHAP [34]) where the aim is to explain individual outputs by the ML

model [5,23]. Their counterparts in causal explanation are *type-level* causality (akin to global explanation it describes general relations amongst variables) and *token-level* causality (like local explanation, focusing on individual events) [61].

2.1 Model-Agnostic Global Explanation

In our work, we focus on *model-agnostic* explanations: querying a black-box on its input-output behavior to derive an explanation. Unlike model-specific explanations, model-agnostic explanations have the benefit of being applicable to any type of ML model for a task type (e.g. classification or regression), and provide flexibility regarding the explanation and its representation [43]. Model-agnostic explanations are a type of post-hoc (pedagogical) explanation [5]. These are explanations that are applied after training the ML model [5].

Some well-known model-agnostic global explanation methods (an overview is given by [23] and [5]) are ones that study the relation between individual features and a model output (typically in tabular data)—*Partial Dependence Plots* (PDPs) [17], *Individual Conditional Expectation* (ICE) plots [21] and *Average Local Effect* (ALE) plots [1]—model-agnostic global feature importance scores—e.g. *Model Class Reliance* (MCR) [16]—, and global surrogates that approximate a black-box $f(\cdot)$ with a more interpretable model $g(\cdot)$ and use that directly for explainability—e.g. TREPAN [12,13], *Model Extraction* [3], *Black-Box Explanations through Transparent Approximations* (BETA) [30] and *Transparent Model Distillation* [54]. In addition, specifically relevant to our work are *Variable Interaction Networks* (VINs) [28] (evaluating the importance of non-additive tabular feature interactions with a graph), causal interpretations of black-box models [64] (showing how PDPs can be used in conjunction with a known causal graph) and LEWIS [18] (analyzing model behavior on tabular data with plots, including the influence of contextual factors such as sex).

Definitions. In ML, we train a model (e.g. a classifier) $f : \mathcal{X}^q \mapsto \mathcal{Y}$ taking inputs $X \in \mathcal{X}^q$ (e.g. texts) and transforming them into outputs $Y \in \mathcal{Y}$ (e.g. class labels). Training can be done in many ways, such as the supervised paradigm—where we provide it with a dataset $D = (X, Y)$ with example input-label pairs—, or clustering—assigning instances X to k clusters based on their similarity. To illustrate our idea, in our paper we focus on supervised classification models.

The *model-agnostic global explanation problem* involves finding an explanatory function $g(\cdot)$ that explains the behavior of $\hat{Y} = f(X)$ over some dataset $D' = (X, \hat{Y})$ [23].[2] From this function (e.g. a surrogate decision tree for global behavior), we then extract a set of explanations E (e.g. rules from the decision tree) that model the behavior of $f(\cdot)$ in a human-understandable way [23].

[2] Note that the dataset D' used for explanation does not have to be the same as the dataset D used for training, but can be e.g. the test set [5].

Evaluation. Several properties are important when considering the quality of a global explanation. Perhaps the most important property of an explanation method is its *fidelity* (faithfulness) to the model it aims to explain [6,29]. If the explanatory model $g(\cdot)$ is also a predictive model (e.g. a decision tree or sparse linear regressor), fidelity is typically estimated by calculating the predictive performance of the predicted labels $\hat{Y}' = g(X)$ of the surrogate model on the predicted labels of black-box $\hat{Y} = f(X)$ [13,44]. Another important property is the explanation *stability* (robustness) [6]. Stability is an indicator for the reliability and generalizability of the explanation [6,58]. A stable explainable ML method minimizes the effect of randomness and sampling on its performance [6]. Stability is either evaluated by applying small perturbations δ to the inputs X and taking the mean distance between $g(X)$ and $g(X + \delta)$ [58], or by drawing subsamples from the data to measure the effect of data distribution shifts [31].

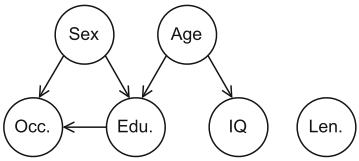
2.2 Causal Graphs

Graphical Markov models use a graph consisting of nodes V and edges E to represent (conditional) independence relations among a set of variables [46]. We discuss four well-established types of graphical Markov models (two assuming no latent variables and two indicating the effect of latent confounders), search algorithms for causal structure learning, and how these algorithms are evaluated.

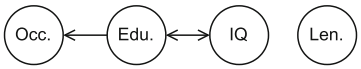
Graphs Assuming Causal Sufficiency. *Directed acyclic graphs* (DAGs) $\mathcal{G} = (V, E)$ consist of directed edges $V_i \rightarrow V_j$ between nodes (at most one between any two nodes), and are not allowed to contain cycles [46]. DAGs imply conditional independencies amongst the variables, where conditional independence $X_i \perp\!\!\!\perp X_j \mid \mathbf{X}$ indicates that the set of elements $\mathbf{X}$ blocks all paths between X_i and X_j.[3] An example DAG is provided in Fig. 2a, with nodes $V = \{$`Age`, `Education`, `IQ`, `Length of Application Letter`, `Occupation`, `Sex`$\}$. The graph implies independencies $\{$`Age` $\perp\!\!\!\perp$ `Occ` $\mid$ (`Edu`, `Sex`), `Age` $\perp\!\!\!\perp$ `Sex`, `Edu` $\perp\!\!\!\perp$ `IQ` $\mid$ `Age`, `IQ` $\perp\!\!\!\perp$ `Occ` $\mid$ `Age`, `IQ` $\perp\!\!\!\perp$ `Occ` $\mid$ (`Edu`, `Sex`), `Len` $\perp\!\!\!\perp \{$`Age`, `Edu`, `IQ`, `Sex`$\}\}$.

DAGs are Markov equivalent if they imply the same independencies. A *completed partially directed acyclic graph* (CPDAG) $\mathcal{C}_{[\mathbf{G}]}$ is a unique representation of the Markov equivalence class (MEC) of DAGs $[\mathbf{G}]$ that have the same skeleton graph (the same graph, where the edge marks are removed from the edges) and v-structures (subgraphs with structure $V_i \rightarrow V_k \leftarrow V_j$). CPDAGs can contain two types of edges: (1) a *directed edge* $V_i \rightarrow V_j$ indicates that $V_i \rightarrow V_j$ in all DAGs in the equivalence class, and; (2) an *undirected edge* V_i—V_j indicates that in some DAGs in $[\mathbf{G}]$ there is an edge $V_i \rightarrow V_j$ and in others an edge $V_i \leftarrow V_j$.

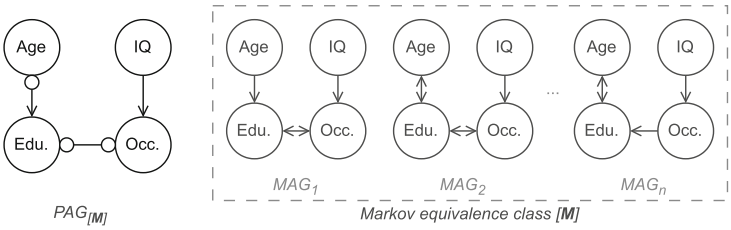
[3] $X_i \perp\!\!\!\perp X_j$ is a short-hand for $X_i \perp\!\!\!\perp X_j \mid \emptyset$ (i.e. $\mathbf{X} = \emptyset$). $X_i \perp\!\!\!\perp \{X_m, X_n\}$ implies $X_i \perp\!\!\!\perp X_m$ and $X_i \perp\!\!\!\perp X_n$.



(a) Directed acyclic graph with 6 variables.



(b) Ancestral graph with 4 variables.



(c) Partial ancestral graph (PAG) for 4 variables and the corresponding Markov equivalence class (MEC) of maximal ancestral graphs (MAGs) it describes.

Fig. 2. Example causal graphs.

Graphs with Confounders. A *mixed graph* is a graph that can contain *directed* edges $\rightarrow$ and *bidirected* edges $\leftrightarrow$ [62]. In the case of graphical Markov models, bidirection $V_i \leftrightarrow V_j$ indicates that there are unmeasured (latent) confounders $V_i \leftarrow U \rightarrow V_j$ (where U may be a single confounder U or represent a network of variables). Graphical Markov models allowing bidirected edges can therefore convey the information that there is a (set of) latent node(s) (i.e. variables not captured in the graph) that influence the (in)dependence relations within the graph. Formally stated, they do not assume *causal sufficiency*—the assumption that there are no unobserved confounders. A mixed graph is an *ancestral graph* if (a) there are no directed cycles, and (b) whenever there is an edge $V_i \leftrightarrow V_j$ then there is no other path from V_i to V_j or from V_j to V_i that is directed. Figure 2b shows an example ancestral graph of the DAG in Fig. 2a where `Age` and `Sex` are unmeasured. This graph implies independencies $\{\texttt{Edu} \perp\!\!\!\perp \texttt{Len},\ \texttt{IQ} \perp\!\!\!\perp \texttt{Len},\ \texttt{Occ} \perp\!\!\!\perp \texttt{IQ} \mid \texttt{Edu},\ \texttt{Occ} \perp\!\!\!\perp \texttt{Len}\}$.

An ancestral graph is said to be maximal—i.e. a *maximal ancestral graph* (MAG) $\mathcal{M}$—if for every pair of nonadjacent nodes (V_i, V_j) there exists a set $\mathbf{W}$ $(V_i, V_j \notin \mathbf{W})$ such that $V_i \perp\!\!\!\perp V_j \mid \mathbf{W}$ [46]. That is, each absent edge corresponds to a conditional independency. Several MAGs can encode the same conditional independencies, forming an MEC of MAGs $[\mathbf{M}]$ described uniquely by a *partial ancestral graph* (PAG) $\mathcal{P}_{[\mathbf{M}]}$. PAG $\mathcal{P}_{[\mathbf{M}]}$ (a) has the same adjacencies as any member of $[\mathbf{M}]$ does; (b) contains a mark of an arrowhead ($<$) iff it is shared by all MAGs in $[\mathbf{M}]$, and; (c) contains a mark of a tail ($-$) iff it is shared by all MAGs in $[\mathbf{M}]$. Arrows may contain a circle ($\circ$) at an end to indicate that this end is in some MAGs in $[\mathbf{M}]$ an arrowhead ($<$) and in some a tail ($-$). Figure 2c depicts an example PAG implying independen-

cies $\{\texttt{Age} \perp\!\!\!\perp \texttt{IQ},\ \texttt{Age} \perp\!\!\!\perp \texttt{Occ} \mid \texttt{Edu},\ \texttt{Edu} \perp\!\!\!\perp \texttt{IQ} \mid \texttt{Occ}\}$, and the MAGs in the Markov equivalence class it describes.

Causal Structure Learning. Many authors have studied the problem of inferring causal models. Causal structure learning (sometimes called *causal search* or *causal* (*structure*) *discovery*) has the goal to infer a causal model from purely observational data, interventional data (e.g. interventions in randomized controlled trials) or a mixture of both [19,59]. These causal models also include causal structures beyond the aforementioned graphical Markov models, such as the popular *structural causal models* (SCMs; also known as *non-parametric structural equation models* [NPSEMs]) [24,25]. We outline some general strategies for causal structure learning from observational data.[4] We describe some causal structure learning algorithms for mixed data in our future work (Sect. 5).

Generally, we distinguish three types of learning methods for causal models: constraint-based, score-based and functional. *Constraint-based* methods use a series of statistical (conditional) independence tests to search for an MEC of graphs that satisfies these independencies [59]. Well-established algorithms in this category include the Peter-Clarke (PC) algorithm [52] that learns a CPDAG $\hat{\mathcal{C}}$[5] from observational data, and the (really) Fast Causal Inference ((r)FCI) algorithm [10,63] for learning a PAG $\hat{\mathcal{P}}$ from observational data. *Score-based* methods aim to maximize a scoring function to find the best graph among candidates [59]. *(Fast) Greedy Equivalence Search* ((f)GES) [8,42] learns a CPDAG $\hat{\mathcal{C}}$ from observational data by iteratively adding edges based on a scoring function, such as the Bayesian Information Criterion (BIC) for continuous variables. *Functional* methods search for *Functional Causal Models* (FCMs)—describing a causal network as a set of functions between variables, e.g. linear relationships and additive noise—by exploiting structural asymmetries in the data when assuming the parametric form of the given FCM (e.g. linearity). An example early FCM is the *Linear Non-Gaussian Acyclic Model* (LiNGAM) [51], which uses statistical analysis to search for a linear SCM when the data is assumed to be non-Gaussian.

Evaluation. Once a causal graph has been generated from the structural learning algorithm, how can it be evaluated? In the case of graphical Markov models, results (e.g. a PAG $\hat{\mathcal{P}}$) are typically evaluated relative to some ground-truth model (e.g. the DAG $\mathcal{G}$ used for data generation). A popular metric is the Structural Hamming distance (SHD) [56] between the two graphs, which is the number of edge insertions, deletions and flips required to change from one graph to another.[6] Other statistics for evaluations are computed based on the confusion matrices of the adjacencies or edges of the two graphs, with the assumption that

[4] For an in-depth overview, we refer the interested reader to [19] and [59].

[5] We use $\mathcal{P}$ as a short-hand for PAG $\mathcal{P}_{[\mathbf{M}]}$ and $\mathcal{C}$ as a short-hand for CPDAG $\mathcal{C}_{[\mathbf{D}]}$.

[6] Note that the SHD was originally defined on CPDAGs, but a similar approach can be applied to other types of graphical Markov models as well.

both graphs include the same features (nodes). Adjacency statistics use the *skeleton graphs* of the two graphs (where all types of edges in a graph are replaced by an undirected edge $-$, thereby reducing each node pair to two types of possible edges), and include statistics such as Adjacency Precision (AP) and Adjacency Recall (AR) [41]. Edge statistics compare the exact edges of the two models (e.g. $\{$*no edge*, —, $\leftarrow$, $\rightarrow\}$ for a CPDAG) to evaluate performance. Arrowhead or tail statistics compare one end of each edge, the head or the tail respectively, and include statistics such as the Arrowhead Precision (AHP) and Arrowhead Recall (AHR) [41]. Note that for each of these statistics, other (computed) statistics of the confusion matrix could be reported instead, such as the number of true positives, the accuracy or the F_1-score.

2.3 Causal Graphs for Model-Agnostic Global Explanation

Sani, Malinsky and Shpitser [48] propose a *global* (type-level) method to summarize the behavior of a black-box model (e.g. a convolutional neural network [CNN]) that uses low-level features $X \in \mathcal{X}$ (e.g. pixels) to predict a label $Y \in \mathcal{Y}$ (e.g. a type of bird). Instead of explaining in the original feature space, the behavior of learned function $\hat{Y} = f(X)$ is explained with the (causal) relationships between some high-level 'human-interpretable' features $Z \in \mathcal{Z}$ (e.g. the belly color, wing pattern & shape, and bill shape & size) and predicted label $\hat{Y} \in \mathcal{Y}$.

In summary, the method works as follows. First, train a black-box $X \rightarrow \hat{Y}$ in a supervised manner with pairs (X, Y) to obtain predictions $\hat{Y}$. Here, $X = (X_1, X_2, \ldots, X_q)$ are the values in input space $\mathcal{X}^q$, $Y \in \mathcal{Y}$ the labels, and $f : \mathcal{X}^q \mapsto \mathcal{Y}$ a black-box function with predictions $\hat{Y} \in \mathcal{Y}$. Next, to explain the global behavior of $f(\cdot)$, estimate a causal *partial ancestral graph* (PAG) $\hat{\mathcal{P}}$ over $V = (Z, \hat{Y})$ with the FCI algorithm [10,63].[7] Other causal estimation methods and graphs (Sect. 2) may be used here instead, but note that PAGs are a good fit for explanation since we can minimize the set of selected variables (PAGs do not assume causal sufficiency) and places where confounding is present are made explicit. $Z = (Z_1, Z_2, \ldots, Z_p)$ $(p \ll q)$ is a set of interpretable features that are given with the data (e.g. additional human-interpretable labels for a bird classifier, or meta-descriptors of the image such as lighting descriptions or when the picture was taken) . The learned PAG forms a family of causal models that indicate (in)dependence relations amongst the selected features $V = (Z, \hat{Y})$, and places for confounding and correlation with the following notation:

- *direction* $Z_i \rightarrow \hat{Y}$ indicates that Z_i causes $\hat{Y}$;
- *bidirection* $Z_i \leftrightarrow \hat{Y}$ indicates Z_i and $\hat{Y}$ share a latent cause $Z_i \leftarrow U \rightarrow \hat{Y}$;
- *partial direction* $Z_i \circ\!\!\rightarrow \hat{Y}$ indicates $Z_i \rightarrow \hat{Y}$, $Z_i \leftrightarrow \hat{Y}$, or both, and;
- *partial bidirection* $Z_i \circ\!\!-\!\!\circ \hat{Y}$ indicates $Z_i \rightarrow \hat{Y}$, $Z_i \leftarrow \hat{Y}$, $Z_i \leftrightarrow \hat{Y}$, or any combination thereof.

[7] The only restriction given to FCI is that $\hat{Y}$ is a non-ancestor of any variable in Z, i.e. all elements in Z can cause each other and $\hat{Y}$ but they cannot be caused by $\hat{Y}$.

The method is evaluated in the computer vision domain with (1) a global PAG for images generated from a known causal diagram; (2) a bird-classification task where a CNN is trained on images with human-annotated features for nine types of bird; (3) a binary pneumonia CNN classifier with annotated features by radiologists, and; (4) a comparison of their techniques' outputs to a sample of local explanations by LIME [44] and SHAP [34].

While promising, the method of Sani, Malinsky and Shpitser [48] has two shortcomings. First, it relies on human-annotated high-level features for explanations. Human annotation is an expensive and time-consuming process. For the image modality, the authors were unable to apply computational methods for inferring high-level features (e.g. using visual object recognition) to address this issue. Second, they do not assess two key properties of global explanations (Sect. 2.1): if the explanatory graph is actually telling of model behavior (*fidelity*) and the generalizability of the explanatory graph (*stability*).

3 Experimental Set-Up

Natural language processing (NLP) has properties similar to computer vision, but has a more well-established body of work we can draw from to computationally infer features. Therefore, we apply Global Causal Analysis (GCA) to a state-of-the-art black-box text classifier on the GoEmotions [14] dataset. Importantly, we also evaluate the stability and faithfulness of the explanatory graph to the black-box model it aims to explain with our novel evaluation method.

Like computer vision, NLP typically uses a large input space $\mathcal{X}^q$ for its tasks. For example, the RoBERTa model [33] studied here converts the input texts into a sequence of tokens from a 30,000 word vocabulary. Using a small set of high-level features would therefore greatly benefit global explanations for NLP. We draw from the extensive literature on computationally inferring linguistic phenomena (from simple methods such as word presence to morphological, syntactic, and semantic information [4]), and illustrate how these features can be related to multiple aspects affecting model behavior (e.g. task-, robustness- and/or fairness-related features). In addition, we take the concept of contrastive explanation and use it to enhance explanatory relevance. Contrastive explanation is applied in local explanation to improve explanatory relevance by setting a one-versus-all class-wise contrast [37,60]. We apply the contrasts globally to limit the edges in the graph to features relevant for a specific class. As a key contribution, we propose a three-step evaluation method that requires no ground-truth causal graph for evaluation—as for black boxes the ground truth is unknown [22]. The method (1) quantifies how faithful GCA is to the model it aims to explain, and (2) evaluates the structural fit and stability of the explanatory graph.

3.1 Data Preprocessing and Model Training

GoEmotions [14] is a dataset containing 58,009 English-language Reddit comments. Each instance comprises a unique identifier, comment text, author, sub-

reddit the comment was posted on, timestamp when it was posted, and a reference to its parent (if applicable). To anonymize the texts, proper names are replaced with a `[NAME]` token and religions with a `[RELIGION]` token [14].

Table 1. GoEmotions dataset descriptives of the high-level sentiment groupings ('positive', 'negative', 'ambiguous', 'neutral') after removing instances with non-unique high-level sentiments, and their corresponding emotion labels.

Sentiment	Emotion label(s)	Train	Test	Validation
Positive	*admiration*, *amusement*, *approval*, *caring*, *desire*, *excitement*, *gratitude*, *joy*, *love*, *optimism*, *pride*, *relief*	15216	1863	1941
Negative	*anger*, *annoyance*, *disappointment*, *disapproval*, *disgust*, *embarrassment*, *fear*, *grief*, *nervousness*, *remorse*, *sadness*	8133	1070	1014
Ambiguous	*confusion*, *curiosity*, *realization*, *surprise*	3858	488	459
Neutral	*neutral*	12823	1606	1592
	+	40030	5027	5006

Table 2. GoEmotions dataset excerpt with Reddit comment `text`, `subreddit` and human-annotated `emotion_labels` and high-level sentiment `label`.

`text`	`subreddit`	`emotion_labels`	`label`
You have a nice bro	pettyrevenge	[admiration]	positive
[NAME] ruled out due to injury. [NAME] starts	rugbyunion	[neutral]	neutral
I would hope the guy is genuine and honest, but	dating	[optimism]	positive
Hi, [NAME]! I thought I would stop by and	atheism	[caring, love, opt.]	positive
Ghost them. It'll drive them crazy and give you	TrueOffMyChest	[neutral]	neutral
Wow, an [NAME] sighting	timberwolves	[surprise]	ambig
i love how the caption implies that the only un	Instagramreality	[amusement]	positive

Each comment in the dataset is labelled by 3–5 human annotators with 27 fine-grained emotions or with the *neutral* label (28 fine-grained labels in total). Of the instances, 83% have one label assigned, 15% two labels, 2% three labels and .2% four or more [14]. All fine-grained labels belong to one of four high-level sentiment labels 'positive' (12 fine-grained labels), 'negative' (11), 'ambiguous' (4) or 'neutral' (1) [14]. For our experiments, we aggregate the 28 fine-grained labels into the four high-level sentiments, where instances that end up with multiple high-level sentiment labels are excluded from further analysis (resulting in 50,063 instances). Furthermore, the data is divided into the predefined 80%–10%–10% train-test-validation splits [14]. Table 1 shows the four high-level sentiment labels, the corresponding fine-grained labels that are grouped under

these labels, and per dataset split the number of instances assigned each sentiment label. Moreover, Table 2 depicts seven example instances, with their corresponding subreddit, human-annotated fine-grained label and inferred high-level sentiment.

For the black-box model, we finetune `DistilRoBERTa-base`: a distilled [47] version of English large language model RoBERTa [33]. We finetune it on the training split with the task to predict labels Y (`label`; 'positive', 'negative', 'neutral', 'ambiguous') based on Reddit comments X (`text`).[8] The most accurate model on the validation split is chosen as the final model. After finetuning, the model achieves an accuracy score of 73.1% (macro-weighted F_1-score 70.3%) on the test split.[9] The black-box model assigned the label 'positive' 1932 times, 'neutral' 1546 times, 'negative' 1002 times, and 'ambiguous' 547 times.

3.2 Procedure: Inferring Features and Global Contrastive Explanation

We extend the human-annotated high-level feature `subreddit` with 22 computationally inferred high-level features. These features serve as an example, to illustrate what type of features one could construct when applying GCA. We group them into three example groupings (so-called *aspects*) relevant for model analysis, where we use GCA to study their effect on model behavior separately, and in conjunction to show how GCA can provide an integrated multi-aspect model behavior analysis. Moreover, we propose to apply contrastive explanation—used to minimize the explanatory factors to the ones distinguishing the actual output from a contrast case in local explanation [37], e.g. by setting a contrast between one class label and all others [60]—to improve GCA's explanatory relevance.

Inferring Features. We study three example aspects of global model behavior: *task*-, *robustness*- and *fairness*-related features. GCA can be applied to study these aspects separately, or an integrated analysis of multiple of these aspects can be performed. Table 3 overviews the human-annotated and inferred features for each aspect, their data type (boolean, categorical, integer, floating point), their description, and if they are human-annotated. What features (and aspects) are relevant in applying GCA depend on the task and application (area).

Task-related features relate to the task at hand. For sentiment analysis, we study the subreddit ('Does sentiment generally differ between subreddits?') and presence of emojis ('Does the model use emojis?') as potential factors. We also include three components traditionally distinguished for word meaning in emotion detection: valence (positiveness-negativeness), arousal (active-passive) and

[8] The model is finetuned for 3 epochs, with a (linear) learning rate of 5×10^{-5}, AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$), a GPU batch size of 16, with seed 42. The Python finetuning uses `Transformers` 4.27.4 with `PyTorch` 2.0.0, `Datasets` 2.11.0 and `Tokenizers` 0.13.2, and is conducted on a Tesla T4 GPU (CUDA 12.0).

[9] The goal is not to get a well-performing model, but to explain model behavior.

dominance (dominant-submissive) [38]. In addition, we include eight basic emotion categorizations by Plutchik [40], with a well-balanced distribution amongst sentiments [39]: anger, anticipation, disgust, fear, joy, sadness, surprise and trust.

Robustness-related features test model sensitivity to noise, distributional artifacts and spurious correlations that may all negatively affect model generalizability [20]. GCA could be used to check if these features are (in)directly related with the black-box output. We include features known to affect earlier sentiment analysis models [32]: the text length (characters, word-level tokens, sentences) and readability (reading grade). In addition, we include features for studying the effect of comment voice (active/passive) and character case (all lowercase).

Table 3. An overview of all included high-level features (grouped by aspect of analysis), their data type, description, and if they are human-annotated.

Aspect	Feature	Type	Description	Annot.
Task	`subreddit`	cat	Subreddit the comment is from	✓
	`has_emoji`	bool	Any emoji in comment*	
	`NRC_valence`	float	Mean valence score†	
	`NRC_arousal`	float	Mean arousal score†	
	`NRC_dominance`	float	Mean dominance score†	
	`NRC_anger`	int	# tokens labelled with anger‡	
	`NRC_anticip.`	int	# tokens labelled with anticipation‡	
	`NRC_disgust`	int	# tokens labelled with disgust‡	
	`NRC_fear`	int	# tokens labelled with fear‡	
	`NRC_joy`	int	# tokens labelled with joy‡	
	`NRC_sadness`	int	# tokens labelled with sadness‡	
	`NRC_surprise`	int	# tokens labelled with surprise‡	
	`NRC_trust`	int	# tokens labelled with trust‡	
Robust.	`len_chr`	int	Length in number of characters	
	`len_tok`	int	Length in number of tokens§	
	`len_snt`	int	Length in number of sentences‖	
	`is_active`	bool	All sentences are in active voice¶	
	`all_lower`	bool	All characters are lowercase	
	`flesch_grade`	float	Flesch-Kincaid reading grade**	
Fairness	`has_name`	bool	Mention of `[NAME]`	
	`has_religion`	bool	Mention of `[RELIGION]`	
	`male_words`	int	Number of male-indicative words††	
	`female_words`	int	Number of female-indicative words††	
	`non-binary_wor.`	int	Number of words indicative of non-binary gender††	

* If any character is a valid emoji according to `emoji`. † Mean human rating of NRC valence/arousal/dominance (VAD) for words [38]. ‡ According to the tokens in NRC *Emotion Lexicon* (EmoLex) [39]. § Total number of tokens over all sentences, according to the `spaCy` tokenizer (`en_core_web_sm`) [27]. ‖ Number of sentences according to `spaCy` [27]. ¶ No passive sentences (`PassivePy` [50]) ** Calculated with `textstat`, where $FKGL = 0.39(\text{words/sentences}) + 11.8(\text{syllables/words}) - 15.59$. †† According to the English *Gender Bias Tool* (GenBit) wordlist [49].

Fairness-related features can be indicative of potential bias with respect to protected attributes [20]. The link between fairness research and NLP explainability has for the most part been limited to local explanations applied to hate speech detection [2], while we focus on global explanations for sentiment analysis. We consider features related to the protected attributes religion (which has been replaced with the `[RELIGION]` token), a person's name (replaced with the `[NAME]` token), and a person's gender (with words indicative of the male, female or non-binary, e.g. waiter versus waitress or herself versus themselves [49]).[10]

Class-Wise Global Contrastive Explanation. Since the relevant features within each aspect may differ for each class label, in addition to describing the overall behavior of the black-box in distinguishing all four classes ($\hat{Y}$) we also perform class-wise global analysis [35]. Contrastive explanation is usually applied to local explanations, where explanatory relevance is increased by setting a one-versus-rest class-wise contrast (e.g. "Why classify X_i as 'positive' rather than 'not positive'?") [37,60]. The explanation can then be limited to factors for distinguishing this class from all others [60]. For example, the task-related feature `NRC_joy` may be indicative in distinguishing the 'positive' label, and less informative for 'neutral'. For each class, we perform a one-versus-rest analysis where the class of interest is encoded as one (1) and the remaining classes as zero (0).[11]

Procedure. For the test split of GoEmotions (containing 5,027 instances) we let the black-box (Sect. 3.1) predict class labels for each instance $\hat{Y} = f(X)$. We then estimate a GCA explanatory graph with the χ^2 independence test, with $\alpha = 0.05$ and the restriction that $\hat{Y}$ may not have a direct arrow towards any variable $Z_i \in Z$ [48] (i.e. $\hat{Y} \not\to Z_i$ and $\hat{Y} \circ\!\!\not\to Z_i$) . We do this separately for each aspect (Z_{task}, Z_{robust}, Z_{fair}) and all aspects combined ($Z = Z_{task} \cup Z_{fair} \cup Z_{robust}$), both for one-versus-rest on each predicted class ($\hat{Y}_{positive}$, $\hat{Y}_{negative}$, $\hat{Y}_{ambiguous}$, $\hat{Y}_{neutral}$) and over all classes ($\hat{Y}$). To work with the χ^2 independence test, non-integer continuous features are binned into 10 equal-sized intervals. In addition, to speed up FCI the `subreddit` feature is re-coded into the 10 most frequently occurring subreddits and the remainder is placed in a category 'other'.

3.3 Evaluation

We propose a three-step evaluation method for applying GCA in practical applications and for our experiments. The method aids domain experts and analysts

[10] We stress that the inferred fairness features here merely serve as an illustration—e.g. of indicators of protected attributes that one can study—, as the actual relevant features depend heavily on the intended application (area) of the ML model.

[11] Note that this same class-contrastive approach to binary encode outputs [60] can also be used to apply GCA to other types of black-boxes, such as ones providing probabilistic class scores, regression analysis and clustering.

in picking a set of variables Z, and indicates the domain fit, faithfulness and stability of the explanatory graph without requiring a ground-truth reference.

1. *Z*-fidelity The fidelity (faithfulness) of the global explanatory model $g(\cdot)$ is typically estimated with the predictive performance of the predicted labels $\hat{Y}' = g(X)$ of the surrogate on the labels $\hat{Y} = f(X)$ of the black-box it aims to explain [13,29,44] (Sect. 2.1). However, as in this case we use PAGs as an explanatory model—capturing (in)dependence relations among variables rather than being predictive models—we cannot compute $\hat{Y}'$. Nevertheless, fidelity is still important because if there is no good fit between the features in Z and $\hat{Y}$, then it might not be telling on how they are related.

To get a general sense of the explanatory power of the set of variables Z on $\hat{Y}$, we fit an ML model that is generally well-performing (predictive accuracy) and has few assumptions on the data: a Random Forest.[12] The Random Forest is merely instrumental in measuring how much information the high-level features contain about the black-box output; other methods can be used here instead. The model is trained with stratified k-fold cross validation (we use 5 folds), and estimates $\hat{Y}'$ on the folds are compared to predictions $\hat{Y}$ of the black-box model. We report the F_1-*score* as an estimate for Z-fidelity (other metrics can be used here instead).[13] Note that this step should be performed before fitting the global (causal) explanatory model, and can even be informative in feature selection.

2. Sanity Checks. Once the GCA explanatory graph has been estimated, we first perform some general sanity checks. We distinguish two types: (1) *automatic* checks ensure that the *background knowledge* imposed on the PAG generation algorithm is indeed in the final GCA explanatory graph (e.g. there are no edges from $\hat{Y}$ to Z,[14] or if a directed edge that is required is indeed there), and; (2) *manual* checks consider the GCA explanatory graph and if the expected relations amongst Z (according to e.g. a domain expert or when the functions are known) are indeed present, regardless of their effect on $\hat{Y}$ (e.g. the number of words and number of characters will be related regardless of whether the model uses these as causal influence or not). In our results, we discuss the manual and automatic sanity checks for three example cases.

3. Structural Fit and Stability. *Modal Value of Edges Existence* (MVEE) is a method for evaluating the quality of generated PAGs when no ground-truth graph is known [26]. It is an extension of *Intersection-Validation* (InterVal) [57], which evaluates the quality of generated CPDAGs $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ learned by n different algorithms on the same dataset. In InterVal, the idea is to generate an *agreement graph*—obtained by taking the strict intersection between graphs

[12] The Random Forest uses default hyperparameters for `scikit-learn` 1.2.2 (100 trees, Gini impurity) with seed 42.

[13] We use F_1-score to account for non-equal distributions of predicted labels (Sect. 3.1).

[14] That is, the outdegree of $\hat{Y}$ for the GCA explanatory graph should always be zero.

$\mathcal{C}$: copy an edge (or the absence thereof) iff it is agreed upon by all graphs, else place a special edge ($\cdots$)—that is then used as proxy for the ground-truth graph. Evaluation metrics of CPDAGs (Sect. 2.2) are then computed relative to the agreement graph: for our experiments we use the SHD [56] towards the agreement graph, called the *Partial Hamming Distance* (PHD) [57].

MVEE [26] takes the notion of agreement graphs from InterVal, but addresses the issue that since there are three more types of edges possible for PAGs, InterVal may be too strict. The edge values for any pair of variables for a CPDAG can take on four values {*no edge*, —, ←, →}, while in a PAG the edges between a pair of nodes can take on seven {*no edge*, ∘–∘, ←, ←∘, →, ∘→, ↔}. Instead of the strict intersection between graphs, MVEE finds a *skeleton* agreement graph using a majority vote from the set of skeletons of a set of input graphs [26]. First, for each PAG $\mathcal{P}$ the skeleton S is calculated (i.e. removing the ends of the edges, such that each pair of nodes can only have an edge value of {*no edge*, —}), and then for these the InterVal method is applied to obtain an agreement graph.

To measure the structural fit & stability (in absence of a ground-truth PAG) we compute the PHD of the GCA explanatory graph with MVEE, where the agreement graph is generated from five explanatory graphs fitted on random 80% subsamples of $(Z, \hat{Y})$ (stability within subsamples; see Sect. 2.1). For MVEE, the PHD indicates the number of edge deletions and additions (↓ lower is better) between the explanatory graph and the agreement graph. Since the number of nodes differs for the aspects in our experiment (and thereby the total possible number of edges between nodes), we report the *relative partial Hamming distance* with the MVEE strategy (*relative* MVEE; ranging from 0 to 1; ↑ higher is better):

$$1 - \frac{\text{MVEE}}{n_{nodes} \times (n_{nodes} - 1)/2} \tag{1}$$

where n_{nodes} is the number of nodes in the GCA explanatory graph and $\binom{n_{nodes}}{2} = n_{nodes} \times (n_{nodes} - 1)/2$ the maximum number of edge values over all nodes.

4 Results and Discussion

We generate GCA explanatory graphs for different sets of high-level features Z (*task*-related, *robustness*-related, *fairness*-related, and all combined) for each predicted label (for each class separately, and all four combined) to explain the behavior of the finetuned `DistilRoBERTa` model on the test split of GoEmotions (Sect. 3.1 & Sect. 3.2).[15] We evaluate their Z-fidelity and structural fit & stability (Sect. 3.3), and discuss three example graphs in detail.[16]

[15] The mean wall-time to generate the GCA explanatory graphs is 0.12s for the fairness aspect (5 features), 0.72s for the robustness aspect (6 features), 2.37s for the task aspect (13 features), and 220.11s for all aspects combined. Wall-time was measured with `causal-learn` 1.3.3 (no depth limit) on Python 3.9.16, on a MacBook Pro with macOS Monterey 12.6.3 (16 GB 2.3 GHz 8-Core Intel Core i9).

[16] Source code available at https://github.com/MarcelRobeer/GlobalCausalAnalysis.

4.1 Quantitative Results

Table 4 summarizes the Z-fidelity scores for our experiments and Table 5 the relative MVEE scores. In summary, we observe the following findings: (1) a high Z-fidelity for the behavior on $\hat{Y}$ (all four aspects), the 'positive' one-versus-rest label (each aspect except fairness), 'neutral' label (each aspect except fairness) and 'negative' label (Z and Z_{task}) shows that for these combinations the selected features Z are very informative of model behavior, and (2) a mean relative MVEE of 0.988 ($SD = 0.016$) over all aspect-label combinations indicates that the method is structurally well-fitting and stable.

Table 4. Z-fidelity (↑ F_1-score, 0–100%) as an estimate for the explanatory power of the chosen variables in Z on $\hat{Y}$. Higher scores indicate that the variables are more telling of $\hat{Y}$, and thereby of the (absence) of edges in the explanatory graph.

Aspect	features	label ($\hat{Y}$)	positive	neutral	negative	ambiguous
all	24	41.95	59.80	32.27	31.79	3.48
fairness	5	21.33	1.11	10.27	0.00	0.00
robustness	6	28.48	40.67	32.43	13.02	8.31
task	13	39.04	57.99	32.69	34.14	6.60

Table 5. The GCA method has a high structural fit & stability (↑ relative MVEE, 0–1) for all combinations of aspects and predicted class labels.

Aspect	label ($\hat{Y}$)	positive	neutral	negative	ambiguous
all	0.98	0.99	0.98	0.99	0.98
fairness	1.00	1.00	0.93	1.00	1.00
robustness	1.00	1.00	1.00	1.00	1.00
task	0.97	0.99	0.99	0.99	0.99

***Z*-fidelity.** Table 4 cross-tabulates the Z-fidelity for the aspects (task-, robustness- & fairness-related features, and all features combined) and the predicted class label behaviors (four contrastive one-versus-rest and one combined). Especially the global behavior of the black-box on the 'positive' class ($\hat{Y}_{positive}$) can be captured well with all features (F_1-score of 59.80%) and just the task-related high-level features (57.99%). Describing its behavior with few features (24 and 13 respectively) is commendable given the black-box model complexity: the `DistilRoBERTa-base` input space is large as the model uses a vocabulary of 30,000 tokens, the model itself consists of 82 million parameters, and the model was pretrained on five datasets totalling 160GB of text [33]. The same

can be said for the Z-fidelity of the features in distinguishing all four labels (label $\hat{Y}$), where all features (41.95%) and the task-related features (39.04%) are able to capture a large portion of the overall behavior (distinguishing four classes). The explanatory graph for each of these combinations should therefore provide valuable insights into what features are (not) related to model behavior.

Moreover, interestingly the model seems to be barely affected by any indicators for protected attributes (fairness aspect) for the one-versus-rest model behavior. This indicates that any arrows towards $\hat{Y}$ in the explanatory graph do not represent a substantial predictive value. The same low Z-fidelity scores hold for the 'ambiguous' label. However, in this case it indicates we have not selected/inferred variables telling of model behavior. Features to study the robustness aspect, however, seem to have a relatively large effect on model behavior—especially across the 'positive' and 'neutral' classes. These scores indicate that robustness-related (unlike fairness-related) features might substantially affect black-box model behavior. Studying the explanatory graphs in more detail can help in distinguishing if these effects are directed or merely due to confounding.

Structural Fit and Stability. The generated GCA explanatory graphs are very stable and have a high structural fit. Table 5 shows the relative MVEE scores (ranging from 0 to 1; higher is better) for all aspect-label combinations. Across all combinations, the mean relative MVEE is 0.988 ($SD = 0.016$). Nine out of 20 combinations are perfectly stable and have a good fit (relative MVEE 1.00), while 11 combinations have a near-one relative MVEE score. Note that the lowest scoring combination (0.93; fairness-related features for the 'neutral' label) has an absolute MVEE of 1 (one edge insertion/deletion to the agreement graph).

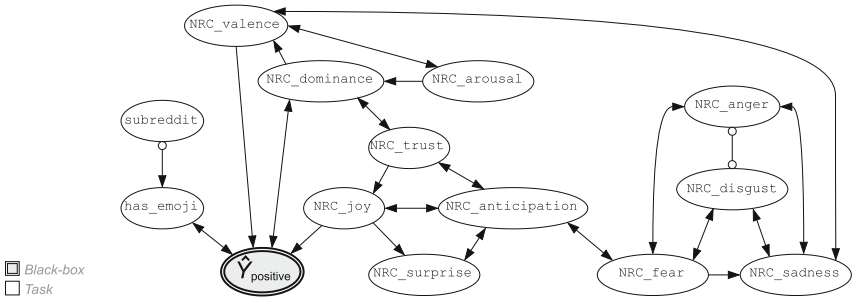


Fig. 3. GCA explanatory graph for task-related features Z_{task} on the 'positive' label $\hat{Y}_{positive}$ (one-versus-rest). The features directly related ($Z_i \rightarrow \hat{Y}_{positive}$) to $\hat{Y}_{positive}$ are `NRC_valence` and `NRC_joy`, while $\hat{Y}_{positive}$ is related due to confounding ($\leftrightarrow$) with `has_emoji` and `NRC_dominance`.

4.2 Empirical Results

We discuss three GCA explanatory graphs in detail, where we consider the three-step evaluation method, the resulting graph and how these interrelate.

Task-Related Features for Label 'Positive'. Figure 3 depicts the explanatory graph for the 13 task-related features for the 'positive' label (one-versus-rest). The features are able to capture the black-box behavior on distinguishing 'positive' ($\hat{Y}_{positive}$) from other classes very well. The model has a Z-fidelity of 57.99 (Table 4) and high structural fit & stability (relative MVEE 0.99; Table 5). $\hat{Y}_{positive}$ does not have any outgoing arrows and thus passes the automatic sanity check.

The class-wise contrast for the 'positive' label does not only quantitatively improve the explanatory relevance, but studying the explanatory graph in detail also provides additional qualitative insights. Two features have a direct effect ($\rightarrow$) on $\hat{Y}_{positive}$: `NRC_joy` (number of words indicative of 'joy' according to NRC EmoLex [39]) and `NRC_valence` (mean human scores of positiveness-negativeness [38]). `NRC_trust` has an indirect effect on $\hat{Y}_{positive}$ through `NRC_joy`, and `NRC_arousal` through `NRC_dominance` and `NRC_valence`. In addition, `has_emoji` (presence of any emojis) and `NRC_dominance` (inferred based on mean human score of dominant-submissive [38]) share unmeasured confounders with $\hat{Y}_{positive}$ ($\leftrightarrow$), possibly providing spurious correlations with model outputs.

Separate from model behavior, we also observe strong interrelatedness between indicators of VAD scores (`NRC_valence`, `-arousal` and `-dominance`), between emotions with a positive sentiment focus (`NRC_trust`, `-joy`, `-anticip.`), and between emotions with a relatively negative sentiment (`NRC_anger`, `-disgust`, `-fear` and `-sadness`). These subgroups largely correspond to the positive and negative sentiment emotions [40], indicating expected behavior from the manual sanity check. Important to note, however, is that VAD scores are usually considered independent aspects of emotion [38]. The subgraph with emotions with positive sentiment is related through `NRC_joy` to the label 'positive' ($\hat{Y}_{positive}$), while the subgraph with emotions with negative sentiment is not on a causal path to $\hat{Y}_{positive}$. This could be a good indicator that the black-box indeed uses positive task-related features for its classification (e.g. indicators of joy), which may enhance trust that the model will generalize well and is relatively robust.

Robustness-Related Features for Label 'Neutral'. We also study the robustness aspect for distinguishing the predicted 'neutral' label from all other classes. Figure 4 shows the GCA explanatory graph with six robustness features and $\hat{Y}_{neutral}$. The graph has a Z-fidelity of 32.43 and a relative MVEE of 1.00. It passes the automatic sanity check that $\hat{Y}_{neutral} \not\rightarrow Z_i$ and $\hat{Y}_{neutral} \circ\!\!\not\rightarrow Z_i$.

Two features are independent from model behavior ($\hat{Y}_{neutral}$) and from all other robustness-related features: `all_lower` (if all characters are lowercase) and `is_active` (if all sentences in the comment are in active voice). The length in number of sentences (integer `len_snt`) is directly indicative of the 'neutral' label.

The number of characters (`len_chr`), number of tokens (`len_tok`), the number of sentences (`len_snt`) and the Flesch-Kincaid reading grade (`flesch_grade`; calculated based on number of syllables, words and sentences) form a clique. This is as expected, as the lengths all positively correlate (longer comments consist of more characters, tokens and sentences) and the reading grade is functionally related to the lengths. Thus, the graph passes the manual sanity check.

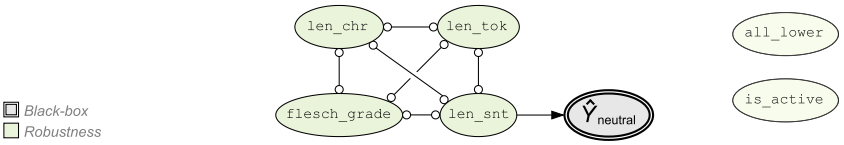


Fig. 4. GCA explanatory graph for robustness-related features Z_{robust} on $\hat{Y}_{neutral}$. `len_snt` is directly related ($Z_i \rightarrow \hat{Y}_{neutral}$) to $\hat{Y}_{neutral}$. `all_lower` and `is_active` are independent from $\hat{Y}_{neutral}$ and all other features in Z_{robust}.

FCI cannot distinguish the direction of this relationship ($Z_i \rightarrow Z_j$, $Z_i \leftrightarrow Z_j$ or $Z_i \leftarrow Z_j$) and if there are any confounders. It assigns a partial bidirection relationship between all four variables $Z_i \circ\!\!-\!\!\circ Z_j$. Including additional robustness features or combining the analysis with other aspects (e.g. fairness-related or task-related features) may help in clarifying these relations, and to see how strong the connection is between the robustness features and the 'neutral' label.

Task-, Fairness- and Robustness-Related Features Combined. Figure 5 shows the GCA explanatory graph over all aspects combined, for the whole black-box model behavior (distinguishing all four labels). The graph scores a Z-fidelity of 41.95 and a relative MVEE (structural fit & stability) of 0.98. The graph passes the sanity check that $\hat{Y}$ has no outgoing directed arrows. To foster multi-aspect analysis, the features related to different aspects in the explanatory graph are color-coded, and the node $\hat{Y}$ is shown in gray with a double bolded border.

Three things immediately stand out. First, the model behavior is directly affected by the mean arousal score (inferred based on the NRC VAD Lexicon [38]), the mean dominance score (also inferred using [38]) and the sentence length. Second, `male_words` (fairness-related), `has_name` (fairness), `female_words` (fairness) and `subreddit` (task-related) form a subgraph, with behavior separate from the behavior of $\hat{Y}$. Male- and female-indicative words, and the presence of the `[NAME]` token are indicative of the `subreddit`. Third, several task-, robustness- and fairness-related features are unconnected in the explanatory graph: `has_emoji`, `is_active`, `all_lower`, `non-binary_words` and `has_religion`.

Many features either share a common confounder or have a directed relationship in line with expected behavior (as studied in detail in Fig. 3 and Fig. 4). For

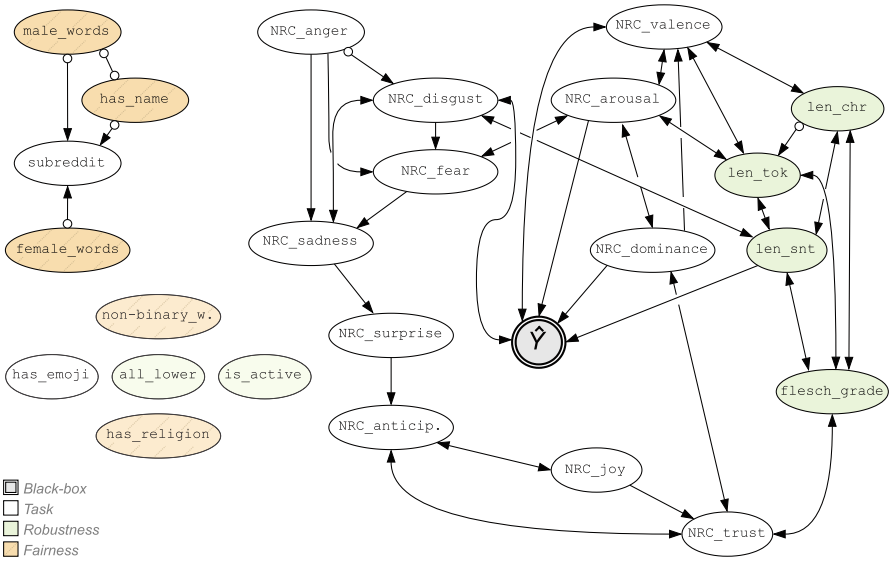


Fig. 5. GCA explanatory graph for all features $Z = Z_{task} \cup Z_{fair} \cup Z_{robust}$ (24 high-level features) on the overall black-box predictive behavior $\hat{Y}$ (four classes). The features directly related ($Z_i \rightarrow \hat{Y}$) to $\hat{Y}$ are `NRC_arousal`, `NRC_dominance` and `len_snt`. `NRC_disgust` and `NRC_valence` are directly related due to confounding ($\leftrightarrow$) with $\hat{Y}$. `has_emoji`, `is_active`, `all_lower`, `non-binary_words` and `has_religion` are independent from $\hat{Y}$ and all other features in Z.

task-related features, we observe strong connections with negative emotion features (`NRC_anger`, `-disgust`, `-sadness` and `-fear`), the VAD components of emotions (`NRC_valence`, `-arousal` and `-dominance`), and the three features catered towards positive emotions (`NRC_joy`, `-anticip.` and `-trust`). For robustness, the lengths in characters (`len_chr`), tokens (`len_tok`) and sentences (`len_snt`) are correlated as expected, and the reading grade (`flesch_grade`) also functionally relates to the instance length. Moreover, we observe how many task-related and robustness features share confounders: {`NRC_disgust` $\leftrightarrow$ `len_snt`, `NRC_valence` $\leftrightarrow$ `len_chr`, `NRC_arousal` $\leftrightarrow$ `len_tok`, `NRC_trust` $\leftrightarrow$ `flesch_grade`}.

5 Conclusion and Future Work

We presented Global Causal Analysis (GCA) as a method for global model-agnostic explanations for text classification, explaining model behavior with a causal explanatory graph. The GCA explanatory graph interprets black-box functioning over a dataset with high-level human-interpretable features (either over all classes or contrastively by setting a one-versus-rest class-wise contrast), revealing if and how these features affect each other and the black-box output.

GCA is a strong addition to global explanation methods. GCA can distinguish causal relations from effects due to (spurious) correlations, and explicitly

shows where latent confounders are. The explanatory graph not only shows relations with the model output, but also between the high-level features themselves. We show how these features can be inferred computationally, avoiding costly human annotation to explain model behavior at a higher level of abstraction.

The three-step evaluation method that is a key part of GCA (1. Z-fidelity; 2. Sanity checks; 3. Structural fit & stability) proves useful in quantitatively and empirically assessing (a) the explanatory power of the selected high-level features and (b) the quality of the explanatory graph. GCA can summarize large parts of model behavior with few human-interpretable features, is structurally stable and well-fitting, and has high explanatory relevance with its ability to explain behavior over all classes or with class-wise one-versus-rest contrasts.

We consider three interesting avenues for future research. The first is using *global interventions* to provide a stronger link with causality research. NLP offers several computational approaches (e.g. [20,45,53]) to intervene upon specific attributes (e.g. for gender replacing all female names with male ones). GCA can then be applied with a mixed causal learning method (e.g. [11,15,55]), and then further enhanced by estimates of the effect sizes and directions (positive/negative) of features (e.g. *Average Causal Effect* [7]). Second, we want to study applying causal high-level feature explanations to *local explanations*. We could benefit from the wealth of desidarata, definitions and methods for counterfactuals and contrastive explanation at the local level (see [9]). Moreover, we could simplify local explanations by summarizing behavior with high-level features, study them from multiple aspects, and also explore locally explaining at various levels of linguistic structures (e.g. explanations at the phrase or word level). Third, we want to apply GCAs in practical contexts to perform *human evaluations* with domain experts, model developers and model end-users.

Acknowledgements. This study has been partially supported by the Dutch National Police. The authors would like to thank Elize Herrewijnen and Gizem Sogancioglu for their valuable feedback on earlier versions of this work.

References

1. Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. J. Royal Stat. Soc. Ser. B: Stat. Methodol. **82**(4), 1059–1086 (2020). https://doi.org/10.1111/rssb.12377
2. Balkir, E., Kiritchenko, S., Nejadgholi, I., Fraser, K.: Challenges in applying explainability methods to improve the fairness of NLP models. In: Proceedings 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022), pp. 80–92. ACL, Seattle, U.S.A. (2022). https://aclanthology.org/2022.trustnlp-1.8
3. Bastani, O., Kim, C., Bastani, H.: Interpreting Blackbox models via model extraction. CoRR abs/1705.08504 (2017)
4. Belinkov, Y., Glass, J.: Analysis methods in neural language processing: a survey. Trans. Assoc. Comput. Linguist. **7**, 49–72 (2019). https://doi.org/10.1162/tacl_a_00254

5. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. J. Artif. Intell. Res. **70**, 245–317 (2021). https://doi.org/10.1613/jair.1.12228
6. Carvalho, D.V., Pereira, E.M., Cardoso, J.S.: Machine learning interpretability: a survey on methods and metrics. Electronics **8**(8), 832 (2019). https://doi.org/10.3390/electronics8080832
7. Chattopadhyay, A., Manupriya, P., Sarkar, A., Balasubramanian, V.N.: Neural network attributions: a causal perspective. In: International Conference on Machine Learning, pp. 981–990. PMLR (2019)
8. Chickering, D.M.: Optimal structure identification with greedy search. J. Mach. Learn. Res. **3**, 507–554 (2002)
9. Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C., Jorge, J.: Counterfactuals and causability in explainable artificial intelligence: theory, algorithms, and applications. Inf. Fusion **81**, 59–83 (2022). https://doi.org/10.1016/j.inffus.2021.11.003
10. Colombo, D., Maathuis, M.H., Kalisch, M., Richardson, T.S.: Learning high-dimensional directed acyclic graphs with latent and selection variables. Ann. Stat. **40**(1), 294–321 (2012). https://doi.org/10.1214/11-AOS940
11. Cooper, G.F., Yoo, C.: Causal discovery from a mixture of experimental and observational data. In: Proceedings of the 15th Conf. on Uncertainty in Artificial Intelligence, pp. 116–125. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1999)
12. Craven, M.W., Shavlik, J.W.: Using sampling and queries to extract rules from trained neural networks. In: Eleventh International Conference on Machine Learning (ICML), Proceedings, pp. 37–45 (1994). https://dl.acm.org/doi/10.5555/3091574.3091580
13. Craven, M.W., Shavlik, J.W.: extracting tree-structured representations of trained neural networks. In: Advances in Neural Information Processing Systems (NIPS), vol. 8, pp. 24–30 (1996)
14. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: a dataset of fine-grained emotions. In: 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4040–4054. Online (2020). https://doi.org/10.18653/v1/2020.acl-main.372
15. Eaton, D., Murphy, K.: Exact Bayesian structure learning from uncertain interventions. In: Artificial Intelligence and Statistics, pp. 107–114. PMLR (2007)
16. Fisher, A., Rudin, C., Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J. Mach. Learn. Res. **20**, 1–81 (2019)
17. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Ann. Stat. **29**(5), 1189–1232 (2001). https://doi.org/10.1214/aos/1013203451
18. Galhotra, S., Pradhan, R., Salimi, B.: Explaining black-box algorithms using probabilistic contrastive counterfactuals. In: Proceedings of the 2021 International Conference on Management of Data, pp. 577–590. SIGMOD 2021, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3448016.3458455
19. Glymour, C., Zhang, K., Spirtes, P.: Review of causal discovery methods based on graphical models. Front. Genet. **10**, 524 (2019)
20. Goel, K., Rajani, N.F., Vig, J., Taschdjian, Z., Bansal, M., Ré, C.: Robustness gym: unifying the NLP evaluation landscape. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technologies: Demonstrations, pp. 42–55. ACL, Online (2021). https://doi.org/10.18653/v1/2021.naacl-demos.6

21. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. J. Comput. Graph. Stat. **24**(1), 44–65 (2015). https://doi.org/10.1080/10618600.2014.907095
22. Guidotti, R.: Evaluating local explanation methods on ground truth. Artif. Intell. **291**, 103428 (2021). https://doi.org/10.1016/j.artint.2020.103428
23. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018). https://doi.org/10.1145/3236009
24. Halpern, J.Y.: A modification of the Halpern-Pearl definition of causality. In: International Joint Conference on Artificial Intelligence (IJCAI), pp. 3022–3033 (2015)
25. Halpern, J.Y., Pearl, J.: Causes and explanations: a structural-model approach - Part I: Causes. In: 17th Conference on Uncertainy in Artificial Intelligence, Proceedings, pp. 194–202. Morgan, San Francisco, CA (2001). https://doi.org/10.1093/bjps/axi147
26. Handhayani, T., Cussens, J.: Kernel-based approach for learning causal graphs from mixed data. In: Jaeger, M., Nielsen, T.D. (eds.) Proceedings of the 10th International Conference on Probabilistic Graphical Models. Proceedings of the Machine Learning Research, vol. 138, pp. 221–232. PMLR (2020). https://proceedings.mlr.press/v138/handhayani20a.html
27. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: industrial-strength natural language processing in Python (2020). https://doi.org/10.5281/zenodo.1212303
28. Hooker, G.: Discovering additive structure in black box functions. In: Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2004)
29. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: how should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4198–4205. ACL, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.386
30. Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J.: Interpretable & explorable approximations of black box models. In: KDD 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning (2017)
31. Lakkaraju, H., Arsov, N., Bastani, O.: Robust and stable black box explanations. In: Proceedings of the 37th International Conference on Machine Learning (ICML). JMLR.org (2020). https://proceedings.mlr.press/v119/lakkaraju20a/lakkaraju20a.pdf
32. Li, L., Goh, T.T., Jin, D.: How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. Neural Comput. Appl. **32**(9), 4387–4415 (2018). https://doi.org/10.1007/s00521-018-3865-7
33. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019)
34. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems, pp. 4765–4774 (2017)
35. Madsen, A., Reddy, S., Chandar, S.: Post-hoc interpretability for neural NLP: a survey. ACM Comput. Surv. **55**(8), 1–42 (2022). https://doi.org/10.1145/3546577
36. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2019). https://doi.org/10.1016/j.artint.2018.07.007
37. Miller, T.: Contrastive explanation: a structural-model approach. Knowl. Eng. Rev. **36**, e14 (2021). https://doi.org/10.1017/s0269888921000102

38. Mohammad, S.: Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 174–184. ACL, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1017
39. Mohammad, S.M., Turney, P.D.: Crowdsourcing a Word-Emotion Association Lexicon. Comput. Intell. **29**(3), 436–465 (2013)
40. Plutchik, R.: A general psychoevolutionary theory of emotion. In: Theories of Emotion, pp. 3–33. Elsevier (1980). https://doi.org/10.1016/b978-0-12-558701-3.50007-7
41. Raghu, V.K., Poon, A., Benos, P.V.: Evaluation of causal structure learning methods on mixed data types. In: Le, T.D., Zhang, K., Kıcıman, E., Hyvärinen, A., Liu, L. (eds.) Proceedings of the 2018 ACM SIGKDD Workshop on Causal Disocvery. Proceedings of the Machine Learning Research, vol. 92, pp. 48–65. PMLR (2018). https://proceedings.mlr.press/v92/raghu18a.html
42. Ramsey, J., Glymour, M., Sanchez-Romero, R., Glymour, C.: A million variables and more: the fast greedy equivalence search algorithm for learning high-dimensional graphical causal models, with an application to functional magnetic resonance images. Int. J. Data Sci. Anal. **3**(2), 121–129 (2017). https://doi.org/10.1007/s41060-016-0032-z
43. Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. In: 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016), pp. 91–95 (2016)
44. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should i trust you?": explaining the predictions of any classifier. In: 22nd ACM SIGKDD Intl. Conf. on Knowledge Discovery in Data Mining (KDD 2016), Proceedings, pp. 1135–1144 (2016). https://doi.org/10.1145/2939672.2939778
45. Ribeiro, M.T., Wu, T., Guestrin, C., Singh, S.: Beyond accuracy: behavioral testing of NLP models with CheckList. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4902–4912. ACL, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.442
46. Richardson, T., Spirtes, P.: Ancestral graph Markov models. Ann. Stat. **30**(4), 962–1030 (2002). http://www.jstor.org/stable/1558693
47. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) (2019)
48. Sani, N., Malinsky, D., Shpitser, I.: Explaining the behavior of black-box prediction algorithms with causal learning. CoRR abs/2006.02482 (2020)
49. Sengupta, K., Maher, R., Groves, D., Olieman, C.: GenBiT: measure and mitigate gender bias in language datasets. Microsoft J. Appl. Res. **16**, 63–71 (2021)
50. Sepehri, A., Markowitz, D.M., Mir, M.: PassivePy: a tool to automatically identify passive voice in big text data (2022). https://doi.org/10.31234/osf.io/bwp3t
51. Shimizu, S., Hoyer, P.O., Hyvärinen, A., Kerminen, A.: A linear non-gaussian acyclic model for causal discovery. J. Mach. Learn. Res. **7**(72), 2003–2030 (2006). http://jmlr.org/papers/v7/shimizu06a.html
52. Spirtes, P., Glymour, C.N., Scheines, R., Heckerman, D.: Causation, Prediction, and Search. MIT Press, Cambridge (2000)
53. Tan, S., Joty, S., Baxter, K., Taeihagh, A., Bennett, G.A., Kan, M.Y.: Reliability testing for natural language processing systems. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL) and the 11th International Joint Conference on Natural Language Processing, pp. 4153–4169. ACL, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.321

54. Tan, S., Caruana, R., Hooker, G., Lou, Y.: Distill-and-compare: auditing black-box models using transparent model distillation. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, pp. 303–310. AIES 2018, Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3278721.3278725
55. Tian, J., Pearl, J.: Causal discovery from changes. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, pp. 512–521. UAI 2001, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
56. Tsamardinos, I., Brown, L.E., Aliferis, C.F.: The max-min hill-climbing Bayesian network structure learning algorithm. Mach. Learn. **65**(1), 31–78 (2006). https://doi.org/10.1007/s10994-006-6889-7
57. Viinikka, J., Eggeling, R., Koivisto, M.: Intersection-Validation: a method for evaluating structure learning without ground truth. In: Storkey, A., Perez-Cruz, F. (eds.) Proceedings of the 21st International Conference on Artificial Intelligence and Statistics. Proceedings of the Machine Learning Research, vol. 84, pp. 1570–1578. PMLR (2018). https://proceedings.mlr.press/v84/viinikka18a.html
58. Vilone, G., Longo, L.: A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods. Front. Artif. Intell. **4**, 717899 (2021). https://doi.org/10.3389/frai.2021.717899
59. Vowels, M.J., Camgoz, N.C., Bowden, R.: D'Ya like DAGs? a survey on structure learning and causal discovery. ACM Comput. Surv. **55**(4), 1–36 (2022). https://doi.org/10.1145/3527154
60. van der Waa, J., Robeer, M., van Diggelen, J., Neerincx, M., Brinkhuis, M.: Contrastive explanations with local Foil Trees. In: 2018 Workshop on Human Interpretability in Machine Learning (WHI) (2018)
61. Woodward, J.: Making Things Happen. Oxford University Press, Oxford (2004)
62. Zhang, J.: Causal reasoning with ancestral graphs. J. Mach. Learn. Res. **9**(47), 1437–1474 (2008). http://jmlr.org/papers/v9/zhang08a.html
63. Zhang, J.: On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. Artif. Intell. **172**(16), 1873–1896 (2008). https://doi.org/10.1016/j.artint.2008.08.001
64. Zhao, Q., Hastie, T.: Causal interpretations of black-box models. J. Bus. Econ. Stat. **39**(1), 272–281 (2019). https://doi.org/10.1080/07350015.2019.1624293