Causal Graphs and Fairness in Machine Learning: Addressing Practical Challenges in Causal Fairness Evaluation

LEA COHAUSZ*, JAKOB KAPPENBERGER, and HEINER STUCKENSCHMIDT, University of Mannheim, Germany

Background: With the discussion of fairness in Machine Learning (ML) gaining traction in recent years, the idea of viewing fairness through the causal lens has become prominent. The main idea behind this is that by looking at the causal structure underlying the data used for an ML model, we can see and evaluate more concisely which influences of the sensitive variables on the target variable are problematic and how they are problematic. Doing so allows not only a nuanced view of fairness and an informed choice of fairness measures but also more targeted approaches (such as path-specific bias mitigation) to handle fairness issues.

Objectives: Mainly, two important points have hindered the practical use of the causal lens and causality-based bias mitigation. First, a classification of different graphical structures with different fairness implications involving a sensitive variable and a target variable is still missing, as is a discussion of how different contexts can shape our evaluation of fairness. Second, the construction of such graphical models is not trivial and error-prone. However, recent work showed that combining background knowledge and data-driven network structure learning may lead to more accurate graphs. In this work, we attempt to address and tackle these two practical shortcomings.

Methods: Our first contribution is a classification and discussion of causal structures with different fairness implications and how contexts shape our assessment. Our second contribution is an advancement in learning more accurate graphs by adapting structure learning algorithms, and a detailed evaluation of graph correctness and subsequent fairness implications.

Results: We show that when including background knowledge naturally available in fairness settings, graph learning becomes more accurate, which also has positive implications for accurate fairness assessments.

Conclusions: Our work may pave the way for a broader adoption of causal ML fairness by providing concrete suggestions about the implications of causal structures and contexts, and learning more accurate graphs. We also address current limitations and highlight the need for stakeholder inclusion.

JAIR Track: Fairness and Bias in AI

JAIR Associate Editor: Virginia Dignum

JAIR Reference Format:
Lea Cohausz, Jakob Kappenberger, and Heiner Stuckenschmidt. 2025. Causal Graphs and Fairness in Machine Learning: Addressing Practical Challenges in Causal Fairness Evaluation. *Journal of Artificial Intelligence Research* 84, Article 15 (October 2025), 34 pages. DOI: 10.1613/jair.1.19082

1 Introduction

As the topic of fairness in Machine Learning (ML) has gained prominence, so has the notion that an evaluation of fairness necessarily requires a look at the causal connections underlying the data-generating process and at the context in which a particular ML-based system is deployed (Makhlouf et al., 2024). Kilbertus et al. (2017) showed

*Corresponding Author.

Authors' Contact Information: Lea Cohausz, lea.cohausz@uni-mannheim.de, ORCID: 0000-0002-6164-3988; Jakob Kappenberger, ORCID: 0000-0003-1290-0199, jakob.kappenberger@uni-mannheim.de; Heiner Stuckenschmidt, ORCID: 0000-0002-0209-3859, heiner.stuckenschmidt@uni-mannheim.de, University of Mannheim, Mannheim, Baden-Wuerttemberg, Germany.

DOI: 10.1613/jair.1.19082

that the common fairness measures employed in ML (e.g., Demographic Parity, Equalized Odds, etc.) cannot be used sufficiently and sensibly unless the variables and their connections are viewed through the causal network. Chiappa and Isaac (2019) investigated how looking at the Causal Bayesian Network (CBN) underlying the data can help us determine and handle potential fairness problems, although their approach remains theoretical with no broad practical evaluation. As fairness in ML usually relates to the absence of an influence of sensitive variables (e.g., age, gender, ethnicity) on the decision we try to make, looking at the relationship between variables through CBNs is helpful. A CBN, or rather the Directed Acyclic Graph (DAG), i.e., the graphical part of a CBN, can show whether a sensitive variable is relevant at all and how it is connected (directly, indirectly through other variables, or both). Thus, we can make informed decisions of whether a sensitive variable has a problematic influence and through which paths, even allowing path-specific bias mitigation (Chiappa and Isaac, 2019; Madras et al., 2019).

Although Chiappa and Isaac (2019) showed many different networks and how different connections have different implications, the concrete implications regarding what different DAG structures may mean for ML are still missing. This aspect is relevant, however, as some nuanced implications for ML are not obvious by purely looking at the data or even the DAG. For example, a sensitive variable may not be correlated with the target (which appears unproblematic initially) but with other variables that are correlated with the target. The inclusion of these variables can then lead to fairness concerns, as ML models will likely place weight on the variables correlated with the sensitive variable at the very least. Furthermore, especially for settings in which the sensitive variable may have problematic influences in some way, a detailed look at how the different contexts (application area, aim of the model, history of the data) of ML models may determine our decisions on whether data used for the model is fair even when the underlying structure is the same, is mostly still missing (Cohausz et al., 2024; Selbst et al., 2019). Hence, the first objective of this paper is to provide a) a classification of different causal structures and their fairness implications when using the generated data for ML, and b) a discussion of how different contexts shape our decision-making even for the same causal connections.

A major hurdle of using the causal lens, in general, is that the CBNs underlying real-life data are usually unknown. While the DAGs can be constructed using expert background knowledge, data-driven structure learning methods, or a combination of both, each of these methods has downsides. Expert background knowledge (i.e., knowledge about the existence or prohibition of certain relationships between variables) requires an intimate understanding of all relationships, which is unrealistic in many settings. Data-driven structure learning is known to be unreliable, especially on realistic data (Constantinou et al., 2021). Binkytė-Sadauskienė et al. (2022) showed that using different structure learning methods on the same dataset leads to very different results that also have different fairness implications (e.g., a variable is not connected according to one method and is connected according to another one). Combining expert-driven and data-driven methods seems to be promising (Constantinou et al., 2023) but has been researched to a very limited extent.

In the fairness context, some background knowledge about the data-generating mechanism is necessarily known a priori: namely, we know that, usually, sensitive variables cannot be influenced by non-sensitive variables (e.g., gender is not influenced by passing a course) and that the target variable cannot influence other variables. With this automatic background information, we can adapt the data-driven structure learning methods to produce more accurate results. Hence, the second objective of this paper is an adaptation of structure learning methods in the fairness context and an evaluation of the accuracy with which the CBNs are learned when background information is available compared to when it is not. Concretely, we constructed the following research questions (RQs) to evaluate our second contribution:

RQ1: Does the correctness of the learned graphs increase when background information is present? With this research question, we aim to check whether the graphs learned using background information are closer to the ground-truth graphs than those learned without information. While we expect this research question to be answered affirmatively, in the interest of the self-containment of this work and due to the remaining research question building on a positive result, we still need it. Past research on the helpfulness of background knowledge (Constantinou et al., 2023) arrived at mixed results for this question.

RQ2a: Do different methods benefit differently from background information? With this research question, we aim to check whether background information benefits methods from different structure learning families differently.

RQ2b: Do different data sizes benefit differently from background information? With this research question, we aim to check whether background information is more helpful for larger or smaller numbers of data instances. Past research has clearly shown that structure learning methods struggle with less data being available (Constantinou et al., 2021); hence, background information may be particularly valuable in those settings.

RQ3: Are mistakes introduced through providing background information? With this research question, we aim to check whether some previously correct edges are now missing or some previously correctly not included edges are now wrongly included when background information is provided versus when background information is not provided. The results can help us understand whether there is also a downside to using background information.

RQ4: Do variables get classified more correctly when background information is available? With this research question, we aim to check whether the sensitive variables that can be classified according to the structures of which they are part (defined later in the paper as part of our first contribution) are classified accordingly in the learned graph and whether this happens more accurately when including background information. In doing so, we check that the relevant structures have been correctly learned to the extent that the classification is possible.

RQ5: Do the paths through which problematic influences get propagated get detected more reliably when background information is available? With this research question, we aim to check whether all paths through which a problematic variable influences the target are detected (and not just some that would be sufficient for **RQ4**) and whether this happens more accurately when including background information. We focus on those structures that we later define as problematic, as the handling of the variables involved in them is path-specific.

RQ6: To what extent does the estimation of causal fairness change when background information is available? With this research question, we aim to investigate whether the introduction of background knowledge and the corresponding effect on learning DAGs leads to estimates of causal fairness that are closer to the estimates achieved when using the ground-truth DAGs compared to when not using background information.

As a result, our paper is structured as follows:

- We provide an introduction to fairness in ML as well as to CBNs, how to learn them, and how to think about fairness using DAGs in Section 2.
- We introduce our classification of different structures in Section 3.
- We demonstrate how contexts may shape our decisions in Section 4.
- We discuss how different methods for learning DAGs can be adapted to include background knowledge in Section 5. The implementation of the adapted structure learning methods, as well as of the automatic detection of the structures according to Section 3, can be found online: https://github.com/lea-cohausz/causalfair.
- We evaluate the adapted methods according to our research questions by looking at how well they recover networks for which we know the ground truth and how well the different classes are subsequently found. Our evaluation setup is described in Section 6, and the results are presented in Section 7. We also show that our methods are applicable to real-life data as well in Section 8.
- We discuss the limitations and implications of our paper and point out future work in Section 9.

Proceeding as mentioned, we hope to make the causal view on fairness and the subsequent handling of fairness issues more readily and practically applicable.

2 Background on Fairness and Causality

We will start by introducing important concepts relevant to this work. Concretely, we will introduce and critically discuss ML fairness, introduce causal graphs and how to construct them, and how they can be used for fairness assessments.

2.1 Fairness

Fairness is a highly contested theoretical construct, subject to multiple interpretations across diverse social and scientific domains (Jacobs and Wallach, 2021). For example, while some define fairness as providing equal opportunities for all, others argue that it requires either equal treatment or equitable outcomes—which may then involve redistributive measures—to truly promote fairness (Dolata et al., 2022). With the rapid proliferation of ML in various applications, concerns regarding the fairness of decisions made by these systems have become increasingly prominent (Gerdon et al., 2022). In the quest for more just and equitable ML-based decision-making, researchers have proposed a wide range of operationalizations of fairness. Many of these approaches rely on the notion of *algorithmic bias*, which denotes systematically varying model performance for different individuals affected by the decisions of a ML-based system, resulting in disadvantages for some of them (Mitchell et al., 2021).

In general, measures proposed to quantify algorithmic bias either aim to produce similar predictions for similar individuals (i.e., individual fairness) or, more commonly, for different social groups (i.e., group fairness) (Mehrabi et al., 2022). To meaningfully cluster individuals into such groups, these measures rely on sensitive variables, which are typically demographic characteristics (Castelnovo et al., 2022; Baker et al., 2023). Demographic features include, among others, attributes such as gender, age, and socioeconomic status. Baker et al. (2023) define demographic features as variables that remain immutable within the context of the ML application. For instance, in an educational setting, features like gender are considered demographic and potentially sensitive because they cannot be altered within the setting (unlike, for example, educational attainment).

In the remainder of this paper, we will focus on the following most often used measures of group-specific algorithmic bias, where $\hat{Y}$ represents the binary prediction of Y, Y is the binary ground truth for a given data record, and A is a binary sensitive (e.g., demographic) attribute.[1] Later, we will also discuss causal fairness notions.

- **Demographic Parity (DP)** codifies the notion that predictions should be independent of the sensitive attribute. Thus, it requires that the positive prediction rate is equal between all groups across the sensitive attribute (Dwork et al., 2012), i.e.:

$$P(\hat{Y} = 1 | A = 0) = P(\hat{Y} = 1 | A = 1) \tag{1}$$

  If the values are not roughly equal, a bias has been detected according to DP.
- **Equalized Odds (EO)** takes the ground truth into account and defines fairness as the equality of false positives and false negative rates across the groups examined (Hardt et al., 2016), i.e.:

$$P(\hat{Y} = 1 | A = 0, Y = y) = P(\hat{Y} = 1 | A = 1, Y = y) \tag{2}$$

  Again, if the values are not roughly equal, a bias has been detected according to EO.

DP belongs to the group of *independence* measures that only consider the distribution of features and predictions. In contrast, the *separation* measure EO also incorporates the ground truth Y (Räz, 2021). Different fairness measures require different degrees of absence of statistical relationships between a sensitive feature and the prediction of the target variable. If something contradicts this absence, one usually employs a bias mitigation technique (Stinar and Bosch, 2022; Deho et al., 2022). Bias mitigation tries to eradicate a specified relationship between the sensitive feature and the target either in pre-processing (e.g., by trying to de-correlate features) or post-processing (e.g., by

[1] All the following definitions can be and already have been extended for the non-binary case.

shifting the threshold that determines the prediction for a demographic group) or within the ML-method (e.g., by also reducing a loss term that includes the relevant fairness measures) (Stinar and Bosch, 2022; Deho et al., 2022).

It is worth noting that framing fairness exclusively in terms of algorithmic bias has attracted substantial criticism in the literature. In particular, Selbst et al. (2019) contend that many researchers neglect the broader context in which an ML system operates, focusing narrowly on the model itself as well as its inputs and outputs. When using measures of algorithmic bias to quantify fairness, it is essential to consider both the context in which the dataset was generated and the application domain where the ML system is deployed (Cohausz et al., 2024). For example, suppose that historically biased training data is used to inform college admission decisions. In that case, some group fairness measures (e.g., EO) may fail to detect fairness disparities because—even if admission decisions are skewed toward a particular social group (thus replicating the training data distribution)—no bias is flagged by these measures.

These criticisms gave rise to *causal fairness* as an alternative paradigm of evaluating the fairness of ML, which utilizes concepts of causal modeling (Makhlouf et al., 2024). Causal models allow for explicitly encoding independence relationships and information on which variables influence one another. Thus, they are well-suited for uncovering algorithmic bias in the data. This advantage is particularly important as a mismatch between the particular notion of fairness examined and the causal structures present in the data may produce undesirable outcomes (Loftus et al., 2018).

Crucially, by neglecting these intricacies, even ML systems that optimize common definitions of algorithmic bias risk diverting attention from the underlying social and legal factors contributing to fairness issues in the data or model. This happens because the initial focus becomes "fixing" the model and its predictions (Green and Hu, 2018). Moreover, when quantifying algorithmic bias, researchers and ML practitioners must often rely on unobservable theoretical constructs–such as educational attainment–that cannot be directly measured to group individuals for fairness assessments. In doing so, critical aspects like the reliability and validity of the instruments used to approximate these constructs are frequently overlooked (Jacobs and Wallach, 2021). Therefore, deploying measures of algorithmic bias alone does not guarantee the fairness of ML-based decisions. Instead, a holistic approach that considers the structure and limitations of the data and model, as well as the context in which they were created, is essential to minimize unfairness.

2.2 CBNs

Having established the advantages of deploying causal modeling to advance fairness in ML, we will briefly detail what CBNs are as we rely on them to model how variables influence each other as well as the underlying data-generating mechanisms (Chiappa and Isaac, 2019). As already mentioned, the graphical parts of CBNs consist of DAGs. A DAG is a graph with nodes (also called vertices) $\mathcal{X}$ that, in the case of a Bayesian Network, encode random variables and directed edges $\mathcal{E}$ connecting the vertices (Pearl, 2009). An edge from one node to another, i.e., $x_i \rightarrow x_j$, means that the first node causally influences the second node; we can call the first node the parent and the second the child node. A path in a DAG encompasses a sequence of directed edges, i.e., $x_i \rightarrow x_j \rightarrow \ldots \rightarrow x_t$. Furthermore, it holds for a CBN that a variable x_i is only dependent on its parents and independent of all other non-descendant variables given its parents, i.e.:

$$P(x_i) = \prod_{Pa(x_i)} P(x_i | Pa(x_i)) \tag{3}$$

where $Pa(x_i)$ is a parent of x_i. Therefore, CBNs encode independence information. In DAG 6 in Figure 1, A and Y are conditionally independent given X, which we write as $A \perp\!\!\!\perp Y | X$. This statement is equivalent to saying that all information relevant for Y is encoded in X, and we do not need to know A to learn something about Y (Pearl, 2009). Note, however, that A and Y are only conditionally independent, which means that the variables are correlated. An imperfect ML model may use this correlation, even though all information necessary is encoded in X.

Before we can work with DAGs, they first need to be constructed as the true underlying graph is usually unknown. There are several ways to do so:

(1) **Expert background knowledge.** While we can construct the DAG using background knowledge (Hicks et al., 2022), we usually do not know about causal relationships, or our ideas might not match the data. Still, expert knowledge is important: We often know about the temporal ordering of variables and, therefore, know that certain relationships cannot exist (e.g., grades cannot influence ethnicity). That is, we know a few things about the DAG structure.

(2) **Data-driven methods.** Research on causal structure learning has produced several methods to learn CBNs from data (Kitson et al., 2023a). If certain assumptions hold and data is sufficient, these methods work rather well (Scutari et al., 2019a). In more realistic cases, however, the methods cannot reliably produce accurate DAGs (Scanagatta et al., 2019).

(3) **Combining expert background knowledge and data-driven methods.** We may know that some relationships in the data are impossible or must exist, but we do not know about all relationships. We can feed this knowledge to the structure learning algorithms. Although combining both methods seems to lead to better results, doing so has been researched comparatively poorly, and some data-driven methods do not even allow the incorporation of background knowledge (Constantinou et al., 2023). In part, this lack of research is because there is no general procedure for its evaluation (i.e., which relationships are assumed to be known), and the background inclusion greatly depends on the data, knowledge, and general situation.

In this paper, we argue that combining data-driven and expert background knowledge is the best option. Before we later discuss how we can adapt the structure-learning methods to include background knowledge, we will now briefly explain which families of methods exist for this task. We will later evaluate one method from each of the three most popular families: constraint-based methods, score-based methods, and methods from functional causal modeling (Kitson et al., 2023a).[2]

2.2.1 Constraint-Based Structure Learning. Constraint-based structure learning consists of two stages. During the first stage, edges are removed iteratively from an initially complete undirected graph by performing independence tests (Kitson et al., 2023a). Edges can be removed when two variables are (conditionally) independent of each other. Whenever an edge is removed, the variables that make these variables conditionally independent are stored in a separating set S. For example, if A and B are independent given C, i.e., $A \perp\!\!\!\perp B|C$, then C is stored in $S_{A,B}$. During the second stage, as many edges as possible are oriented. To do this, we look at groups of three variables A, B, C, and their separating sets. If we have two variables A, B that are conditionally independent and both are dependent on the same third variable C and their separating set does not include C, i.e., $C \notin S_{A,B}$, then we have that $A \longrightarrow C$ and $B \longrightarrow C$. C is a so-called collider, and A, B, C form a v-structure. After all v-structures are identified and the corresponding edges are oriented, other edges are oriented to avoid new v-structures. This process concludes the second stage. It has to be noted that not all edges are usually oriented, as only those edges that are part of a v-structure or directly avoid a v-structure can be oriented. Therefore, constraint-based methods do not return a DAG but a Complete Partial DAG (CPDAG). Constraint-based methods are guaranteed to return the correct CPDAG if the independence tests return correct results (Kitson et al., 2023a; Pearl, 2009). Constraint-based algorithms are known to miss more edges than other methods, but also insert fewer incorrect edges (Scanagatta et al., 2019; Scutari et al., 2019a). We use the PC-Stable (abbreviated in this paper as PC) algorithm, which has been found to work well (Kitson et al., 2023a).

[2] Hybrid methods connecting constraint-based and score-based structure learning also exist (Kitson et al., 2023a). In practice, hybrid methods have been proven to work less well than the mentioned individual methods (Scanagatta et al., 2019; Scutari et al., 2019a). Hence, we will not consider them here.

2.2.2 Score-Based Structure Learning. In score-based structure learning, we aim to find a DAG that maximizes a score, which measures the compatibility between the DAG and the data (Kitson et al., 2023a). Hence, the search space of possible graphs must be searched, and the possible graphs must be compared with a score (e.g., an information-theoretic score). Searching the space of possible graphs is usually (though not always) done with a heuristic approach. Despite its simplicity, one algorithm frequently used directly or in some variants is the Hill-Climber (HC) (Scanagatta et al., 2019; Scutari et al., 2019a). HC starts with an empty graph and iteratively adds or deletes those edges that lead to the highest increase in the chosen score until the score no longer improves. A DAG that is at least a local maxima is returned, but reaching the global maxima is not guaranteed (Kitson et al., 2023a).

2.2.3 Functional Causal Models. Algorithms belonging to this family typically perform two steps (Shimizu, 2014). First, independent component analysis is used to determine which variables are not related to each other. Thus, a skeleton without oriented edges is constructed. For determining the exact causal ordering, functional causal modeling utilizes the property that variables can be determined by a function of their parent variables and a noise term that is independent of their parents. If the function is correctly identified, it is the case that the noise term is only independent of the parent variables for one direction and not for the other. Hence, algorithms belonging to this family search for such relationships between variables to determine the order. It should be noted that this method is usually used for continuous data, although it can also be used for discrete data. One of the most prominent algorithms belonging to this family is the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu, 2014).

2.2.4 Limitations of DAGs. While DAGs enable detailed modeling of causal relationships between variables, using them is not without limitations. First, as mentioned, we typically must learn DAGs from the existing data. Consequently, the learned DAG may not accurately reflect actual causal relationships; important predictive or sensitive variables that exert an effect might simply be absent from the data (Constantinou et al., 2021). Although this is a general challenge in measuring fairness, it is important to stress that our DAGs do not necessarily provide a complete picture of the causal mechanisms underlying the target variable. This limitation is particularly concerning because different methods for constructing DAGs are often evaluated using synthetic datasets that fail to reproduce the complexities of real-world contexts (Kitson et al., 2023b). More on the limitations will follow in Section 9.

2.3 CBNs and Fairness

Having defined CBNs, we can turn to how they support the use of causal fairness notions. As mentioned, the importance of causality for ensuring fairness in ML has gained significant attention (Makhlouf et al., 2024). Chiappa and Isaac (2019) demonstrate that by examining DAGs, we can assess whether sensitive variables have a causal influence on the target variable from a Bayesian Network perspective.

Several scholars have introduced causal fairness notions requiring well-defined DAGs (Makhlouf et al., 2024). Different causal fairness notions make different assumptions regarding the underlying causal graph (e.g., the relationship of unobserved confounders), the functional relationships in the data, and the existence of redlining variables, but all require DAGs to assess the influence of the sensitive variable on the target variable. Redlining variables are variables through which information from sensitive variables is not allowed to pass. In other words, all paths from a sensitive variable toward the target involving a redlining variable are considered problematic. Consequently, other paths are not seen as problematic. For instance, one might consider the influence of gender on salary via interview results as illegitimate while viewing the effect through job category as acceptable. Several fairness notions allow for the definition of such variables, the most prominent of which is the path-specific fairness notion (Madras et al., 2019; Kilbertus et al., 2017). Due to the flexibility of path-specific fairness notions, we believe that these are potentially the causal notions that can best realize the potential of causal fairness. However, all causal fairness notions provide a much more holistic picture of potential ML unfairness compared to relying on group-based measures only.

In cases where path-specific fairness notions are relevant, bias mitigation efforts should focus solely on the problematic pathway, thereby increasing the precision of interventions and potentially improving fairness and model accuracy. Several studies adopting this causal perspective have shown that path-specific bias mitigation is feasible and that not all causal pathways are inherently problematic (Madras et al., 2019; Kilbertus et al., 2017). More recently, some work has even focused on path-specific bias mitigation in settings where no explicit information about the sensitive feature is provided (Grari et al., 2021). However, a detailed discussion regarding the criteria used to classify a pathway as problematic is typically lacking. Moreover, prior studies do not thoroughly differentiate between various causal structures, nor do they examine how these differences may impact fairness and bias mitigation strategies. This absence of a systematic classification limits the practical application of path-specific bias mitigation approaches.

Also, building on Chiappa and Isaac's theoretical work, Binkytė-Sadauskienė et al. (2022) noted that we usually do not have the correct causal graph underlying the data. They employed several Bayesian Network structure learning algorithms on a real-life dataset and showed that the different methods not only learned different networks but that the different networks also had different implications for fairness. Logically, suppose a sensitive variable is not connected to the target at all in one learned graph but connected in some way in another graph. In that case, the resulting implications for fairness are very different. Again, this finding has far-reaching practical implications because it makes the use of path-specific bias mitigation and all causal fairness approaches unreliable or even impossible. Similarly, Maasch et al. (2024) remark that we need methods that are better suited to recover true connections between sensitive variables and the target. They develop an efficient algorithm for local discovery and can, therefore, identify which sensitive features impact the target directly. However, more complex and indirect relationships are not considered, although those are very relevant, as other work already highlighted (Chiappa and Isaac, 2019; Nilforoshan et al., 2022) and as we will highlight as well. In the next section, we will classify four different kinds of structures that we believe have different implications for fairness and how to handle potential bias mitigations. Subsequently, we will talk about the context that may lead to the decision of which paths are legitimate. Afterward, we will discuss how to learn more accurate DAGs using background knowledge.

3 CBN Structures and Fairness

Figure 1 shows different exemplary structures containing a sensitive variable and a target variable that may exist within a DAG.[3] In general, we differentiate between four different kinds of structures that have different implications for fairness or bias mitigation. As already stated, we will later show that different contexts can lead to further differences in the way we evaluate, measure, and potentially deal with unfairness.

In general, we speak of a potential problem for ML if it is likely that an ML model will use the sensitive variable or variables heavily dependent on (observed or unobserved) sensitive variables (proxies) for its prediction. As we will see, DAGs reveal different ways in which this can or cannot be the case. The four different kinds of structures we have identified are:

- Structures that include directed paths from a sensitive variable to a target variable: These are structures in which the sensitive variables directly or indirectly impact the target through an ancestor-descendant relationship (i.e., $A \rightarrow \ldots \rightarrow Y$). Examples of it are DAGs 1, 2, 3, and 4 in Figure 1. Here, it is clear from a network perspective that information about the sensitive variable is transported to the target (either through other variables or directly). Although, from a pure network perspective (e.g., consider DAG 2 in

[3] Appendix A additionally lists the structures with meaningful variable names along with exemplary scenarios.

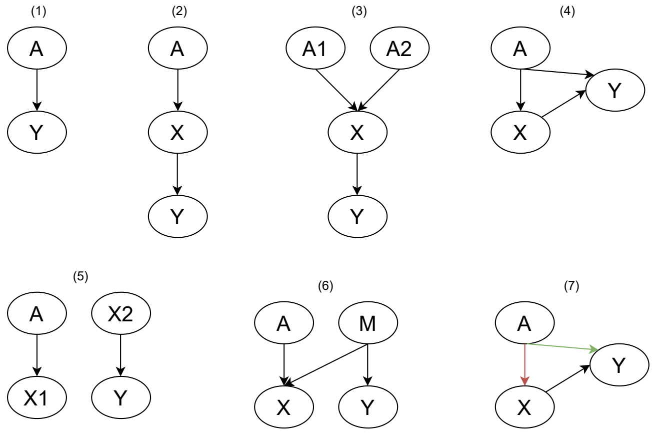


Fig. 1. Different causal structures.

Figure 1), conditioning on a variable that is a descendant of A and an ancestor of Y (here: X) makes the two variables independent, i.e., $A \perp\!\!\!\perp Y|X$, the variables are still correlated. Hence, from an ML perspective, this correlation makes it likely that the information from the sensitive variable is used (either by placing weight on the sensitive variable directly or by placing weight on any of the descendants of the sensitive variable). Moreover, even if we exclude A and try to mitigate any influence of A on other variables, we still have a potential fairness problem as the target itself is correlated with A. We call sensitive variables involved in such structures **problematic variables**. All paths that are involved in transporting information from the sensitive variable to the target are called **problematic structures**. All independence-based group fairness measures (such as DP) and causal-based fairness notions will indicate a fairness problem in such settings, whereas separation-based measures (e.g., EO) will not. As we will discuss later, depending on the setting, making a path-based fairness decision may be valuable. Generally, fairness could only be guaranteed for such a setting by excluding all sensitive features and their descendants. However, practically, and as already remarked by others, this will lead to the exclusion of most variables, rendering the model useless (Kusner et al., 2017). Hence, more advanced bias mitigation strategies need to be followed. However, proxy variables that are very highly correlated with the sensitive feature almost certainly need to be removed.

- The exact opposites of this first structure are structures that are unproblematic both from a network and an ML perspective. They are unproblematic when the sensitive and target variables are parts of different unconnected substructures. An example of such a case is DAG 5 in Figure 1. We call such sensitive variables **unproblematic variables**. It should be noted that some ML-models may still place a small weight on such features (e.g., Neural Networks are known to place a small weight on all variables regardless of their usefulness (Grinsztajn et al., 2022)), so to be absolutely certain that they are not used, we should probably still exclude them and their descendants. Neither group-based nor causal fairness notions are likely to indicate a fairness problem in this setting. Still, it could be a good idea to monitor different fairness measures to ascertain that the models do not place importance on the sensitive variable erroneously if the sensitive variables or their descendants are used for model training.
- Another potential case is that A and Y are, on their own, uncorrelated and independent, but information about A may be transported to Y nonetheless. An example of such a case is DAG 6 in Figure 1. The

transportation of information about the sensitive feature may happen when a child (here X) of A and another variable (here M) that is also a parent of Y are used in the model. In this case, X is a collider, and conditioning on or using a collider makes the parents dependent on each other (which they are otherwise not). In contrast, not conditioning for it ensures that any influence from A to Y is blocked. A prerequisite for this relationship is that there is no outgoing arrow from any variable influenced by A to any variable influencing Y. From a network perspective, this is as unproblematic as if there were no connection at all as long as we do not include (condition on) X, i.e., the collider. Looking at DAG 6 in Figure 1, all information relevant to the target is in M; both $X \perp\!\!\!\perp Y|M$ and $A \perp\!\!\!\perp Y|M$. It is even the case that A is fully independent of and uncorrelated with Y, which, at first glance, makes it look unproblematic. However, without conditioning on M, X is not independent of Y, and in either case, X and Y are correlated. For this reason, if we include X, an ML model may place weight on X, although this is not necessary once M is known. As X is correlated with A, information about A will be used in the prediction after all. However, if X (or generally all colliders and their children) is excluded as well, we no longer have a problem as the target itself is not correlated with the sensitive feature. We call sensitive variables part of such structures **blocked variables**. Interestingly, causal fairness notions will not necessarily indicate a problem in this setting if the DAG was learned using the true target (and not the prediction), as the network structure does not indicate a problem. Hence, this class highlights the importance of thinking about specific structures even when using causal fairness notions. Separation-based fairness notions (e.g., EO) are well-suited to pick up biases introduced through the machine learning model.

- Finally, we may have cases that look problematic from a network perspective but may not be from an ML perspective. These are cases in which the sensitive variable has a direct and an indirect influence on the target, and the two relationships are opposing each other (e.g., a positive relationship of A on X and X on Y, and a negative relationship of A on Y). An example of this case is DAG 7 in Figure 1. From a network perspective only, this is problematic as information about the sensitive variable is transported to the target. However, such a case may actually occur if a relevant variable is missing. If, e.g., we cannot observe M in DAG 6 in Figure 1, then we arrive at such a structure. In this case, A may, in fact, balance out its own effect that is transported through X, and we may arrive at a Y that is unbiased. In such a case, bias mitigation strategies that only remove A may lead to more biased models, as A cannot be used to minimize the impact. We call sensitive variables involved in such structures **opposing effects variables**. Whether the different fairness notions detect a problem depends on the precise fairness notion used; generally, however, separation-based fairness notions (e.g., EO) are well suited, and so are some causal fairness notions as long as they take the specific functional relationships into account (e.g., counterfactual notions).

The different classes of structures have different implications for how to deal with issues regarding fairness. For the *problematic variables*, depending on the context, the sensitive variable and its proxy variables may be removed, and some or all paths through which its information is transported may be subjected to a method of bias mitigation. In general, this class is the one to which all kinds of bias mitigation strategies can be applied. Depending on the context, measures of algorithmic bias need to be monitored. For these variables, knowing the causal structure will help in deciding which relationships are problematic and which are not, but this step still needs to be taken.

For the *unproblematic variables*, we can, to be absolutely certain that they will not be important for an ML model, exclude the sensitive variables and all variables in their subnetwork without having to worry about impacting accuracy (as these variables do not matter anyway), and then we are done with it.

For the *blocked variables*, we have to be aware that if we exclude A, we also have to exclude all variables that are descendants of A to ensure that the sensitive feature truly has no impact. Regarding the accuracy of resulting ML models, we also do not need these variables, so we can and probably should safely exclude them as they and their children are not causally related to the target. It should be noted that a simple analysis of the correlation between features may not reveal a fairness problem at all (as the sensitive feature and the target are not correlated), which makes it a particularly important case to consider.

For the *opposing effects variables*, we must also be aware that we need to exclude the variables that are descendants of A if we exclude A. Otherwise, we risk biased predictions. However, we potentially lose valuable information, which affects the overall model accuracy. If, however, we leave A in, then we should a) check that the target is not biased and b) monitor EO or other measures that are concerned with introducing bias through the prediction.

As we can see, the procedures to investigate and potentially mitigate biases for the last three classes are relatively straightforward, regardless of the context, in ideal circumstances. For the first class, however, the combination of the causal structure and the context determines what needs to be done regarding algorithmic bias. How context may matter will be discussed next.

4 Context and DAGs

DAGs, as already stated, can help us to see whether we need to think about algorithmic bias at all. However, they can also help us with thinking about whether the concrete influences are problematic or not, depending on the context, which then allows us to perform more targeted (potentially path-specific) bias mitigation and to select appropriate measures of algorithmic bias (Chiappa and Isaac, 2019; Cohausz et al., 2024).

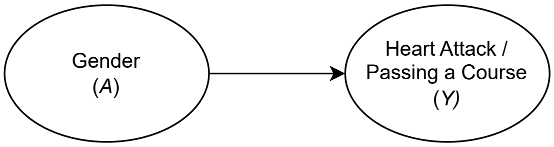


Fig. 2. Causal graph visualizing a direct connection between the sensitive variable A and the target Y.

Suppose we consider a direct connection between A and Y, such as in Figure 2. We may believe that it is okay to use the variable and to include its influence if, e.g., we aim to train a model that predicts heart attacks (for which it is known that symptoms are different for men and women). Then, we need to check using a separation-based measure like EO that the predictive accuracy is equal for all demographic groups. If, however, there is a direct influence of gender on passing a course, we probably do not think that this is okay. In such a case, because the target is biased in a problematic way, too, we need to remove the variable and still check for biases using either a group fairness notion like DP or a similar causal fairness notion. The situation would again be different if we only aim to understand why a person may fail a course. Then, the observation that gender matters should be investigated further. In these cases, the application scenario matters a lot.

Similarly, one has to investigate the roots of potential historical disparities that can cause algorithmic bias. Returning to the example of predicting heart attacks, a decision system trained on historical data will most likely predict more heart attacks for men. This result would cause a measure like DP to indicate biased decision-making, prompting us to take action since this disparity will likely, at least in part, be caused by the well-established underdiagnosis and undertreatment of heart attacks in women in the past (Kuehnemund et al., 2021). Had there been more awareness of this problem historically, the data likely would not exhibit a bias to the same degree, rendering it more readily usable without accounting for its inherent fairness problems. In contrast, when designing a system to assign social aid based on past redistributive social programs, one might view a potential disparity between lower and higher income groups not only as unproblematic but desired, even though DP might still indicate a bias.

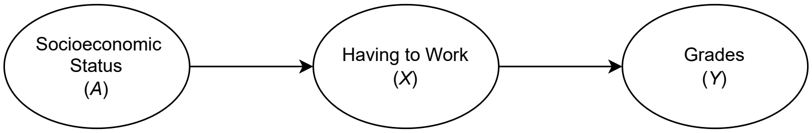


Fig. 3. Causal graph visualizing an indirect connection between the sensitive variable A and the target Y via a third variable X.

Suppose there is an indirect connection between A and Y (as shown in Figure 3) or both a direct and an indirect one; now, the context and our personal beliefs may play an even more fine-grained role. For example, socioeconomic status influences how much people have to work to finance their studies, and this impacts their study time, potentially reducing their study success. Then, there are three possible points of view: One is that this relationship is legitimate, and we should only exclude A so that no direct relationship is possible. An alternative one is that this relationship is illegitimate. And the third one is that it is legitimate to have this path as long as we predict only to help people, whereas it is not okay if we use it to determine who is admitted to the course. If we believe it is problematic, the influence of A on all intermittent variables would need to be mitigated using established bias mitigation strategies, and DP or causal fairness notions would need to be monitored. It is also important to exclude potential proxy variables in such a case. In general, whenever we decide that relationships are not acceptable and whenever we exclude A and potentially perform bias mitigation, we should use fairness notions that do not rely on an unbiased target variable, and we have to accept that accuracy decreases as we acknowledge the target to be biased. Looking at DAGs allows us to make decisions based on the context and the exact paths involved.[4] If we decide that only some paths are problematic, we should turn to path-specific fairness notions and apply corresponding path-specific bias mitigation (Chiappa, 2019). Group-based notions or other causal fairness notions would indicate a stronger bias than is actually justified.

5 Learning CBNs with Background Information

Detecting the above-described structures relies on accurate DAGs. We argue that when constructing a graph to assess fairness, we can use a standard procedure to combine background knowledge and data-driven methods. The reason for this is the background information we automatically have when considering fairness.

5.1 Background Information

As mentioned in Section 2.1, it follows from a definition of sensitive features that non-sensitive variables cannot influence them (Baker et al., 2023). Additionally, the target variable usually follows all other variables temporally. For example, if we try to predict admission to a university, all information that can be used has existed longer than the admission decision. Therefore, we can separate the variables into three groups: target variables (which cannot influence any other variables), sensitive variables (which cannot be influenced by any other variables), and regular predictive variables, for which it logically follows that they cannot influence sensitive variables or be influenced by the target. There may also be situations where sensitive variables can be influenced by other sensitive variables or where we know there is an order within the other predictive variables. However, we generally have at least three tiers: the target, other predictive features, and sensitive variables.[5] With the specification of these tiers, we already have a lot of background knowledge: we can require that the data-driven structure learning methods do not include any edges that are impossible according to this specification. Using this background knowledge is

[4] For a more in-depth discussion of different contexts and the relation to fairness measures and bias mitigation, see Cohausz et al. (2024).

[5] If there is a case for which it does not hold that other variables do not influence one or some sensitive variables, we can simply loosen the restriction again or define the specific relationship we may allow; in the former case, the structure learning then relies more on the data.

particularly helpful, as it is also the knowledge we need to evaluate algorithmic bias, anyhow: knowing which variables are sensitive and what the target is. Additional knowledge we have about the structures can also be specified.[6] It should be noted that during structure learning, the information from background knowledge can "propagate". As already learned edges (or learned absences of edges) influence the subsequent learning of other edges, prohibiting certain edges may lead to more correctly learned edge existence/absence than just the concrete edge allowed or forbidden according to the background knowledge.

5.2 Adapting Structure-Learning Methods to Include Background Knowledge

We will now explain how we can adapt each of the methods introduced in Section 2 to incorporate background knowledge.

5.2.1 Adapting Constraint-Based Structure Learning. Including background information in constraint-based methods is not straightforward, as the first stage (identifying which variables are conditionally independent) cannot really be modified, and no implementation so far allows a user to specify background information (Constantinou et al., 2023). Our approach is to use the background information at the end of the second stage during edge orientation: If we have an undirected edge and our background knowledge does not allow one direction, then the edge is oriented accordingly. Afterward, further edges are oriented to avoid v-structures again. Compared to the adaptations of the other methods, this method makes comparatively little use of the background information. It is also not guaranteed that relationships that go against our background knowledge do not exist because the edge may already have been oriented during the v-structure orientation. However, if the CPDAG is correct until we inject the background knowledge, the resulting graph will also be correct.

5.2.2 Adapting Score-Based Structure Learning. Adapting score-based methods, here HC, to handle the background information is easier, as we can restrict the search space, i.e., edges that are impossible according to our classification will never be added (Constantinou et al., 2023).[7] Constantinou et al. (2023) recently experimented with different kinds of background knowledge and their effect on the accuracy of DAGs but found that restricting edges has only a small effect on accuracy. However, we limit the search space more fundamentally, which may also impact the efficiency with which DAGs are learned.

5.2.3 Adapting Functional Causal Models. Similar to HC, we can include background knowledge by preventing LiNGAM from considering certain directed relationships. These edges are then never allowed to be added, and, therefore, the impact of background knowledge is relatively direct.

6 Evaluation on Artificial Data

To evaluate the research questions formulated in Section 1, we need to compare the learned DAGs to the ground truth DAGs. As such, we need to know the real ground truth, which is only possible when using synthetic data that we sample from networks. In total, we chose 10 networks to evaluate against; five are from the *bnlearn* library Scutari et al. (2019b) and are frequently used to evaluate structure learning methods; we constructed the other five networks ourselves. We did the latter for three reasons: In contrast to the data sampled from the *bnlearn* library (which is discrete), we can also sample continuous data. Second, we can additionally specify non-linear relationships, making it more difficult to learn the networks. Third, we can ensure that all of the different structures in which sensitive variables can be involved, according to Section 3, are included. A summary of the different DAGs can be found in Table 1. Our networks are called Synthetic I-V and are abbreviated as SD I, SD II, etc.

[6] That is, whether certain variables cannot have ingoing edges or cannot be influenced by certain other variables, or whether edges are definitely in place between certain variables.

[7] Similarly, we could also add information that a certain edge needs to exist–then the edge is directly added and can never be removed.

Table 1. This table provides a summary of the DAGs used in the evaluation. It states the number of nodes and edges, problematic variables, blocked variables, opposite effects variables, and unproblematic variables, as well as the number of distinct paths through which problematic information travels. It also states whether the sampled data is discrete or continuous and whether non-linear data was sampled.

Name	Nodes	Edges	Problematic	Blocked	Opposing Effects	Unproblematic	Problematic Str.	Data Type	Non-Linearity
alarm	37	46	9	2	0	0	17	Discrete	-
asia	8	8	2	0	0	0	3	Discrete	-
earthquake	5	4	2	0	0	0	2	Discrete	-
insurance	27	52	1	0	0	0	118	Discrete	-
sachs	11	17	1	0	0	1	5	Discrete	-
SD I	9	7	2	1	0	1	8	Continuous	No
SD II	10	13	4	0	2	0	7	Continuous	Yes
SD III	20	29	3	2	0	1	16	Continuous	Yes
SD IV	9	8	1	1	1	1	1	Continuous	Yes
SD V	13	13	1	1	1	1	1	Continuous	No

For each network from *bnlearn*, we selected some of the root nodes or children of root nodes that we had already declared as sensitive to be sensitive. For all networks, we extracted the class of each sensitive variable (i.e., problematic, unproblematic, etc.) and all ground truth paths through which information from problematic variables is transported.

6.1 Evaluation of the Accuracy of the Learned DAGs and Structures

We use the procedure described in Algorithm 1 to conduct the experiments. We evaluate the structure learning methods for each network using data sizes of 500, 1,000, and 10,000. We sample 30 times for each of the data sizes to achieve reliable results and to be able to compute test statistics. For each sample, we try to learn the network with each method, once without and once with background information. The resulting graph is compared to the ground truth, and the resulting values are stored.

Algorithm 1 The algorithm shows the setup of the experiments for each DAG.

```
for sample size in {500, 1000, 10000} do
    for experiment in range(30) do
        sample data
        for method in {PC, HC, LiNGAM} do
            for information in {No Info, Info} do
                learn DAG
                compare to ground truth
```

When comparing the learned edges to the correct edges, we focus on two measures, True Positives (TP) and False Positives (FP), because they allow for a detailed but easily comprehensible interpretation of the results. We normalize both resulting values by dividing by the number of edges in the ground truth graph in order to compare values across different networks, hence having True Positive (TPR) and False Positive Rates (FPR). Note that, for FPR, this can lead to values larger than 1 as more wrong edges can be inserted than correct edges exist. Note that for PC, we compare against the ground truth CPDAG, which also changes when more information is available. Moreover, note that edges that are undirected in the learned graph but could be directed according to the ground truth CPDAG are considered incorrect. As we sample the data 30 times, we report the mean and the 95%-confidence intervals in our results.

To evaluate the discovery of the different classes of sensitive variables and the problematic structures, we consider accuracy.

6.2 Evaluation of the Impact on Causal Fairness Notions

To capture whether the differences in the constructed DAGs also impact the estimation of fairness notions (**RQ6**), we employ the following causal fairness notions:

- The **Natural Direct Effect (NDE)** quantifies the direct effect of a binary sensitive attribute A on the target variable Y (i.e., the pathway $A \longrightarrow Y$) independent of any intermediate or mediating variables Z. It captures the change in the probability of $Y = 1$ when switching the sensitive attribute from a baseline level $A = 0$ (written as a_0) to an active level $A = 1$ (written as a_1) while keeping any mediating variables Z fixed at the levels they would naturally attain under $A = 0$ (denoted as Z_{a_0}) (Pearl, 2009). Formally, we write:

$$\text{NDE}_{a_1,a_0}(Y) = P(Y_{a_1,Z_{a_0}}) - P(Y_{a_0}). \tag{4}$$

- Conversely, the **Natural Indirect Effect (NIE)** estimates the indirect effect of A on Y via the mediating variables Z (i.e., the pathway(s) $A \longrightarrow Z \longrightarrow Y$). Thus, the change in the probability of $Y = 1$ when setting all mediating variables Z to the values they would have if $A = 1$ (Z_{a_1}) while keeping the actual sensitive attribute at $A = 0$ (written as a_0) is computed (Pearl, 2009). It is defined thusly:

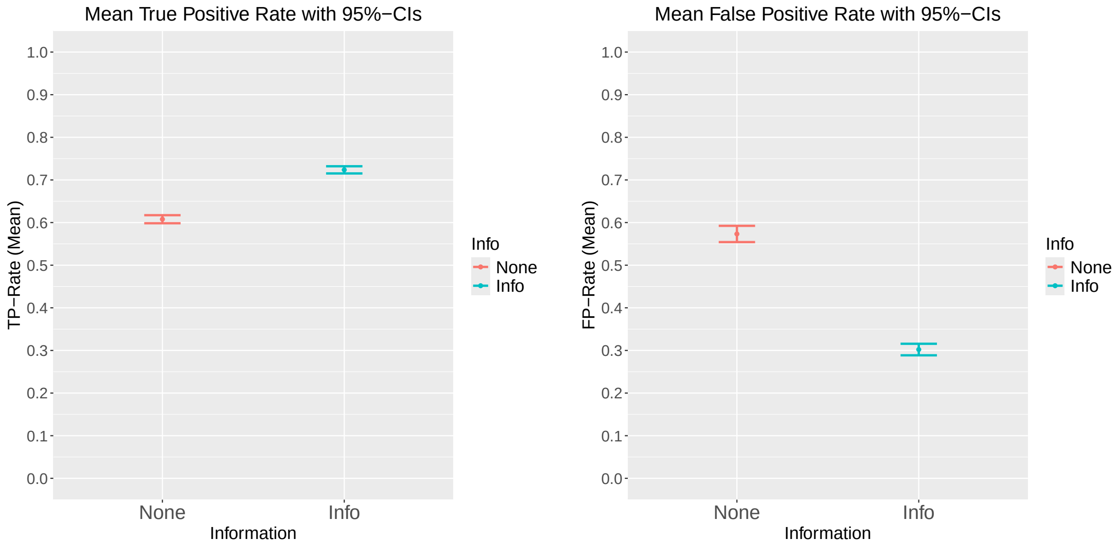
$$\text{NIE}_{a_1,a_0}(Y) = P(Y_{a_0,Z_{a_1}}) - P(Y_{a_0}). \tag{5}$$

Both NDE and NIE rely on CBNs to derive how A affects Y (directly and indirectly). Compared to other notions of causal fairness (e.g., counterfactual fairness, path-specific counterfactual effects), apart from their straightforward interpretation, these two measures have the advantage of requiring fewer assumptions, for instance regarding the existence of proxy or redlining variables (i.e., NIE does not discriminate between potentially "fair" indirect effects and discriminative ones) or the relationships of unobserved variables (Makhlouf et al., 2024). While we generally argue that path-specific bias mitigation notions are often a better choice, they require us to define redlining variables. As we want to keep our evaluation of causal fairness as robust as possible and do not want to make assumptions, we decided to go with the above-mentioned measures. It is important to highlight, however, that other causal fairness measures exist and may be better suited in specific contexts (Makhlouf et al., 2024).

In order to gauge how background information influences the assessment of causal fairness, we estimate the NDE and NIE five times for each data sample: using the ground-truth DAG, using the DAGs learned by applying LiNGAM and HC, and using the DAGs learned by applying LiNGAM and HC while having background information. We do not estimate the NDE and NIE for PC because we require directed edges. Then, for each data sample and each of NDE and NIE, we compute:

$$\begin{aligned} D_1 &= \left|\hat{\theta}_{\text{gt}} - \hat{\theta}_{\text{no\_info}}\right|, \\ D_2 &= \left|\hat{\theta}_{\text{gt}} - \hat{\theta}_{\text{info}}\right|, \\ \text{change\_ratio} &= \frac{D_2}{D_1}, \end{aligned} \tag{6}$$

where θ denotes the causal-effect quantity of interest—either the NDE or the NIE, as defined in Equations (4) and (5). We write $\hat{\theta}_{\text{gt}}$, $\hat{\theta}_{\text{no\_info}}$, and, $\hat{\theta}_{\text{info}}$ for the empirical estimates of θ obtained under, respectively, the ground-truth DAG, the learned DAG without background information, and the learned DAG with background information. Hence, the resulting ratio can range from 0 (i.e., no change) to infinity. A value larger than 1 indicates that the



(True Positives) **(False Positives)**

Fig. 4. Results summarized across all DAGs.

fairness estimates from a DAG learned without information were closer to the estimate using the ground-truth DAG than the estimate using the DAG learned with information was to the estimate using the ground-truth DAG. Values smaller than 1 mean that the estimate using the DAG learned with information is closer to the estimate using the ground-truth DAG. After computing the values for each sample, we average the ratios across the 30 data samples. The reason for computing the ratio is that we are interested in the changes introduced by the background knowledge and not so much in the absolute values of NDE and NIE. These absolute values are naturally of very different magnitudes, considering the different example networks.

7 Results on Artificial Data

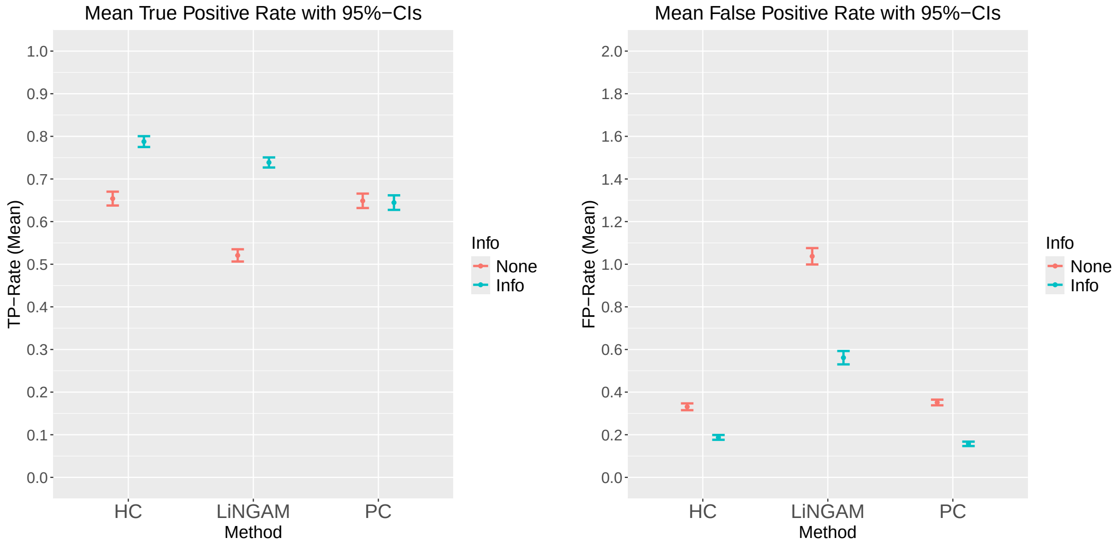
Figure 4 and Table 2 show the results for **RQ1** for each ground truth DAG and summed over all DAGs. Generally, we can see that the TPR increases across all settings (Figure 4) when information is provided. The increase is significant, as indicated by the confidence intervals, but not extremely large. A larger difference can be seen for the FPR. Here, providing information leads to a large decrease in incorrect edges being learned, making the overall learned graph much more correct. This conclusion generally holds for all individual ground truth DAGs (table 2) with a few exceptions. For the TPR in SD I, we can see that providing information does not lead to a significant improvement. However, this may be the case because even before providing any information, the percentage of correct edges was quite large. For insurance and alarm, the impact on the FPR is not significant. In general, these appear to be difficult networks to learn from data. All in all, **RQ1** is answered positively, though, and as expected[8].

Figure 5 and Table 3 show the results for **RQ2a**. Across all DAGs, we can see that providing information increases the TPR for HC and LiNGAM but not for PC. For PC, there is no effect on TPR. For the FPR, we can observe that all methods benefit from having information available, though LiNGAM seems to benefit the most.

[8] Plots showing the relative change ratio between using and not using background information as an alternative way to present the results are shown in Appendix C.

Table 2. The detailed results for **RQ1**. The numbers in brackets indicate the lower and upper bounds of the surrounding confidence interval. Significant differences are bold.

Network	% True Pos. without Info	% True Pos. with Info	% False Positives without Info	% False Positives with Info
alarm	0.58 (0.56,0.60)	**0.65 (0.63,0.66)**	0.74 (0.67,0.81)	0.63 (0.56, 0.71)
asia	0.49 (0.46, 0.51)	**0.67 (0.65,0.70)**	0.67 (0.63,0.71)	**0.31 (0.27,0.34)**
earthquake	0.68 (0.65,0.71)	**0.86 (0.83,0.89)**	0.45 (0.41, 0.50)	**0.1 (0.08,0.12)**
insurance	0.35 (0.34,0.36)	**0.41 (0.39,0.42)**	0.67 (0.59,0.76)	0.56 (0.49,0.62)
sachs	0.38 (0.36, 0.41)	**0.65 (0.63,0.67)**	0.81 (0.75,0.86)	**0.34 (0.32,0.37)**
SD I	0.86 (0.85,0.87)	0.87 (0.86,0.88)	0.15 (0.13,0.18)	**0.07 (0.06, 0.09)**
SD II	0.50 (0.49,0.52)	**0.58 (0.56,0.59)**	0.35 (0.33,0.38)	**0.28 (0.25,0.30)**
SD III	0.64 (0.63,0.65)	**0.67 (0.66,0.69)**	0.70 (0.65,0.75)	**0.44 (0.42,0.46)**
SD IV	0.82 (0.79,0.85)	**0.97 (0.97,0.98)**	0.53 (0.45,0.60)	**0.07 (0.06,0.09)**
SD V	0.77 (0.75,0.80)	**0.90 (0.89,0.91)**	0.65 (0.59,0.71)	**0.22 (0.20,0.24)**



(True Positives) **(false Positives)**

Fig. 5. Results for RQ2a summarized across all DAGs.

LiNGAM, however, is also the method that generally inserts the most FP edges. Hence, it appears consequential that it benefits the most from a reduction of the search space. It is more interesting that PC also benefits regarding the reduction of FPs, as we do not even restrict the search space there. As undirected edges count as being false positives, this does make sense, though. These findings are also reflected when looking at the results for the individual DAGs in Table 3 by far and large. Another conclusion that we may draw is that HC has the highest TPR (although LiNGAM is comparable with information and PC without information) while having the smallest FPR (together with PC). As such, when information is available, HC seems to be a very good choice. To answer **RQ2a**, HC and LiNGAM benefit more from providing information than PC, but PC also benefits from it.

Table 3. The detailed results for **RQ2a**. The numbers in brackets indicate the lower and upper bounds of the surrounding confidence interval. Significant differences are bold.

Network	Method	% True Pos. without Info	% True Pos. with Info	% False Positives without Info	% False Positives without Info
alarm	HC	0.47 (0.45,0.49)	**0.62 (0.60,0.65)**	0.54 (0.52,0.57)	**0.37 (0.35,0.39)**
	LiNGAM	0.60 (0.58,0.62)	**0.65 (0.63,0.66)**	1.49 (1.40,1.57)	1.45 (1.38,1.52)
	PC	0.8 (0.64,0.71)	0.67 (0.64,0.71)	0.19 (0.17,0.21)	**0.08 (0.06,0.09)**
asia	HC	0.48 (0.43,0.53)	**0.79 (0.74,0.83)**	0.53 (0.47,0.60)	**0.17 (0.11,0.24)**
	LiNGAM	0.46 (0.44,0.47)	**0.71 (0.69,0.73)**	1.01 (0.97,1.04)	**0.50 (0.45,0.54)**
	PC	0.52 (0.48,0.56)	0.52 (0.48,0.56)	0.47 (0.43,0.51)	**0.25 (0.20,0.30)**
earthquake	HC	0.65 (0.57,0.72)	**0.92 (0.88,0.96)**	0.27 (0.29,0.45)	**0.07 (0.03,0.11)**
	LiNGAM	0.72 (0.70,0.74)	**0.98 (0.96, 1.00)**	0.66 (0.62,0.69)	**0.16 (0.12,0.19)**
	PC	0.67 (0.61,0.73)	0.67 (0.61,0.73)	0.34 (0.37,0.41)	**0.08 (0.05,0.10)**
insurance	HC	0.40 (0.38,0.43)	0.44 (0.41,0.46)	0.30 (0.28,0.31)	0.27 (0.26,0.29)
	LiNGAM	0.30 (0.29,0.30)	**0.43 (0.41,0.45)**	1.56 (1.45,1.68)	**1.24 (1.83,1.31)**
	PC	0.35 (0.32,0.37)	0.35 (0.32,0.37)	0.16 (0.15,0.16)	0.15 (0.15,0.16)
sachs	HC	0.44 (0.41,0.47)	**0.77 (0.75,0.79)**	0.42 (0.39,0.46)	**0.12 (0.11,0.12)**
	LiNGAM	0.17 (0.16,0.17)	**0.68 (0.67,0.70)**	1.48 (1.44,1.50)	**0.55 (0.52,0.59)**
	PC	0.54 (0.49,0.58)	0.50 (0.45,0.55)	0.52 (0.49,0.54)	**0.36 (0.34,0.38)**
SD I	HC	0.90 (0.89,0.91)	**0.92 (0.92,0.92)**	0.04 (0.03,0.05)	0.02 (0.02,0.03)
	LiNGAM	0.90 (0.89,0.92)	**0.92 (0.92,0.92)**	0.05 (0.03,0.07)	0.02 (0.01,0.03)
	PC	0.78 (0.76,0.80)	0.78 (0.76,0.80)	0.37 (0.35,0.39)	**0.18 (0.15,0.20)**
SD II	HC	0.61 (0.60,0.63)	0.71 (0.70,0.72)	0.35 (0.33,0.38)	**0.26 (0.24,0.27)**
	LiNGAM	0.47 (0.45,0.48)	**0.59 (0.57,0.61)**	0.52 (0.49,0.55)	0.46 (0.44,0.49)
	PC	0.43 (0.40,0.46)	0.43 (0.40,0.46)	0.19 (0.17,0.21)	**0.11 (0.09,0.13)**
SD III	HC	0.77 (0.76,0.77)	**0.78 (0.78,0.79)**	0.33 (0.32,0.34)	0.33 (0.32,0.34)
	LiNGAM	0.59 (0.58,0.60)	**0.67 (0.66,0.68)**	1.19 (1.12,1.26)	**0.67 (0.65,0.69)**
	PC	0.57 (0.56,0.58)	0.57 (0.56,0.58)	0.59 (0.57,0.60)	**0.33 (0.31,0.34)**
SD IV	HC	0.97 (0.95,0.99)	**1.00 (1.00,1.00)**	0.09 (0.06,0.13)	**0.03 (0.01,0.4)**
	LiNGAM	0.50 (0.45,0.54)	**0.93 (0.91,0.94)**	1.21 (1.08,1.33)	**0.17 (0.14,0.19)**
	PC	1.00 (0.99,1.00)	0.96 (0.95,0.98)	0.28 (0.26,0.31)	**0.03 (0.01,0.04)**
SD V	HC	0.84 (0.84,0.85)	**0.93 (0.92,0.93)**	0.32 (0.31,0.34)	**0.24 (0.22,0.26)**
	LiNGAM	0.51 (0.48,0.54)	**0.82 (0.81,0.84)**	1.22 (1.12,1.32)	**0.39 (0.35,0.42)**
	PC	0.96 (0.95,0.98)	0.96 (0.95,0.98)	0.41 (0.39,0.44)	**0.02 (0.01,0.04)**

For **RQ2b**, we can clearly see that providing information is helpful regardless of data size, as evidenced by Figure 6 and Table 4. All benefit from information regarding both TPR and FPR. Furthermore, we can observe that more data generally leads to more TPs being found. For the FPs, it is not the case that there are fewer of them when more data is available. The trend rather shows that there are more FPs in this case (although the difference is not significant). This finding makes sense, as more data is known to lead to more edges being inserted overall.

For **RQ3**, we can see that providing information only rarely leads to introducing errors by looking at Table 5. On average, only 0.006 edges that were correct before are no longer there when providing information for HC. The numbers for PC (0.005) and LiNGAM (0.059) are also very small. Note that these are absolute numbers. Furthermore, providing information only rarely leads to introducing new edges that are wrong. This result is particularly reassuring for PC, where the late inclusion of background information could, theoretically, lead to many wrong edges being inserted if the graph is not quite correct at this point. Because we only orient edges after all other parts of the algorithm have been concluded, the following v-structure orientation phase can propagate

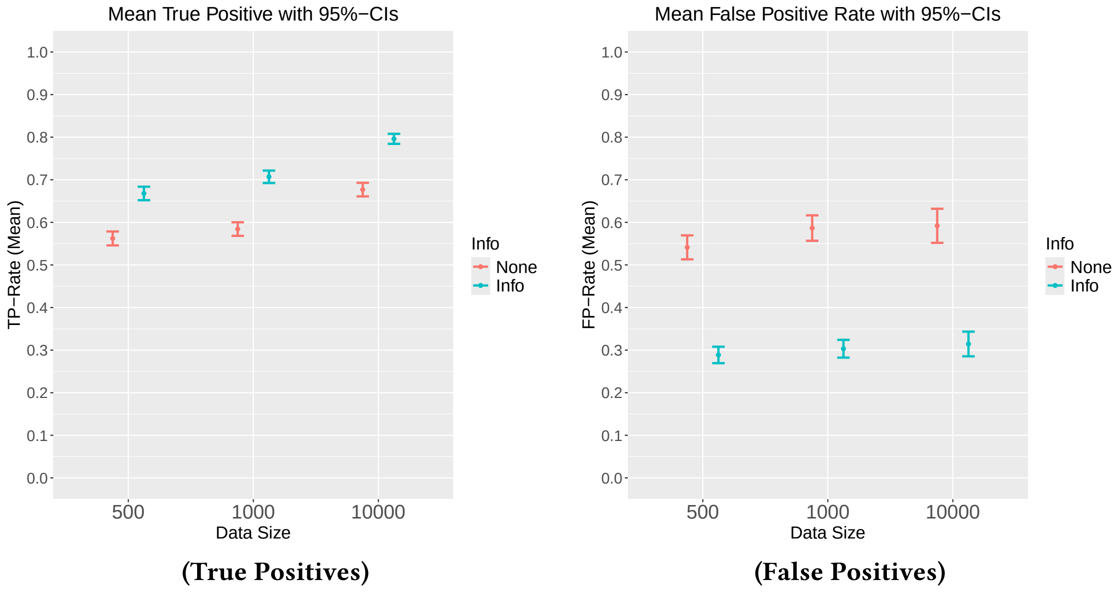


Fig. 6. Results for **RQ2b** summarized across all DAGs.

errors by wrong orientation if some edges are simply incorrectly added. As PC tends not to add wrong edges, the result is not entirely surprising, but it is still reassuring.

For **RQ4**, we can see in Figure 7a that problematic variables are detected better when information is available for HC and LiNGAM. For PC, the background information does not increase the percentage of problematic variables found. Moreover, we can see that HC and, to a slightly smaller degree, LiNGAM, are able to find most of the problematic variables, with HC being able to find, on average, roughly 87% when using information.

Somewhat surprisingly, Figure 7b reveals that blocked variables are also detected reliably, at least by HC. This finding is surprising because several edges need to be found correctly to classify a variable as blocked. Moreover, alarm and SDIII proved to be difficult to learn when looking at Table 2 again, which makes these results all the more surprising. Having information seems to lead to a better detection of blocked variables, but the difference is not significant.

Figure 8a shows how many variables that truly have no connection were classified as such. Only four datasets (sachs and SDIII, SDIV, SDV) contained such variables. PC found all of these variables even without information available. This result is expected because PC tends to place too few edges in general and hardly any wrong edges. LiNGAM and HC were also fairly successful at finding the variables; using information seemed to have a positive impact on LiNGAM but not HC.

Figure 8b shows how many variables with opposing effects are correctly detected. Correctly classifying these variables is challenging because it requires the correct insertion of several edges. Interestingly, we can observe that all methods perform similarly well when information is available and reach very good levels in this case. In part, this result is explainable through the fact that only a few datasets (three) contained opposing effects variables. When no information is available, HC still performs about as well, but LiNGAM and PC perform much worse. This finding is a little surprising for PC, for which the inclusion of information usually does not add much.

Figure 9 shows the results for **RQ5**. As a brief reminder, for **RQ5**, we check whether all paths through which a potentially problematic sensitive variable impacts the target are correctly recovered (not just that a variable is

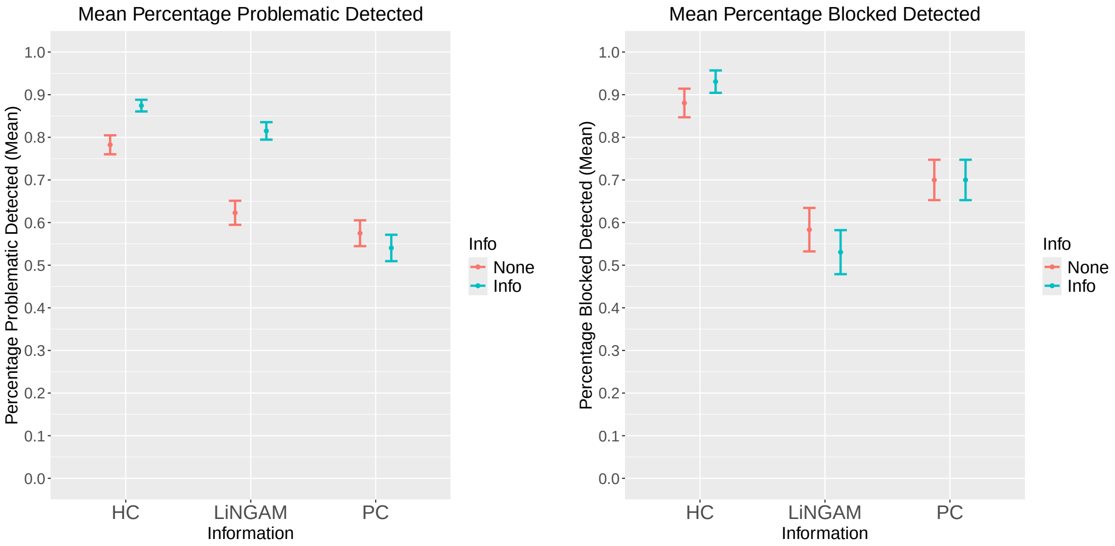
Table 4. The detailed results for **RQ2b**. The numbers in brackets indicate the lower and upper bounds of the surrounding confidence interval. Significant differences are bold.

Network	Data Size	% True Pos. without Info	% True Pos. with Info	% False Positives without Info	% False Positives without Info
alarm	500	0.49 (0.47,0.51)	**0.54 (0.52,0.55)**	0.66 (0.57, 0.75)	0.55 (0.45,0.65)
	1000	0.56 (0.54,0.58)	**0.62 (0.61,0.63)**	0.67 (0.58,0.76)	0.57 (0.47, 0.68)
	10000	0.70 (0.66,0.73)	**0.79 (0.77,0.81)**	0.89 (0.73, 1.06)	0.77 (0.61,0.94)
asia	500	0.48 (0.45,0.52)	**0.66 (0.61,0.70)**	0.64 (0.56,0.71)	**0.30 (0.25,0.36)**
	1000	0.43 (0.39,0.46)	**0.58 (0.54,0.63)**	0.72 (0.65,0.79)	**0.42 (0.36,0.48)**
	10000	0.54 (0.50,0.58)	**0.77 (0.74,0.81)**	0.32 (0.59,0.72)	**0.20 (0.15,0.25)**
earthquake	500	0.56 (0.50,0.61)	**0.72 (0.66,0.79)**	0.47 (0.41,0.53)	**0.13 (0.09,0.16)**
	1000	0.60 (0.56,0.65)	**0.85 (0.81,0.89)**	0.63 (0.57,0.69)	**0.13 (0.09,0.18)**
	10000	0.88 (0.84,0.92)	**1 (1,1)**	0.27 (0.19,0.34)	**0.04 (0.02,0.06)**
insurance	500	0.28 (0.27,0.29)	**0.31 (0.30,0.32)**	0.51 (0.42,0.59)	0.46 (0.38,0.64)
	1000	0.31 (0.30,0.32)	**0.35 (0.34,0.36)**	0.57 (0.47,0.68)	0.50 (0.41,0.59)
	10000	0.46 (0.43,0.48)	**0.56 (0.55,0.57)**	0.94 (0.73,1.15)	0.71 (0.57,0.85)
sachs	500	0.30 (0.27,0.33)	**0.53 (0.49,0.57)**	0.71 (0.61 0.81)	**0.31 (0.27,0.35)**
	1000	0.34 (0.31,0.37)	**0.63 (0.59,0.66)**	0.84 (0.74,0.93)	**0.35 (0.32,0.39)**
	10000	0.50 (0.44,0.56)	**0.80 (0.78,0.81)**	0.87 (0.76,0.98)	**0.36 (0.31,0.42)**
SD I	500	0.83 (0.80,0.85)	0.85 (0.82,0.87)	0.19 (0.15,0.23)	**0.09 (0.07,0.11)**
	1000	0.84 (0.82, 0.86)	0.86 (0.84,0.88)	0.17 (0.13,0.21)	**0.08 (0.05,0.10)**
	10000	0.91 (0.90,0.92)	0.91 (0.90,0.92)	0.11 (0.08,0.14)	0.06 (0.04,0.08)
SD II	500	0.48 (0.46,0.51)	**0.57 (0.54,0.60)**	0.39 (0.35,0.44)	**0.28 (0.25,0.32)**
	1000	0.53 (0.51,0.55)	**0.62 (0.59,0.64)**	0.36 (0.32,0.39)	**0.26 (0.23,0.30)**
	10000	0.50 (0.46,,0.53)	0.54 (0.50,0.58)	0.31 (0.28,0.34)	0.28 (0.24,0.32)
SD III	500	0.63 (0.60,0.65)	0.65 (0.63,0.68)	0.71 (0.62,0.80)	**0.45 (0.41,0.49)**
	1000	0.64 (0.62,0.66)	0.67 (0.65,0.69)	0.71 (0.62,0.80)	**0.45 (0.41,0.49)**
	10000	0.66 (0.64,0.68)	**0.70 (0.69,0.72)**	0.68 (0.61,0.76)	**0.43 (0.39,0.46)**
SD IV	500	0.82 (0.77,0.88)	**0.98 (0.97,0.99)**	0.51 (0.39,0.63)	**0.07 (0.05,0.10)**
	1000	0.81 (0.75,0.87)	**0.98 (0.97,0.99)**	0.55 (0.42,0.68)	**0.07 (0.05,0.09)**
	10000	0.83 (0.77,0.88)	**0.97 (0.96,0.98)**	0.52 (0.39,0.66)	**0.08 (0.06,0.11)**
SD V	500	0.75 (0.71,0.79)	**0.87 (0.85,0.89)**	0.63 (0.53,0.73)	**0.24 (0.20,0.28)**
	1000	0.78 (0.73,0.83)	**0.92 (0.91,0.94)**	0.65 (0.55,0.75)	**0.20 (0.17,0.24)**
	10000	0.79 (0.74,0.83)	**0.92 (0.90,0.93)**	0.67 (0.57,0.78)	**0.21 (0.57,0.78)**

Table 5. The results for **RQ3**. The column *Avg. Correct to Missing* shows the average number of edges that were correctly found without information but not longer once information was provided. The column *Avg. New False Positive* shows the average number of edges that were incorrectly found only when providing information but not when not providing information.

Method	Avg. Correct to Missing	Avg. New False Positive
HC	0.006	0.049
LiNGAM	0.059	0.310
PC	0.005	0.000

correctly classified as problematic). Detecting all paths through which potentially problematic sensitive variables impact the target is a very challenging task, as large parts of the network must be recovered correctly. In the figure,



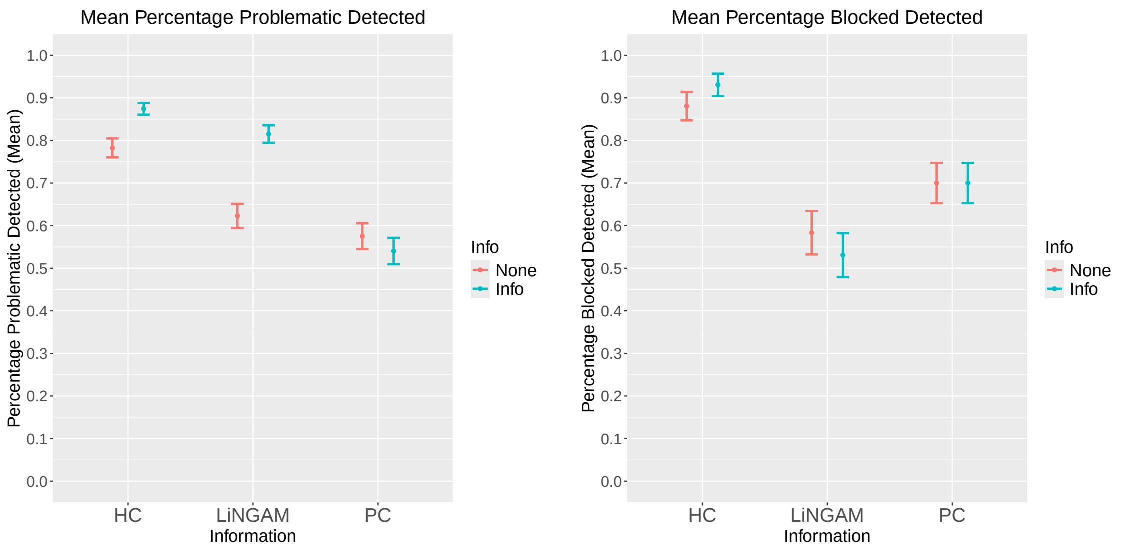
(Problematic) **(Blocked)**

Fig. 7. Results for **RQ4** and problematic variables as well as blocked variables summarized across all DAGs separated by learning methods.

we can observe that for HC and LiNGAM, providing information leads to a huge benefit. As before, information does not really help for PC.

Table 6 displays the results of **RQ6**, which shows the impact of using background knowledge on the estimation of causal fairness notions. We only show the results for HC, as we have already seen that HC outperforms the other methods. We need to compute NDE and NIE separately for each sensitive variable, which we have abbreviated here. We show the results for all variables that we defined as sensitive in Table 1. As explained before, the values represent the ratio of differences of the estimated values between learned DAGs with and without information compared to the ground-truth DAG. Hence, values smaller than 1 indicate that the estimate of the fairness values using the DAGs learned with background information is closer to the estimate using the ground-truth DAG, and values larger than 1 indicate that the fairness values using the DAGs learned without background information are closer to the estimate using the ground-truth DAG. If using the ground-truth DAG and both of the learned DAGs estimate an effect of 0, we wrote “Effect Size 0 in all”. An effect of 0 occurs if the sensitive variable is unproblematic, i.e., completely disconnected from the subnetwork of which the target variable is a part.

Generally, we can observe that for 52 variable-value combinations, the ratio is smaller than 1, indicating that the estimate from the learned DAG using background information is closer to the estimate using the ground-truth DAG. 19 values are larger than 1; the remaining are either exactly 1 (one occurrence) or no effect was calculated using any DAG. These results have two important implications. The first is that our results support previous research that already highlighted that the DAGs are very important for estimating causal fairness (Binkytė-Sadauskienė et al., 2022). Different DAGs lead to very different results (as indicated by values much smaller and larger than 1). The second is that our **RQ6** can generally be answered affirmatively, although there are some exceptions.



(Unproblematic) **(Opposing Effects)**

Fig. 8. Results for **RQ4** and unproblematic variables as well as opposing effects variables summarized across all DAGs separated by learning methods.

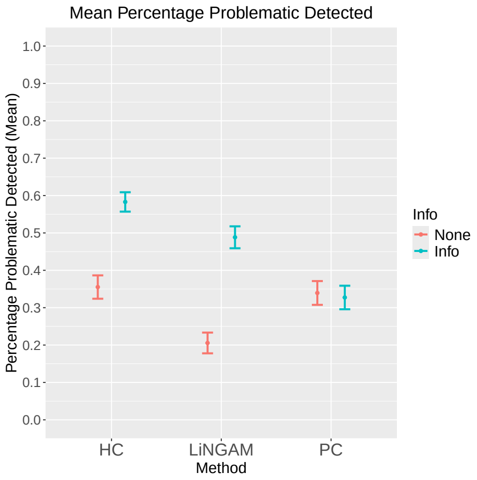


Fig. 9. Results for **RQ5** summarized across all DAGs separated by learning methods.

8 Evaluation on Real Data

As already mentioned, our evaluation of the methods on synthetic data has limitations concerning the artificiality with which the data is produced. Unbalanced, non-linear data is particularly challenging for structure learning algorithms. To show that our methods are also applicable in realistic settings, we now turn to a real dataset that

Table 6. Results for **RQ6** and HC, summarized across all DAGs. The values represent the ratio of differences of the estimated values between the learned DAGs with and without information compared to the ground-truth DAG.

Network	Variable	Relation of Error NDE	Percentage of Error NIE
alarm	v1	0.705	0.998
	v2	0.941	6.357
	v3	0.742	0.028
	v4	1.241	0.669
	v5	0.959	0.968
	v6	0.908	0.986
	v7	0.744	0.229
	v8	6.059	0.684
	v9	0.870	1.001
	v10	Effect Size 0 in all	Effect Size 0 in all
	v11	2.167	Effect Size 0 in all
asia	v1	2.107	6.697
	v2	0.0824	0.031
earthquake	v1	15.236	0.030
	v2	4.289	0.031
insurance	v1	0.470	0.629
sachs	v1	0.077	0.317
	v2	Effect Size 0 in all	Effect Size 0 in all
SD I	v1	1.019	0.838
	v2	0.026	0.037
	v3	0.810	1.299
	v4	0.043	0.041
SD II	v1	3.770	2.909
	v2	0.967	20.454
	v3	0.117	0.021
	v4	0.935	0.931
SD III	v1	0.598	0.379
	v2	0.475	0.337
	v3	0.993	1.146
	v4	5.082	2.071
	v5	0.003	0.133
	v6	0.881	1.853
SD IV	v1	0.005	0.004
	v2	Effect Size 0 in all	Effect Size 0 in all
	v3	0.537	0.758
	v4	0.000	0.000
SD V	v1	0.000	Effect Size 0 in all
	v2	0.709	0.589
	v3	1.000	1.000
	v4	0.126	0.000

we know well enough to state whether the results appear sensible. The results for two other prominent datasets used in the fairness literature can be found in Appendix B.

The dataset is concerned with predicting whether students pass a particular course on the first try. The course is a mandatory part of a Bachelor's program at our university and should be taken in the third semester. It is

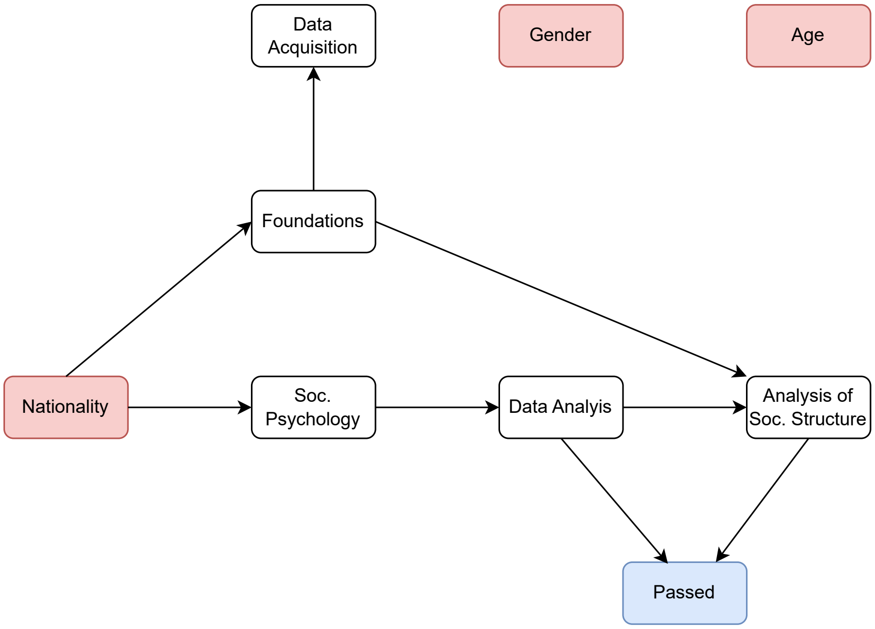


Fig. 10. The part of the returned graph that shows the demographic features and the subnetwork that includes the target.

concerned with statistics. 16% of students do not pass this course on the first try. To predict whether a person passes the course on the first attempt, we can use information regarding which other mandatory courses were passed already by the time the student takes the exam. To check whether fairness concerns may come into play when training an ML model on this data, we include demographic information we have available: age, gender, and nationality. We have student data available for the past ten years, which has led to 811 students in our final sample.

Because of our previous results, we used HC on the data and provided information on the target as well as the demographic information. For a better overview, Figure 10 shows only a part of the returned graph, including the component that includes the target (the target is in blue) and the demographic features (the demographic features are in red). The other courses not displayed are not connected to any of the courses or demographic features displayed here, but are interconnected with each other.

We can observe that gender and age are disconnected from all other nodes; no edge has been learned that connects them to any of the other variables on a path toward the target node or the target node itself. Hence, they are unproblematic according to our classification in Section 3, although leaving them out of ML models is likely still sound advice. Further, we can observe that although the target is not directly connected to nationality, an indirect relationship exists. Nationality seems to directly influence the courses *Foundations* and *Social Psychology I*. Both courses are supposed to be taken before taking the target course. Both courses also require more language skills than the other courses. *Foundations* teaches basic theories and is not concerned with statistics at all; *Social Psychology I* teaches both basic theories and employs statistical methods. It is likely that, in reality, people of other nationalities than the domestic nationality have poorer language skills and struggle in these courses. This theory points to another aspect and limitation we already mentioned. We need to rely on the variables that we actually have and construct graphs from them. If other relevant variables (language skills in this case) are missing, then the resulting graph may not necessarily depict real causal relationships.

Going back to the graph, *Social Psychology I* is influencing *Data Analysis*. It is not surprising that these two are connected as *Social Psychology I*, as already stated, introduced statistical methods as well. Both *Data Analysis* and *Foundations* influence *Analysis of the Social Structure*. For us, this also appears reasonable as this course builds on these two and uses the theories taught in *Foundations* and combines them with statistics taught in *Data Analysis*. *Data Analysis* and *Analysis of the Social Structure* both influence the target course. For *Data Analysis*, any other result would have been surprising as the target course is the de facto advanced course of *Data Analysis*. That *Analysis of the Social Structure* also directly influences the target course is interesting. This finding shows a well-known advantage of structure learning algorithms: They show which variables are relevant and which are not (many courses are also not relevant at all for this course).

What do the results mean from a fairness perspective? For age and gender, we have already stated that these are unproblematic as they are unconnected to the subnetwork, which includes the target node. Still, as mentioned, they should probably be excluded. For nationality, the situation is obviously different. While there is no direct relationship, information about the nationality is transported to the target variable via other variables. This essentially affects all variables that are connected directly and indirectly with the target, and we suspect that this will often be the case in real-life data. If we plan to learn an ML model on the data that is supposed to predict whether a student passes, we need to control for biases introduced by nationality. What exactly we control for depends both on the aim of the model and the stakeholders' perspectives. In this case, the stakeholders are the courses' teachers and previous students. If the model is aimed at identifying who is struggling with the course and then providing personalized assistance, keeping all variables in the model may be a valid decision. If the model is used to prioritize who gets accepted to the course, the stakeholders may decide that at least nationality needs to be taken out of the data and, potentially, that the remaining bias through the other variables needs to be mitigated. As an example, stakeholder discussion may reveal that the effect of nationality through *Foundations* is seen as illegitimate (e.g., because it requires more and different language skills), whereas the effect of nationality through the other courses is seen as fair. Then, path-specific bias mitigation could be employed and monitored with a path-specific bias measure.

One aspect that should be highlighted in this scenario is the following: Stakeholders are probably much more likely to agree that language skills are a legitimate reason for different outcomes compared to nationality. Here, however, it is likely that nationality is a proxy for language skills. Then again, we cannot be absolutely certain of this. This consideration highlights that knowing more about the setting and the data and having many variables available (particularly demographic ones) is very valuable. It also highlights that we have to be careful when talking and thinking about our decisions.

9 Limitations, Implications, and Future Work

Although our results are promising, this paper has also revealed a number of limitations, aspects that require caution, implications for applying the methods discussed, and future research directions.

9.1 Limitations and Implications

While we have shown that including background helps to learn more accurate DAGs, which in turn helps to classify the variables according to the classes we identified and to estimate unfairness better, it is still not the case that the DAGs are always learned correctly. Learning DAGs is tricky, and several factors contribute to the ease with which they can be learned. In particular, complex relationships or a lack of data hinder the learning of DAGs, which is amplified in settings with many variables. As we have seen in our results, having more data available makes it more likely that accurate DAGs are learned. Thus, causal structure learning may be unreliable in sparse data settings. Moreover, we must always be aware that our learned DAG may not be accurate and that we should be careful when using it to inform future actions. As our approach is supposed to improve the estimation of fairness and to allow better bias mitigation, wrong DAGs can have problematic outcomes.

Moreover, as mentioned before, DAGs can only be learned from existing data. If variables that would be part of the causal network are not observed, the edges in the learned network do not directly or necessarily imply a causal relationship. As we see in Figure 1, the absence of M changes the graph and the relationships drastically. To a degree, we deal with this issue by having the opposing effects class, which exists precisely for this. Nonetheless, we should be very careful with conclusions drawn from the graph and should think about potentially missing variables and how they would be included in the graph or change its construction.

Another limitation, particularly relevant for real-life data, and as observed with the data example given above, is that it may be the case that the sensitive variables influence all paths leading to the target. Even if we decide that some of the paths are fine, this still heavily impacts what we can actually do with the predictive model and what limitations we impose. As Nilforoshan et al. (2022) write, this can lead to extreme cases, where, e.g., we decide that a model is only fair if everyone receives the same treatment. We believe that future work should take a holistic approach to investigate this problem and make recommendations. If the data is very biased, the decision may have to be that we cannot create ML models using this data.

Just as important as learning accurate graphs, and connected to it, is the discussion of who decides which paths are unproblematic. In general, the process we describe in this work requires human involvement, and this involvement can influence the results drawn heavily. Clearly, in order to decide which paths are allowed, background knowledge concerning the domain is required. What follows is that domain experts need to be included in the decision-making process. However, we argue that this should not be seen as a nuisance but as a strength. The more background knowledge we have available, the more accurate the learned DAGs will be, and the better our estimation of ML unfairness will be. Hence, including domain experts in all parts of the process is very valuable. However, given that the consequences of the insights provided by domain experts can be quite far-reaching, it is important to discuss who should be consulted as a domain expert. What we need to avoid is the danger of "fair-washing". Concretely, we need domain experts who do not define each path in the DAG as unproblematic, which would, consequently, lead to a ML defined as fair despite potentially not being fair. We believe that the best strategy to avoid this problem is not to have a singular domain expert but rather a group of different stakeholders. For example, if our ultimate aim is to train an ML model to detect potential university drop-outs early on, stakeholders could be current and former students, teaching personnel, and administrative staff. With each group having slightly different interests and perspectives, the most and probably the best background knowledge can be gathered, and resulting fairness estimations are the most likely to be of real value. The main obvious issue with including different groups of stakeholders in the process is not just that members of each group need to agree to participate–if they all see value in the ML tool, their participation is likely a given– but rather that they need some introduction into causal modeling and DAGs. This requirement presents a time-consuming hurdle, although some workshops have been constructed aimed to quickly teach stakeholders the tools necessary for causal modeling Rodrigues et al. (2022). In general, we argue that the involved domain experts should be part of different stakeholder groups. However, we need to achieve a good balance between including different perspectives and limiting the temporal effort for all involved. Of course, an issue may then be that stakeholders disagree. However, varying points of view can even be seen as valuable insights, and there are strategies to reconcile differing opinions Rodrigues et al. (2022), although this problem is still an ongoing research area. We consider a deeper discussion of how to obtain expert knowledge as out of scope of this paper, but point interested readers to Rodrigues et al. (2022), Taylor et al. (2024), and Barbrook-Johnson and Penn (2022).

Taking a step back to view an ML project more globally, the guided involvement of domain experts provides many advantages Cohausz (2025). Ideally, a first fairness assessment is done on the data before a ML model is even constructed, and domain experts, who are also much better able to assess the role of the context in which the model is employed, should be consulted then. As such, they are involved from a very early point onward, which also allows them to express their views on whether the ML tool is even a good idea, on whether all important variables are part of the dataset, and on which variables are likely to be very important (allowing a partially manual feature selection). The fairness assessment should then continue until the final model to be deployed is found. Here, again, it is probably very helpful to use the stakeholders' insights not only concerning the fairness of the model but also their views on the most important measures (e.g., whether recall is more important than precision). Hence, the involvement of stakeholders can be of even broader interest than just from the perspective of fairness and ideally encompasses all time steps of ML model development and deployment. In general, it may be helpful to think of fairness not as a step of ML model development but rather as one dimension concerning the whole project and intimately related to other dimensions. Then, it also becomes obvious that thinking about fairness and asking for domain experts' perspectives also touches other aspects of ML such as performance or feature selection.

9.2 Future Work

Precisely due to the problem that different people may provide different information, we believe that it is very important for future research to define some ground truth DAGs with agreed-upon background information to evaluate fairness and test bias mitigation methods. The need for such agreed-upon settings has already been highlighted by others (Plečko and Meinshausen, 2020). Additionally, it would be interesting to research the extent to which the background information provided varies between different stakeholder groups. Such research could enable us to estimate how reliable experts are and how to reconcile varying views.

Moreover, future research should focus on achieving even more accurate DAGs. Our results show that the inclusion of background information is very helpful. Hence, future work should attempt to develop adaptations to include background information even more effectively. Additionally, and as indicated by our results, a focus should be placed on score-based methods as these achieve the most promising results. In a similar direction, a thorough investigation of how difficult (sparse, complex) data settings impact the accurate learning of DAGs and to what extent background knowledge can help counter it, would be interesting.

Lastly, as it is unrealistic to always achieve perfect DAGs, more research should be conducted to investigate the extent to which fairness measures are affected by varying levels of errors in structure learning.

10 Conclusion

In our paper, we made two major contributions. The first is to provide a classification of different structures in DAGs in which sensitive variables can be involved. We highlighted that one class of structures (problematic variables), especially, requires a thorough and case-by-case evaluation regarding whether and to what extent the influence of the sensitive variable is okay. As others have already stated, there may be cases for which we believe some paths to transport legitimate and some paths to transport illegitimate information, which shows the importance of path-specific bias mitigation (Chiappa and Isaac, 2019). Some work already deals with path-specific bias mitigation in a sophisticated way (Madras et al., 2019; Grari et al., 2021). We believe that our work makes it more apparent when such a strategy should be used. Our paper's second contribution is the adaptation of structure learning algorithms by using background information available in fairness scenarios. We show that the adaptations lead to more correctly learned DAGs, which greatly improves the usefulness of DAGs for fairness, as no reliable conclusions can be drawn otherwise (Binkytė-Sadauskienė et al., 2022). We also show that the method can be used on real data. Although there are still some limitations, we contributed important results for a more causal-driven fairness evaluation.

Acknowledgments

Lea Cohausz is funded by the grant "Consequences of Artificial Intelligence for Urban Societies (CAIUS)" by the Volkswagen Foundation. We thank our colleague Ricarda Link for her helpful feedback and the reviewers for their insightful comments.

References

Baker, R. S., Esbenshade, L., Vitale, J., Karumbaiah, S., et al. (2023). Using demographic data as predictor variables: a questionable choice. *Journal of Educational Data Mining*, 15(2):22–52.

Barbrook-Johnson, P. and Penn, A. S. (2022). Participatory systems mapping. In *Systems Mapping: How to build and use causal models of systems*, pages 61–78. Springer.

Becker, B. and Kohavi, R. (1996). Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

Binkytė-Sadauskienė, R., Makhlouf, K., Pinzón, C., Zhioua, S., and Palamidessi, C. (2022). Causal Discovery for Fairness. arXiv:2206.06685 [cs, stat].

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1):4209.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7801–7808.

Chiappa, S. and Isaac, W. S. (2019). A causal bayesian networks viewpoint on fairness. *Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data: 13th IFIP WG 9.2, 9.6/11.7, 11.6/SIG 9.2. 2 International Summer School, Vienna, Austria, August 20-24, 2018, Revised Selected Papers 13*, pages 3–20.

Cohausz, L. (2025). Why the future of aied is causal: Arguments for creating a tradition based on causal thinking. In *International Conference on Artificial Intelligence in Education*, pages 17–31. Springer.

Cohausz, L., Kappenberger, J., and Stuckenschmidt, H. (2024). What Fairness Metrics Can Really Tell You: A Case Study in the Educational Domain. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, pages 792–799, Kyoto Japan. ACM.

Constantinou, A. C., Guo, Z., and Kitson, N. K. (2023). The impact of prior knowledge on causal structure learning. *Knowledge and Information Systems*, pages 1–50.

Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K. (2021). Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning*, 131:151–188.

Deho, O. B., Zhan, C., Li, J., Liu, J., Liu, L., and Duy Le, T. (2022). How do the existing fairness metrics and unfairness mitigation algorithms contribute to ethical learning analytics? *British Journal of Educational Technology*, 53(4):822–843.

Dolata, M., Feuerriegel, S., and Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal*, 32(4):754–818.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, Cambridge Massachusetts. ACM.

Gerdon, F., Bach, R. L., Kern, C., and Kreuter, F. (2022). Social impacts of algorithmic decision-making: A research agenda for the social sciences. *Big Data & Society*, 9(1):205395172210893.

Grari, V., Lamprier, S., and Detyniecki, M. (2021). Fairness without the sensitive attribute via causal variational autoencoder. *arXiv preprint arXiv:2109.04999*.

Grari, V., Lamprier, S., and Detyniecki, M. (2023). Adversarial learning for counterfactual fairness. *Machine Learning*, 112(3):741–763.

Green, B. and Hu, L. (2018). The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Hicks, B., Kitto, K., Payne, L., and Buckingham Shum, S. (2022). Thinking with causal models: A visual formalism for collaboratively crafting assumptions. In *LAK22: 12th International Learning Analytics and Knowledge Conference*, pages 250–259.

Jacobs, A. Z. and Wallach, H. (2021). Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada. ACM.

Kilbertus, N., Rojas Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. *Advances in neural information processing systems*, 30.

Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., and Chobtham, K. (2023a). A survey of bayesian network structure learning. *Artificial Intelligence Review*, pages 1–94.

Kitson, N. K., Constantinou, A. C., Guo, Z., Liu, Y., and Chobtham, K. (2023b). A survey of Bayesian Network structure learning. *Artificial Intelligence Review*, 56(8):8721–8814.

Kuehnemund, L., Koeppe, J., Feld, J., Wiederhold, A., Illner, J., Makowski, L., Gerß, J., Reinecke, H., and Freisinger, E. (2021). Gender differences in acute myocardial infarction—A nationwide German real-life analysis from 2014 to 2017. *Clinical Cardiology*, 44(7):890–898.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. (2018). Causal Reasoning for Algorithmic Fairness. arXiv:1805.05859 [cs].

Ma, J., Guo, R., Zhang, A., and Li, J. (2023). Learning for counterfactual fairness from observational data. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1620–1630.

Maasch, J., Gan, K., Chen, V., Orfanoudaki, A., Akpinar, N.-J., and Wang, F. (2024). Local Causal Discovery for Structural Evidence of Direct Discrimination. arXiv:2405.14848 [cs, stat].

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2019). Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 349–358.

Makhlouf, K., Zhioua, S., and Palamidessi, C. (2024). When causality meets fairness: A survey. *Journal of Logical and Algebraic Methods in Programming*, 141:101000.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6):1–35.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2021). Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163.

Nilforoshan, H., Gaebler, J. D., Shroff, R., and Goel, S. (2022). Causal conceptions of fairness and their consequences. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th*

International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, pages 16848–16887. PMLR.

Pearl, J. (2009). *Causality*. Cambridge university press.

Plečko, D. and Meinshausen, N. (2020). Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44.

Rodrigues, D., Kreif, N., Lawrence-Jones, A., Barahona, M., and Mayer, E. (2022). Reflection on modern methods: constructing directed acyclic graphs (dags) with domain experts for health services research. *International Journal of Epidemiology*, 51(4):1339–1348.

Räz, T. (2021). Group Fairness: Independence Revisited. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 129–137. arXiv:2101.02968 [cs].

Scanagatta, M., Salmerón, A., and Stella, F. (2019). A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence*, 8:425–439.

Scutari, M., Graafland, C. E., and Gutiérrez, J. M. (2019a). Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253.

Scutari, M., Scutari, M. M., and MMPC, H.-P. (2019b). Package 'bnlearn'. *Bayesian network structure learning, parameter learning and inference, R package version*, 4(1).

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 59–68, Atlanta GA USA. ACM.

Shimizu, S. (2014). Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41:65–98.

Stinar, F. and Bosch, N. (2022). Algorithmic unfairness mitigation in student models: When fairer methods lead to unintended results. Publisher: Zenodo.

Taylor, H., Crabbe, H., Humphreys, C., Dabrera, G., Mavrogianni, A., Verlander, N. Q., and Leonardi, G. S. (2024). Development and use of a directed acyclic graph (dag) for conceptual framework and study protocol development exploring relationships between dwelling characteristics and household transmission of covid-19–england, 2020. *Building and Environment*, 250:111145.

Varshney, K. (2018). Introducing ai fairness 360. *IBM Research blog*.

Zhang, L., Wu, Y., and Wu, X. (2016). A causal framework for discovering and removing direct and indirect discrimination. *arXiv preprint arXiv:1611.07509*.

A Examples

Figure 11 shows the same structures as Figure 1 but with meaningful variable names. We will now briefly explain the example scenarios that the networks encode. Most of the examples are taken from Cohausz et al. (2024) and are only supposed to provide more intuitively understandable explanations of the structures.

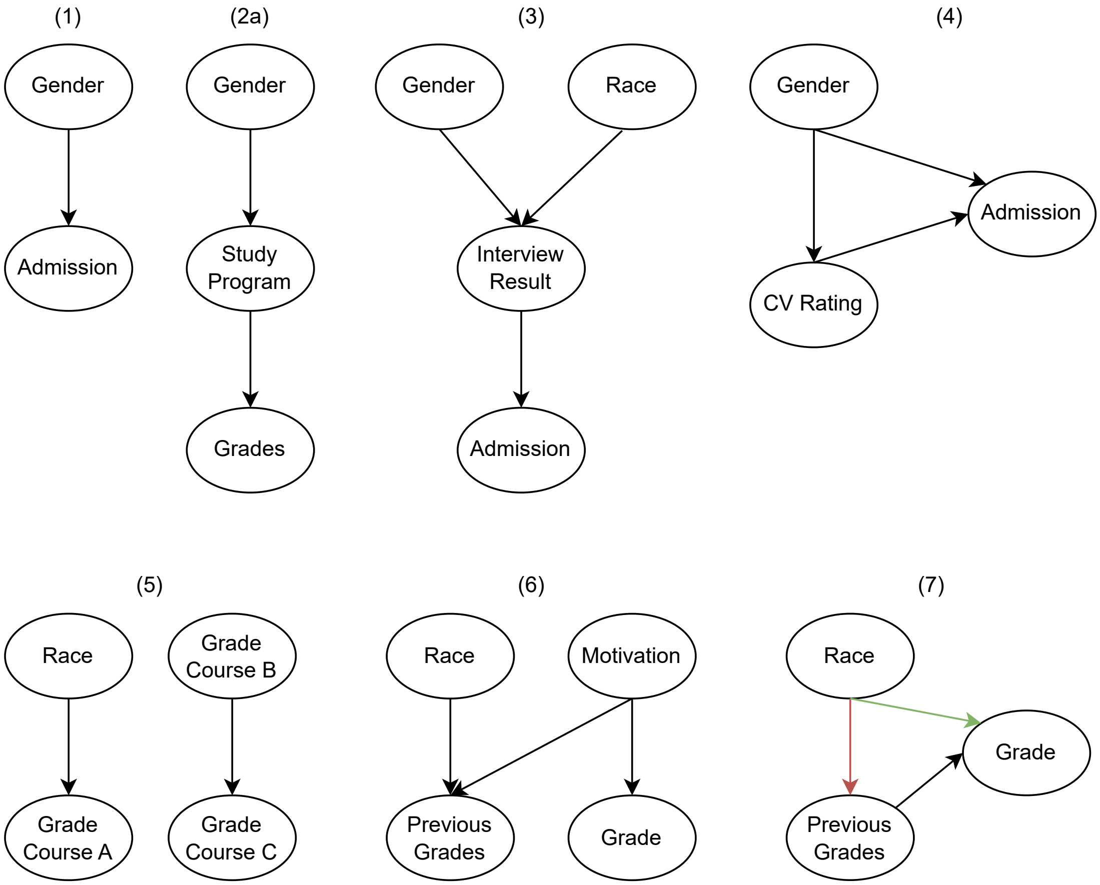


Fig. 11. The structures from Figure 1 but with meaningful variable names.

The first structure (1) may represent a university admission process in which the person making a decision discriminates against a gender, thus directly impacting admission into a program. The second structure encodes a scenario where gender impacts the choice of study program, which affects the grades as different study programs may be more or less difficult or award better or worse grades on average. The third structure again represents a university admission scenario. Gender and race impact the interview rating, and the interview rating determines admission. One may observe that when considering these specific scenarios, the second may be seen as unproblematic, and the third may be seen as problematic despite being very similar. The fourth structure is again a university admission scenario where gender impacts the rating of the CV (e.g., due to hobbies mentioned), and the CV impacts admission; gender impacts admission directly due to discrimination. The fifth structure shows a scenario in which race impacts the grade in one course, but for our target course (course C), only the grade from another course (course B) is relevant. The sixth structure shows a scenario in which–due to discrimination–race impacts the grades in previous courses. The grades in both previous and current courses are also affected by students' motivation. The current course's grade is not determined by race at all (no discrimination against a race happens). Finally, the seventh structure represents the same scenario but with motivation being unobservable, which–when applying structure learning algorithms–will result in the structure depicted here.

B Additional DAGs

The two DAGs below in Figures 12 Becker and Kohavi (1996) and 13 Varshney (2018) were learned using HC and using background knowledge (i.e., specifying the target and the sensitive variables). Figure 12 depicts the DAG learned using the adult dataset, and Figure 13 depicts the DAG learned using the law school dataset. Both are frequently used in the (causal) fairness literature (Kusner et al., 2017; Ma et al., 2023; Grari et al., 2023). Note that our learned adult DAG is very close to the adult causal DAG created by Zhang et al. (2016) and frequently used in the literature.

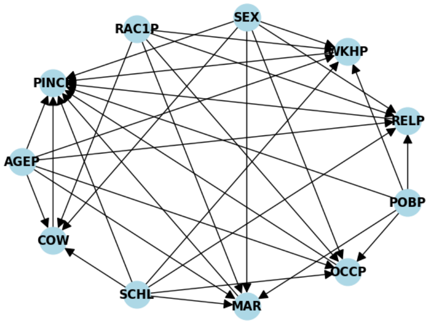


Fig. 12. The DAG created from the adult dataset.

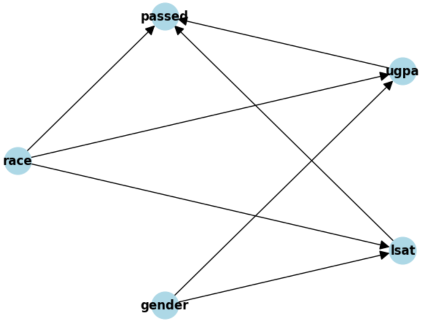


Fig. 13. The DAG created from the law school dataset.

C Additional Results: Relative Change Ratio

As an additional way to analyze the impact of having background information available compared to not having it available, we compute the relative change ratio for the true positive (TPR) and false positive (FPR) rates, respectively. We define the relative change ratio for the TPR as:

$$C_{TPR} = \frac{TPR_{info} - TPR_{noinfo}}{TPR_{noinfo}}$$

The relative change ratio for the FPR is computed accordingly. We then average across all data settings and DAGs but separate by method. We compute the average and the 95% confidence interval.

The results can be seen in Figures 14 and 15. We can clearly see that HC and LiNGAM benefit heavily from background information. Interestingly, the change in the TPR is more pronounced than the change in the FPR–although for both there are clearly significant effects–highlighting that the effect of having information propagates to parts of the structure on which we do not have information.

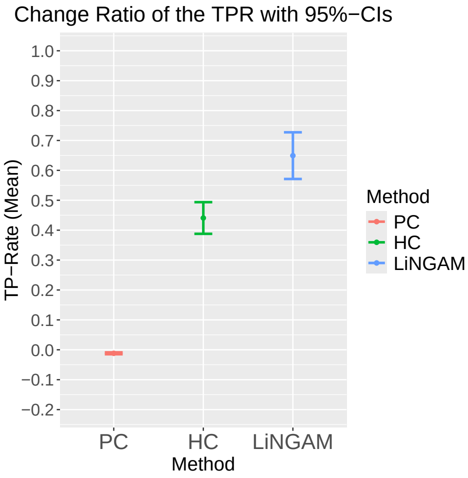


Fig. 14. The plot shows the average change (95% confidence intervals) in the TPR when having background information available compared to not having it for each of the three methods.

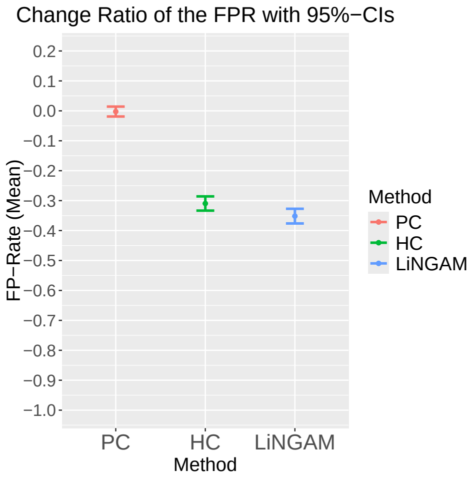


Fig. 15. The plot shows the average change (95% confidence intervals) in the FPR when having background information available compared to not having it for each of the three methods.

received 7 May 2025; accepted 23 September 2025