Counterfactual Fairness with Imperfect Causal Graphs

Cong Su[1,2], Qiaoyu Tan[3,4], Carlotta Domeniconi[5], Lizhen Cui[1,2], Jun Wang[1,2], Guoxian Yu[1,2*]

[1]School of Software, Shandong University, Jinan, China
[2]SDU-NTU Joint Centre for AI Research, Shandong University, Jinan, China
[3]NYU Shanghai Center for Data Science, Shanghai, China
[4]New York University Shanghai, Shanghai, China
[5]Department of Computer Science, George Mason University, VA, USA
csu@mail.sdu.edu.cn, qiaoyu.tan@nyu.edu, {kingjun, clz, gxyu}@sdu.edu.cn, carlotta@cs.gmu.edu

Abstract

Fairness-aware machine learning aims to build predictive models that comply with fairness requirements, particularly concerning sensitive attributes such as race, gender, and age. Among causality-based fairness notions, counterfactual fairness is widely adopted for its individual-level guarantees, requiring that an individual's predicted outcome remains unchanged in a counterfactual world where its sensitive attribute is altered. However, existing methods critically assume that the true causal graph is fully known, which is rarely the case in practice. Moreover, counterfactual fairness suffers from inherent identifiability limitations, as counterfactual quantities cannot always be uniquely estimated from observational data, especially under incomplete causal knowledge. To address these challenges, we propose a principled framework (`CF-ICG`) for counterfactual fairness under imperfectly known causal graphs, e.g., Completed Partially Directed Acyclic Graphs (CPDAGs). We first introduce a criterion to determine the identifiability, and bound the counterfactual quantities under CPDAGs. Building upon this, we develop an efficient local algorithm that avoids the exhaustive enumeration of all DAGs, ensuring robustness against worst-case fairness violations. Experimental results on synthetic and real-world datasets demonstrate the practical effectiveness and theoretical soundness of `CF-ICG`.

Extended version and code —
https://www.sdu-idea.cn/pubDetail?pubId=305

Introduction

Ensuring algorithmic fairness has become imperative as machine learning models are increasingly deployed in sensitive decision-making domains. Numerous fairness-aware methods have been proposed to mitigate discrimination against individuals or groups characterized by sensitive attributes such as gender, race, and age (Su et al. 2022; Rabonato and Berton 2025). These methods can be broadly divided into two categories. The first category focuses on *statistical fairness*, which aim to enforce statistical independence between sensitive attributes and outcomes (Chouldechova 2017; Jin et al. 2023; Ma et al. 2023; Jin, Li, and Feng 2024). The second category is grounded in *causal inference frameworks* (Pearl et al. 2000), which explicitly model and mitigate the causal effects of sensitive attributes on model predictions, using tools like the *do* operator (Wang et al. 2024; Su et al. 2025).

Among causality-based fairness notions, counterfactual fairness (Kusner et al. 2017) stands out for its individual-level fairness guarantee, requiring that a model's prediction for an individual, specified by context variables, remains invariant under hypothetical counterfactuals on its sensitive attribute. This formulation is more general and challenging than interventional fairness, as it necessitates reasoning across both the factual and counterfactual worlds. However, this also introduces intrinsic identifiability challenges, where counterfactual quantities may not be computable from observational data alone, posing a serious barrier to its real-world application (Pearl 2009).

Although several efforts have been made to address identifiable issues of counterfactual fairness, most of them assume the availability of a *well-defined causal Directed Acyclic Graph* (DAG) (Kusner et al. 2017; Wu, Zhang, and Wu 2019; Xu et al. 2019; Chiappa 2019; Huang, Zhang, and Wu 2022; Robertson et al. 2024), an assumption that is often unrealistic in practice. More recently, a few approaches have attempted to relax this requirement by utilizing imperfect causal graph (e.g., Completed Partially Directed Acyclic Graphs (CPDAGs)), which can be obtained using causal discovery algorithms (Zanga, Ozkirimli, and Stella 2022; Zheng et al. 2024). They can be categorized into: 1) Feature selection approaches (Zuo et al. 2022), which identify non-descendants of the sensitive attribute and limit model inputs accordingly; and 2) Causal inference approaches (Grari, Lamprier, and Detyniecki 2023), which leverage variational autoencoders (VAEs) to make counterfactual inference. However, these methods either drastically reduce model utility by excluding all descendants of sensitive attributes, or rely on strong structural assumptions to handle identifiability. More discussion of related work is presented in Extended version.

Consequently, it is imperative to develop more practical and robust approaches that can responsibly incorporate certain descendants of sensitive attributes under CPDAGs without compromising counterfactual fairness. However, achiev-

*Corresponding author: Guoxian Yu

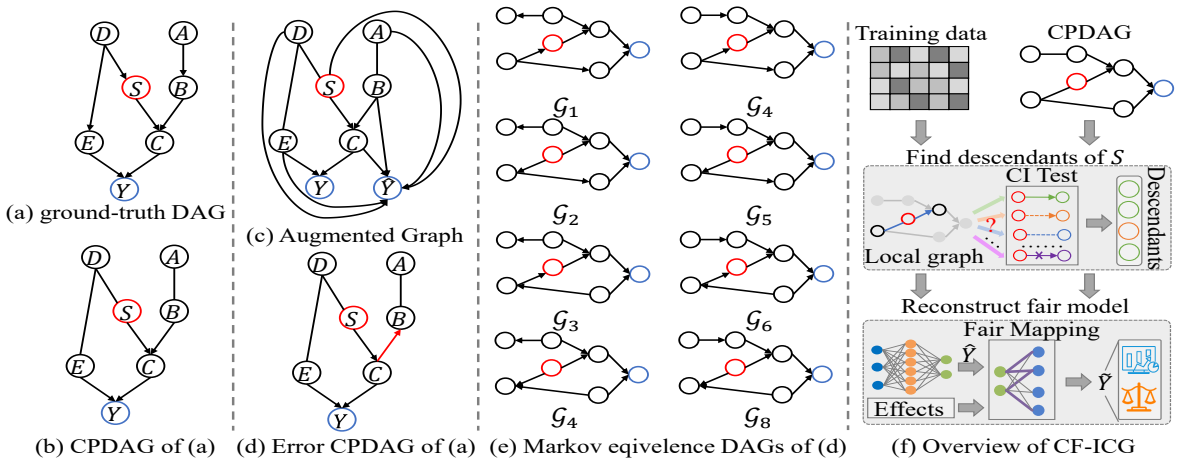


Figure 1: Illustrative example where S is the sensitive attribute (highlighted in red) and Y is the outcome attribute (highlighted in blue). (a) is the true causal graph, (b) is its CPDAG, (c) is the augmented graph with $\hat{Y}$, (d) is the error CPDAG of (b) (highlighted in red directed edge), and (e) lists all Markov equivalence DAGs of (d); and (f) shows our `CF-ICG` framework, where it first finds the descendants of sensitive attribute S using only local causal structure around S, and then reconstructs the models with worst-case counterfactual fairness guarantees.

ing this is challenging due to several fundamental obstacles. *First*, the identifiability of counterfactual quantities under CPDAGs remains a major hurdle. A straightforward yet impractical strategy is to enumerate all DAGs within the Markov equivalence class to evaluate all possible counterfactual effects (Wu, Zhang, and Wu 2019; Wu et al. 2019). As illustrated in Figure 1, where (a) depicts the true causal graph, (b) shows its CPDAG, and (e) enumerates all corresponding DAGs ($\mathcal{G}_1, \cdots, \mathcal{G}_8$), counterfactual quantities can be estimated by exhaustively analyzing each graph. This strategy, however, is computationally prohibitive and becomes infeasible as the size of the equivalence class increases. *Second*, identifying counterfactual quantities relies on the accuracy of descendant recognition as shown in Proposition 1, which is sensitive to noise or errors in the graph structure. Even subtle mistakes, such as the erroneous reversal of the edge between B and C in Figure 1, can lead to incorrect classification of variables as descendants or non-descendants of the sensitive attribute S (Zuo et al. 2022). This misclassification leads to biased or invalid counterfactual estimations. Together, these challenges underscore the difficulty of estimating counterfactual effects without full causal knowledge, where both the complexity of graph enumeration and the brittleness of descendant identification critically limit the practicality of existing counterfactual fairness methods.

To address these challenges, we propose `CF-ICG`, a principled and efficient local framework for achieving Counterfactual Fairness under Imperfect Causal Graphs, such as CPDAGs. The overview of `CF-ICG` is depicted in Figure 1 (f). Our key **innovations** are given as below:

- **Identification Criterion**. Identifiability is a critical barrier for counterfactual fairness to be applied in real applications. We are the first to develop a theoretical criterion (Proposition 1) to determine whether counterfactual fairness is identifiable under CPDAGs. Importantly, our analysis reveals that it is sufficient to traverse the descendant sets of the sensitive attribute (e.g., $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_4$ in Figure 1 (e)) , rather than exhaustively enumerating all DAGs within the CPDAG.
- **Efficient Local Bounding Algorithm**. Building on above insight, we introduce a local method to find the descendants of the sensitive attribute and compute the lower and upper bounds of counterfactual fairness violations, leveraging only the local structures around the sensitive attribute, thus avoiding global graph search.
- **Robust Model Reconstruction**. We further propose a robust learning algorithm that reconstructs decision models with worst-case counterfactual fairness guarantees, ensuring responsible feature usage even under causal ambiguity.
- **Empirical Validation**. Extensive experiments on synthetic and real-world datasets demonstrate the effectiveness of our approach, achieving strong fairness guarantees while maintaining competitive predictive performance.

Preliminaries

Structural Causal Model

A structural causal model (SCM) (Pearl 2009) is a framework to model causal relations between variables. It is defined as a triple $(\mathcal{U}, \mathcal{V}, \mathcal{F})$, where $\mathcal{V}$ is the set of observable endogenous variables and $\mathcal{U}$ is the set of unobservable exogenous variables that cannot be caused by any variable in $\mathcal{V}$. $\mathcal{F}$ is the set of functions $\{f_1, f_2, \cdots, f_{|V|}\}$, each of which is associated with a variable $V_i \in \mathcal{V}$ describing how V_i depends on its direct causes, i.e., $V_i = f_i(pa(V_i), \mathcal{U}_i)$, where $pa(V_i)$ denotes the observable direct causes of V_i and $\mathcal{U}_i$ is the set of unobservable direct causes of V_i. The set of equations $\mathcal{F}$ induces a causal graph $\mathcal{G}$ that represents the relationships between variables, typically in the form of a directed acyclic graph (DAG), where the direct causes of V_i

correspond to its parent set in the causal graph.

A DAG $\mathcal{D}$ is represented by directed edges without any directed cycles. When all edges in the causal graph are a mixture of directed and undirected edges, we say the graph $\mathcal{G}$ is a partially directed graph (PDAG). Any two DAGs are Markov equivalent if they share the same conditional independence relations (Pearl 2014). A Markov equivalence class is uniquely represented by a completely partially directed acyclic graph (CPDAG) $\mathcal{G}^*$, denoted as $[\mathcal{G}^*]$, which includes all DAGs equivalent to $\mathcal{G}^*$.

The skeleton of $\mathcal{G}$ is an undirected graph obtained by removing all arrowheads from $\mathcal{G}$. Given a graph $\mathcal{G}$, V_i is called a parent of V_j and V_j is called a child of V_i if $V_i \to V_j$ in $\mathcal{G}$. Also, V_i is a sibling of V_j if $V_i - V_j$ in $\mathcal{G}$. We denote $pa(V_i, \mathcal{G})$, $ch(V_i, \mathcal{G})$, $sib(V_i, \mathcal{G})$ and $adj(V_i, \mathcal{G})$ as the sets of parents, children, siblings, and adjacent vertices of V_i in $\mathcal{G}$, respectively. A variable V_i is an ancestor of V_j and V_j is a descendant of V_i if there exists a directed path from V_i to V_j or $V_i = V_j$.

Counterfactual Fairness

Given a DAG $\mathcal{G}$ and two distinct variables X and Y, the causal effect of X on Y can be interpreted by the post-intervention distribution of Y intervening on X via do operator (Pearl 2009). Formally, given a distribution $P(\mathcal{U})$ over the exogenous variables $\mathcal{U}$, an intervention on X, $do(X = x)$, which forces variable X to take certain value x, is defined as the substitution of the structural equation $X = f_X(pa(x), \mathcal{U}_X)$ with $X = x$, and the post-interventional density of Y is denoted as $f(Y = y|do(X = x))$ or $f(Y_{X \leftarrow x} = y)$ for short. However, if we only know a CPDAG $\mathcal{G}^*$, the causal effects of X on Y may not be identifiable from observational data (Perković et al. 2015; Wu et al. 2019).

Let S, Y and X denote the sensitive attributes, outcome attribute, and other attributes, and $\hat{Y}$ be the predictions of a decision model produced by a learning algorithm. Based on do operator, we say that the decision model is counterfactual fairness with respect to the sensitive attribute S if it would have been the same had S been s in the counterfactual world as in the real world that S is s':

Definition 1 (Counterfactual Fairness)**.** *The model prediction $\hat{Y}$ satisfies counterfactual fairness with respect to the sensitive attributes S if under any context $\mathcal{Z} = \mathbf{z}$:*

$$P(\hat{Y}_{S \leftarrow s'} | S = s', \mathcal{Z} = \mathbf{z}) = P(\hat{Y}_{S \leftarrow s} | S = s', \mathcal{Z} = \mathbf{z}) \quad (1)$$

for all possible values of y, any value that S can take, and any context $\mathcal{Z} \subseteq \mathcal{X}$. The context $\mathcal{Z} \subseteq \mathcal{X}$ specifies certain sub-groups or individuals. Notably, when $\mathcal{Z} = \mathcal{X}$, counterfactual fairness is assessed at the individual level.

The Proposed Method

Overview

Given only observational data, the underlying causal DAG may not be recoverable without strong assumptions, such as linearity (Shimizu et al. 2006) or additive noise (Hoyer et al. 2008; Peters et al. 2014). Instead, we can use causal discovery algorithms (Zanga, Ozkirimli, and Stella 2022; Zheng et al. 2024) to obtain a CPDAG that contains the underlying causal DAG. Previous methods have exploited the CPDAG to achieve counterfactual fairness. They build on variable selection (Zuo et al. 2022), and suffer a reduced prediction accuracy and a strong assumption for identification.

In contrast with the work outlined above, we propose a novel and efficient method which uses all available variables to bound the counterfactual quantities of the sensitive attribute on the outcome, thereby ensuring counterfactual fairness. Our method does not require a search of all possible DAGs, but can estimate all possible counterfactual effects using CPDAG. In this section, we begin by presenting an identifiability criterion to bound the counterfactual effects under CPDAGs. We then provide an efficient method for implementing the proposed criterion by identifying the ancestral relations between two distinct nodes in a CPDAG. Finally, we propose a post-processing strategy to make the decision model satisfying counterfactual fairness, even when the counterfactual effect is not identifiable.

Counterfactual Inference and Identification

Our goal is to train a decision model $h(\mathbf{X}, S; \theta)$ that maps the observed variables to the outcome $\hat{Y}$ based on the parameters θ, so as to make accurate predictions while achieving counterfactual fairness. To formalize this objective, we first introduce the term augmented-$\mathcal{G}^*$ (Zuo et al. 2024) in CPDAGs as follows:

Definition 2 (Augmented-$\mathcal{G}^*$ with $\hat{Y}$)**.** *For a CPDAG $\mathcal{G}^* = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \mathcal{S} \cup \mathcal{X}$, let $\mathcal{G}'$ augment $\mathcal{G}^*$ by (i) adding an additional node $\hat{Y}$; and (ii) adding the edge $V \to \hat{Y}$ for each node $V \in \mathcal{V}$ in $\mathcal{G}^*$. The resulting graph is denoted by $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ where $\mathcal{V}' = \mathcal{V} \cup \{\hat{Y}\}$ and $\mathcal{E}' = \mathcal{E} \cup \{V \to \hat{Y} | V \in \mathcal{V}\}$. We call $\mathcal{G}'$ the augmented-$\mathcal{G}^*$ with $\hat{Y}$.*

Figure 1 (c) illustrates an example of augmented graph with $\hat{Y}$. For $\mathcal{G}'$, the joint probability distribution $P(\mathbf{v}, \hat{y})$ over $\mathcal{V} \cup \{\hat{Y}\}$, denoted as $P(\mathbf{v}, \hat{y}) = P(\hat{y}|\mathbf{v})P(\mathbf{v})$, is consistent with CPDAG $\mathcal{G}'$, as the following theorem (Theorem 4.1 in (Zuo et al. 2024)) states:

Theorem 1. *For a DAG $\mathcal{D}$, let $\mathcal{G}^*$ be a CPDAG such that $\mathcal{D} \in [\mathcal{G}^*]$. Let $\mathcal{D}'$ be the augmented-$\mathcal{D}$ with $\hat{Y}$ and $\mathcal{G}'$ be the augmented-$\mathcal{G}^*$ with $\hat{Y}$. Then the graph $\mathcal{G}'$ is a CPDAG with $\mathcal{E}' = \mathcal{E} \cup \{V \to \hat{Y} | V \in \mathcal{V}\}$ such that $\mathcal{D}' \in [\mathcal{G}']$.*

Theorem 1 implies that once we obtain the CPDAG $\mathcal{G}^*$ from the observational data $(\mathbf{X}, S)$, the augmented-$\mathcal{G}^*$ with $\hat{Y}$, denoted as $\mathcal{G}'$, is an exact CPDAG such that $\mathcal{D}' \in [\mathcal{G}']$. Therefore, we can directly model $\hat{Y}$ on any CPDAG $\mathcal{G}^*$.

With the augmented CPDAG $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$, where $\mathcal{V}' = \mathcal{X} \cup \mathcal{S} \cup \{\hat{Y}\}$, we next discuss the identifiability of counterfactual fairness in Eq. (1) on $\mathcal{G}'$, i.e., when and how we can uniquely estimate counterfactual quantities from observable distribution $P(\mathbf{v})$ over $\mathcal{G}^*$ and conditional distribution $P(\hat{y}|\mathbf{v})$ over $\mathcal{G}'$. Recall that counterfactual fairness concerns an individual from the disadvantaged group $S = s'$ who is specified by context $\mathbf{Z} = \mathbf{z}$. Obviously, the term $P(\hat{Y}_{S \leftarrow s'} = \hat{y} | S = s', \mathbf{Z} = \mathbf{z})$ in Eq. (1) is equivalent to

$P(\hat{Y} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z})$, as the intervention $do(S = s')$ does not change the value of S for the individual.

However, the term $P(\hat{Y}_{S\leftarrow s} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z})$ may suffer from the identifiable issue, as it involves two worlds: a real world where $\hat{Y} = \hat{y}$ and $S = s'$, and a counterfactual world where $\hat{Y} = \hat{y}$ and $S = s$. Given an augmented CPDAG $\mathcal{G}'$ and the sensitive attribute S, the ancestral relations of any other attribute with S can be divided into three types: the definite non-descendants, the definite descendants and the possible descendants of S. Without loss of generality, we denote with $de(S)$ the definite descendants of S, with $\mathrm{PosDes}(S)$ the possible descendants of S, and with $A^{(k)}(S)$ the any subset of $\mathrm{PosDes}(S)$. We define with A_S the descendant set of the sensitive attribute S. If there exists a DAG $\mathcal{D}$ in the equivalence class $\mathcal{G}'$ which includes a direct path from S to D, $D \in A_S$. We provide a graphical condition on $\mathcal{G}'$ and represent the identifiability formula as follows:

Proposition 1. *Given a CPDAG $\mathcal{G}^*$, let S be the sensitive node, and $\mathcal{G}'$ be the augmented-$\mathcal{G}^*$ with $\hat{Y}$. Then for any context $\mathcal{Z} \subseteq \mathcal{X}$, we have:*

1. *The counterfactual quantity $P(\hat{Y}_{S\leftarrow s} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z})$ is identifiable, if the intersection between the possible descendant set and the context is an empty set, i.e., $\mathcal{Z} \cap A_S = \emptyset$.*
2. *If $\mathcal{Z} \cap A_S = \emptyset$, the term $P(\hat{Y}_{S\leftarrow s} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z})$ can be uniquely computed from observed distribution $P(\mathbf{v})$ over $\mathcal{G}^*$ and conditional distribution $P(\hat{y}|\mathbf{v})$ over $\mathcal{G}'$, and the expression is as follows:*

$$P(\hat{Y}_{S\leftarrow s} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z}) = \frac{\sum_{\mathbf{x}\setminus\mathbf{z}}[P(s', \mathbf{x})P(\hat{y}|s, pa(\hat{y}))]}{P(s', \mathbf{z})} \quad (2)$$

3. *If $\mathcal{Z} \cap A_S \neq \emptyset$, the counterfactual quantity $P(\hat{Y}_{S\leftarrow s} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z})$ is unidentifiable. In this case, we can bound the counterfactual quantity as follows:*

$$P(\hat{Y}_{S\leftarrow s} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z}) \leq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}}\left[\max_{A^{(k)}(S)\in\mathbf{Z}\cap A_S}^{P(s',\mathbf{x})}\{P(\hat{y}|s, A^{(k)}(s), pa(\hat{y})\setminus A^{(k)}(s))\}\right]}{P(s', \mathbf{z})}$$

$$P(\hat{Y}_{S\leftarrow s} = \hat{y}|S = s', \mathbf{Z} = \mathbf{z}) \geq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}}\left[\min_{A^{(k)}(S)\in\mathbf{Z}\cap A_S}^{P(s',\mathbf{x})}\{P(\hat{y}|s, A^{(k)}(s), pa(\hat{y})\setminus A^{(k)}(s))\}\right]}{P(s', \mathbf{z})} \quad (3)$$

where specifically $\mathbf{Z} \cap A_S = \{A^{(k)}(S)|A^{(k)}(S) \subseteq de(S) \cup \{\mathcal{Z} \cap \mathrm{PosDes}(S)\}\}$ with intervention $do(S = s)$, and $k \in \{1, \cdots, K\}$ where $K = 2^{|\mathbf{z}\cap\mathrm{PosDes}(S)|}$.

The proof of Proposition 1 is given in Extended version. As shown in proof, the derived bounds derived reflect attainable worst and best-case causal effects; as such the width measures unavoidable causal uncertainty under unidentifiable scenarios in CPDAG. More importantly, Proposition 1 states that the estimation of counterfactual quantities requires only traversing the descendant sets of the sensitive attribute, avoiding full enumeration of DAGs in CPDAG.

Algorithm 1: Identify the type of the causal relationship between S and other covariates

Input: The training data $\mathcal{D} = \{(\mathcal{S}, \mathcal{X}, \mathcal{Y})\}$, local structure $\mathcal{G}^*$ of the sensitive attribute S.
Output: The set of definite descendants, definite non-descendants, and possible descendants of the sensitive attribute S

```
Initialize DefDes = ch(S, G*), DefNonDes = pa(S, G*), PosDes = ∅.
Let M be the set of maximal cliques of sib(S, G*).
for each variable X ∈ X\(pa(S, G*) ∪ ch(S, G*)) do
    if S ⫫ X|pa(S, G*) then
        Add X to DefNonDes
    else if S ⫫̸ X|pa(S, G*) ∪ sib(S, G*) or S ⫫̸ X|pa(S, G*) ∪ M for any M ∈ M then
        Add X to DefDes
    else
        Add X to PosDes
    end if
end for
return DefDes, DefNonDes, PosDes.
```

Finding Possible Descendant Sets

Given a CPDAG obtained from observational data, calculating all possible counterfactual effects in Eq. (1) by enumerating all DAGs is infeasible, since the number of possible DAGs can grow in a super-exponential manner as the number of vertices increases (He, Jia, and Yu 2015). In fact, from Proposition 1, we only need to find the possible descendant sets of the sensitive attribute to bound the unidentifiable counterfactual effects. Therefore, our objective becomes how to find all possible descendant sets of the sensitive attribute.

Recently, Zuo et al. (Zuo et al. 2022) proposed a sufficient and necessary graphical condition to judge causal relationships between any pair of vertices (S, Y) in a CPDAG by examining the critical set of S with respect to Y in $\mathcal{G}^*$. The critical set of S with respect to Y in $\mathcal{G}^*$ consists of all adjacent vertices of S lying on at least one chordless b-possibly causal path from S to Y (Fang and He 2020). However, the definition of critical set is based on the entire CPDAG. In order to check whether a vertice adjacent to S is in the critical set of S with respect to Y in $\mathcal{G}^*$, we need to list all chordless b-possibly causal paths from S to Y, which may also be time-consuming. In this paper, we propose to estimate possible counterfactual quantities by regressing each possible descendant set on the sensitive attribute. This provides a more efficient solution since enumerating possible descendant sets only requires the local structure around the sensitive attribute, with some additional conditional independence tests.

Given an augmented CPDAG $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ contains all DAGs equivalent to $\mathcal{G}^*$ and a sensitive attribute S, the lo-

cal structure around S can be divided into three cases: parents $pa(S, \mathcal{G}') \rightarrow S$, children $ch(S, \mathcal{G}') \leftarrow S$, and siblings $sib(S, \mathcal{G}') - S$ with undirected edges. We say $\mathcal{V}^* \subset \mathcal{V}'$ is a *clique* of graph $\mathcal{G}'$, if the induced subgraph of $\mathcal{G}'$ over $\mathcal{V}^*$ is complete. Furthermore, if there is no $\mathcal{V}'' \subset \mathcal{V}'$ such that $\mathcal{V}^* \subset \mathcal{V}''$ and $\mathcal{V}''$ is a clique, then $\mathcal{V}^*$ is denoted as a *maximal clique*. Then a sufficient and necessary condition for determining whether a variable $X \in \mathcal{X}$ is a definite descendant of the sensitive attribute S is shown as follows:

Theorem 2 ((Fang et al. 2022)). *Given a CPDAG $\mathcal{G}^* = (\mathcal{V}, \mathcal{E})$, let $\mathcal{M}$ be the set of maximal cliques of the induced subgraph of $\mathcal{G}^*$ over $sib(S, \mathcal{G}^*)$. Then for any $X \in \mathcal{V} \backslash \{S\}$, X is a definite descendant of S iff $S \not\perp\!\!\!\perp X | pa(S, \mathcal{G}^*) \cup sib(S, \mathcal{G}^*)$ or $S \not\perp\!\!\!\perp X | pa(S, \mathcal{G}^*) \cup \mathbf{M}$ for any $\mathbf{M} \in \mathcal{M}$.*

From above Theorem 2, we introduce Algorithm 1 to efficiently determine whether a variable is a descendant of sensitive attribute S using only local structure around S. Algorithm 1 only requires the local structure around S, specifically the tuple $(pa(S, \mathcal{G}^*), ch(S, \mathcal{G}^*), sib(S, \mathcal{G}^*))$, the skeleton of the induced subgraph of $\mathcal{G}^*$ over $sib(S, \mathcal{G}^*)$, and a limited number of d-separation queries, which can be answered by performing statistical independence tests. The local structure around S can be obtained using the variant of the MB-by-MB method proposed by Liu et al. (2020). Algorithm 1 outputs the set of definite descendants, definite non-descendants, and possible descendants of S.

The complexity of Algorithm 1 can be measured based on the maximum number of conditional independence tests, which is $(|Cl(sib(S, \mathcal{G}^*))| + 2) \cdot |\mathcal{V}(\mathcal{G}^*)|$, where $|Cl(sib(S, \mathcal{G}^*))|$ is the number of maximal cliques of $sib(S, \mathcal{G}^*)$, and $|\mathcal{X}|$ is the dimensionality of non-sensitive attributes. Since the number of maximal cliques scales linearly with respect to the number of nodes in a CPDAG, the complexity of Algorithm 1 is at most $\mathcal{O}(|sib(S, \mathcal{G}^*)| \cdot |\mathcal{V}(\mathcal{G}^*)|)$ (while the most related FairRelax (Zuo et al. 2022) is $\mathcal{O}(|sib(S, \mathcal{G}^*) + ch(S, \mathcal{G}^*)||\mathcal{V}(\mathcal{G}^*)||\mathcal{E}(\mathcal{G}^*)|)$ where $|\mathcal{V}(\mathcal{G}^*)|$ is the number of nodes and $|\mathcal{E}(\mathcal{G}^*)|$ is the number of edges in $\mathcal{G}^*$), which is more efficient than global enumeration and existing methods.

Counterfactually fair learning approach

We now return to the problem of learning counterfactually fair models. Directly imposing counterfactual fairness as a regularization term on prediction loss is not differentiable w.r.t. model parameter. As such, we propose a post-processing method by reconstructing the decision model with the worst-case violations of counterfactual fairness as a penalty term. Specifically, we first train a decision model $\hat{Y} = h(\mathbf{x}, s; \theta)$ with all available variables $(\mathbf{x}, s)$, which is parameterized by θ. Next, we construct a new decision variable $\tilde{Y}$ from $\hat{Y}$ such that counterfactual fairness with respect to $\tilde{Y}$ is satisfied. To this end, our goal is to find an optimal probabilistic mapping function $P(\tilde{y}|\hat{y}, \mathbf{x}, s)$ that minimizes the difference between Y and $\tilde{Y}$ with the worst-case violations of counterfactual fairness as a penalty term:

$$
\begin{aligned}
&\min \ell(\tilde{y}, y) = P(\tilde{y} = \hat{y})P(\hat{y} \neq y) + P(\tilde{y} \neq \hat{y})P(\hat{y} = y) \\
&s.t.\ U^{(k)}(P(\tilde{Y}_{S \leftarrow s} = \tilde{y}|s', \mathbf{z})) - P(\tilde{Y} = \tilde{y}|s', \mathbf{z}) \leq C, \\
&L^{(k)}(P(\tilde{Y}_{S \leftarrow s} = \tilde{y}|s', \mathbf{z})) - P(\tilde{Y} = \tilde{y}|s', \mathbf{z}) \geq -C, \\
&\sum\nolimits_{\tilde{y}} P(\tilde{y}|\hat{y}, \mathbf{x}, s) = 1, \quad 0 \leq P(\tilde{y}|\hat{y}, \mathbf{x}, s) \leq 1,
\end{aligned}
\tag{4}
$$

where $U^{(k)}(P(\tilde{Y}_{S \leftarrow s} = \tilde{y}|s', \mathbf{z}))$ and $L^{(k)}(P(\tilde{Y}_{S \leftarrow s} = \tilde{y}|s', \mathbf{z}))$ are the upper and lower bounds of counterfactual quantities derived in Proposition 1, $k = 1, \cdots, K$ and $K = 2^{|\mathbf{z} \cap \mathrm{PosDes}(S)|}$. The hyper-parameter C is introduced to penalize the loss when the counterfactual effect is larger than C, as achieving strict counterfactual fairness, i.e., having zero counterfactual effects of the sensitive attribute, is usually unrealistic and would come at the expense of prediction accuracy. Obviously, Eq. (4) is a convex optimization problem with $P(\tilde{y}|\hat{y}, \mathbf{x}, s)$ as parameters, since all probabilities except $P(\tilde{y}|\hat{y}, \mathbf{x}, s)$ can be obtained from the training set, and the distribution $P(\tilde{y}|\mathbf{x}, s)$ can be obtained by $P(\tilde{y}|\mathbf{x}, s) = \sum_{\hat{y}} P(\tilde{y}|\mathbf{x}, s)P(\tilde{y}|\hat{y}, \mathbf{x}, s)$. Thus, the objective is to find the optimal $P(\tilde{y}|\hat{y}, \mathbf{x}, s)$. The overall procedure of `CF-ICG` is summarized in Algorithm 2 in Extended version.

Experiments

In this section, we conduct experiments on synthetic and real-world datasets (as case study) to evaluate the accuracy and fairness of our approach. The common metric *Accuracy* is used to measure prediction performance. Counterfactual fairness can be measured by the discrepancy of predictions in the real world and the counterfactual world for each subgroup or individual. More results on other performance metrics, e.g., F1-score, are given in Extended version.

Experimental Setup

Compared methods. We compare `CF-ICG` against different types of competitive methods. (i) Baseline methods: *Baseline* trains a predictive model using all the variables, disregards the model fairness. *Oracle* uses all attributes that are non-descendants of the sensitive one, given the ground-truth DAG; (ii) Non-causal methods: *Unaware* (Grgic-Hlaca et al. 2016) uses all the attributes except the sensitive one; *LDF* (Fioretto et al. 2021) leverages Lagrangian duality to enforce fairness constraints (i.e. Demographic Parity (Feldman et al. 2015)) on the decision model. (iii) Causal methods: *FairRelax* (Zuo et al. 2022) uses all definite non-descendants and possible descendants of the sensitive attribute in a CPDAG; *Fair* (Zuo et al. 2024) uses all definite non-descendants of the sensitive attribute in a CPDAG; *IFair* (Zuo et al. 2024) uses all attributes and formulates a constrained fairness optimization problem; and LIFair (Li et al. 2024) estimates interventional effects by adjusting the possible parents of sensitive attribute, and then uses a min-max framework to achieve interventional fairness. Each compared method uses the same two hidden layers ReLU neural network with 64 hidden neurons as the base model.

Hyperparameter settings. For all the datasets, we split the training, validation, and test set as 70%, 10%, and 20%,

# of $\mathbf{z}$	Truth	CF−ICG	
		lower bound	upper bound
1	0.186	0.173	0.446
2	-0.055	-0.098	0.188
3	0.211	-0.023	0.232
4	0.146	-0.102	0.197

Table 1: Bounds and ground truth of counterfactual fairness on Synthetic dataset with 20 nodes & 40 edges.

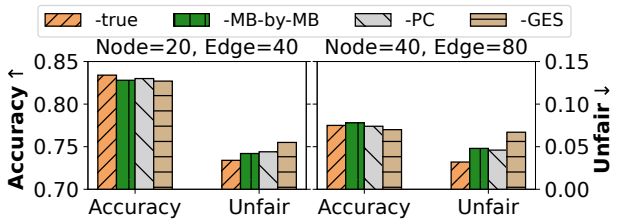


Figure 2: Accuracy (left side of each subgraph) and fairness (right side of each subgraph) of CF−ICG using different causal discovery algorithms on Synthetic datasets.

respectively. We report the average results of ten random splits. For each method we use grid search on the validation set to choose the best values of the hyper-parameters (More details about parameters can be found in Extended version). By default, we set the hyper-parameter $C = 0.05$ (which controls the trade-off between fairness and accuracy), and use the PC algorithm in the causal-learn package (Zheng et al. 2024) to learn a CPDAG with the significant threshold set to 0.01 for conditional independence testing.

Experiments on the Synthetic Datasets

Synthetic data generation. We first randomly generate DAGs with d nodes and $2d$ directed edges, where $d \in \{10, 20, 30, 40\}$ in our experiments. For each DAG, we randomly choose one node as the sensitive attribute and denote the last node in topological order as the outcome. We assume that the data generation mechanism follows a linear additive noise model, with the structural function for each node defined as $V_j = \sum_{V_i \in Pa(V_j)} \omega_{ji} V_i + \epsilon_j$, where ω_{ji} is randomly drawn from a Uniform distribution $U([-2, -0.5] \cup [0.5, 2])$, and $\epsilon \sim \mathcal{N}(0, 1)$, similar to (Zuo et al. 2022, 2024). We generate 2000 records where each node $V_j \in \mathcal{V}$ in a DAG is sampled from a Bernoulli distribution with probability $\sigma(V_j = \sum_{V_i \in Pa(V_j)} \omega_{ji} V_i + \epsilon_j)$, where $\sigma(\cdot)$ is the sigmoid function. For additional experiments based on more complicated structural equations (e.g., non-linear settings), please refer to Extended version.

Quantifying Counterfactual Fairness. We select two nodes, with sensitive attribute and outcome excluded, as the context $\mathcal{Z}$, which form unidentifiable situations. We estimate the bounds of counterfactual fairness using Proposition 1. The ground truth can be computed by exactly executing the intervention under given contexts using the complete causal model. More details about the evaluation of the ground truth can be seen in Extended version. The results are shown in Table 1, where the first column indicates the indices of $\mathbf{z}$'s value combinations. As can be seen, our upper and lower bounds cover the ground truth for all value combinations of contexts $\mathbf{z}$. This validates Proposition 1.

Performance Comparison. Similarly, we also select two nodes as the context set $\mathcal{Z}$ for each DAG to form the unidentifiable situations. Table 2 presents the performance in terms of accuracy (↑) and counterfactual fairness (↓) under different context $\mathbf{z}$. We can conclude that: (i) CF−ICG outperforms causality-based methods in terms of fairness and accuracy. This is because although Fair and FairRelax achieve counterfactual fairness, they only use definite non-descendants or possible non-descendants of the sensitive attribute, which loses information for the training model. LIFair fails to handle counterfactual fairness due to its inherent limitations. IFair can hardly find the true counterfactual quantities when dealing with unidentifiable situation. In contrast, CF−ICG makes predictions with all available variables and reduces unfairness by bounding counterfactual quantities. Moreover, CF−ICG uses only the local structure around sensitive attribute, and thus, it avoids potential errors associated with learning the entire CPDAG. (ii) CF−ICG outperforms statistical methods, i.e., Unaware and LDF, in terms of fairness, while managing comparable accuracy. This is because Unaware mitigates discrimination by excluding the sensitive attribute, but it still struggles to reduce unfair effects caused by descendants of the sensitive attribute. In addition, LDF cannot deal with the spurious correlation between sensitive attributes and predictions.

Robustness on Causal Discovery. To assess the robustness of our CF−ICG when combined with other causal discovery methods, we use a local method, *the variant of MB-by-MB* (Liu et al. 2020), to learn local structure around the sensitive attribute, and two global methods, PC (Colombo, Maathuis et al. 2014) and GES (Greedy Equivalence Search) (Chickering 2002), to learn CPDAGs. We also constructed a true CPDAG by randomly removing the edges from the DAG as a baseline. The results are presented in Figure 2. We observe that our CF−ICG is robust to different causal discovery algorithms. When CF−ICG is combined with different causal discovery algorithms, it exhibits similar performance in terms of both accuracy and fairness. The reason is that CF−ICG only requires the local causal structure of the sensitive attribute, thereby mitigating the estimation errors inherent in learning the entire CPDAGs.

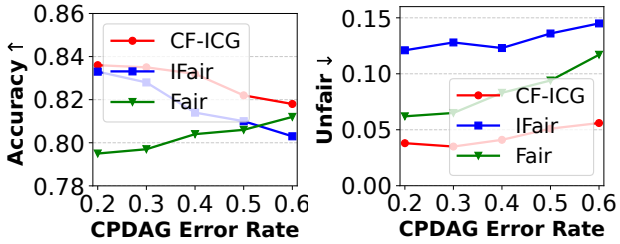
Robustness on noisy CPDAGs. To further evaluate the performance of our CF−ICG with respect to noisy CPDAGs, we investigate the impact of β-level error rate of CPDAGs on CF−ICG, where such noisy CPDAGs are created by randomly removing or reversing directed edges in the true DAG. The corresponding results are shown in Figure 3. As can be seen, as the error rate of CPDAG increases, all methods show a performance drop. Such a decline is much more apparent in the compared methods, as they rely on the entire CPDAG, and are restricted by its inherent estimation errors. In contrast, our CF−ICG alleviates the estimation error issue by focusing on the local structure around the sensitive attribute. Thus, its performance decline is much smaller as

	Nodes = 10, Edges = 20		Nodes = 20, Edges = 40		Nodes = 30, Edges = 60		Nodes = 40, Edges = 80	
	Accuracy↑	Unfairness↓	Accuracy↑	Unfairness↓	Accuracy↑	Unfairness↓	Accuracy↑	Unfairness↓
Baseline	0.842○	0.102●	0.857○	0.211●	0.847○	0.124●	0.816○	0.152●
Oracle	0.807●	0.000○	0.779●	0.000○	0.812●	0.000○	0.740●	0.000○
Unaware	0.839	0.093●	0.855○	0.170●	0.844	0.087●	0.812○	0.145●
LCD	0.833	0.075●	0.842	0.163●	0.835	0.084●	0.787	0.142●
Fair	0.808●	0.052	0.795●	0.057●	0.822●	0.073	0.748●	0.070●
FariRelax	0.820●	0.067●	0.806●	0.062●	0.828●	0.074	0.756●	0.103●
LIFair	0.835	0.082●	0.836	0.166●	0.839	0.080●	0.785	0.143●
IFair	0.832	0.069●	0.834	0.136●	0.836	0.077●	0.784	0.135●
CF−ICG	0.835	**0.053**	0.835	**0.044**	0.838	**0.068**	0.787	**0.059**

Table 2: Accuracy and fairness of compared methods on the synthetic datasets. The best results are highlighted in boldface. ○/● indicates that CF−ICG is statistically worse/better than compared method by student pairwise t-test at 95% confident level.

		Baseline	Oracle	Unaware	LCD	Fair	FairRelax	LIFair	IFair	CF−ICG
Adult	Acc. ↑	0.787○	0.685●	**0.780**○	0.773	0.704●	0.712●	0.770	0.772	0.753
	Unfair. ↓	0.193●	0.000	0.153●	0.149●	0.097●	0.116●	0.148●	0.143●	**0.043**
Compas	Acc. ↑	0.683○	0.614●	**0.680**○	0.667	0.631●	0.632●	0.665	0.664	0.652
	Unfair. ↓	0.232●	0.000	0.226●	0.141●	0.188●	0.203●	0.146●	0.137●	**0.048**
ACSPublicCoverage	Acc. ↑	**0.784**○	0.694●	0.782○	0.772○	0.719●	0.723●	0.769	0.766	0.748
	Unfair. ↓	0.247●	0.000	0.229●	0.157●	0.077●	0.104●	0.154●	0.135●	**0.037**

Table 3: Accuracy and Fairness violation score of each method. The best results are highlighted in boldface. ○/● indicates that CF−ICG is statistically worse/better than the compared method by student pairwise t-test at 95% confident level.



(a) CPDAG error vs. Acc. (b) CPDAG error vs. Unfair

Figure 3: Accuracy and fairness of CF−ICG using different noisy CPDAG on the Synthetic datasets with Node = 20 and Edge = 40.

the CPDAG error rate rises, demonstrating that our CF−ICG is more robust to noise in CPDAGs.

Case Study

To investigate the applicability of our CF−ICG in real-world scenarios, we further carry out experiments on three real-world datasets. Specifically, the Adult dataset (Dua and Graff 2017) contains 48,842 records with 7 variables, and the task is to predict whether the annual income of an individual exceeds $50,000. We take *sex* as the sensitive attribute. The ACSPublicCoverage dataset (Ding et al. 2021) is a *large-scale* dataset with 205,458 records and 12 variables. The task involves predicting whether an individual is covered by public health insurance, with *gender* considered as the sensitive attribute. The COMPAS dataset (Zafar et al. 2017) contains 6,172 records with 9 variables; the task is to predict whether a criminal defendant is likely to reoffend (recidivate). We treat *race* as the sensitive attribute.

For our case study on real-world datasets, we use the PC algorithm to learn a CPDAG using the causal-learn package. We remove all directed edges that point to the sensitive attribute, since *gender* and *race* are not affected by other collected features. Next, we randomly create a DAG as the ground-truth from the learned CPDAG. Counterfactual fairness is measured in almost the same way as in synthetic data. The details on counterfactual data generation can be referred to Extended version. We report the results in Table 3. Compared to other methods, CF−ICG exhibits a superior performance w.r.t. fairness, while achieving a higher or comparable accuracy, which further supports its capability to handle cases where counterfactual quantities are unidentifiable.

Conclusion

In this paper, we are the first to consider the problem of identifiability for counterfactual fairness under unknown or partially known DAGs, which makes a steady step toward more practical counterfactual fairness. We delve into an efficient and general framework to achieve counterfactual fairness on imperfect causal graphs, e.g., CPDAGs. An interesting finding is that we can bound the counterfactual quantities under unidentifiable cases by enumerating possible descendants of the sensitive attribute *only locally*, which is computationally more feasible than exhaustively listing all the DAGs. By analyzing the identification criteria and formulating a convex optimization problem, we achieve counterfactual fairness while maximizing data utility. Our method is applicable to a wide range of settings, as reliable CPDAGs are typically more accessible in practice than complete DAGs, and can effectively tackle unidentifiability issues under the CPDAGs.

Acknowledgments

This work is supported by NSFC (62031003, 62432006 and 62272276), Shandong Provincial Natural Science Foundation (No. ZR2024JQ001), Taishan Scholars Program (No. tsqn202306007 and tsqn202408317), the Youth Student Fundamental Research Funds of Shandong University (SDUQM2509).

References

Chiappa, S. 2019. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence*, 7801–7808.

Chickering, D. M. 2002. Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research*, 2(2): 445–498.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2): 153–163.

Colombo, D.; Maathuis, M. H.; et al. 2014. Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, 15(1): 3741–3782.

Ding, F.; Hardt, M.; Miller, J.; and Schmidt, L. 2021. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, 6478–6490.

Dua, D.; and Graff, C. 2017. UCI machine learning repository. http://archive.ics.uci.edu/ml.

Fang, Z.; and He, Y. 2020. Ida with background knowledge. In *Uncertainty in Artificial Intelligence*, 270–279.

Fang, Z.; Liu, Y.; Geng, Z.; Zhu, S.; and He, Y. 2022. A local method for identifying causal relations under Markov equivalence. *Artificial Intelligence*, 305: 103669.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 259–268.

Fioretto, F.; Van Hentenryck, P.; Mak, T. W.; Tran, C.; Baldo, F.; and Lombardi, M. 2021. Lagrangian duality for constrained deep learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 118–135.

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

Grgic-Hlaca, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NeurIPS Symposium on Machine Learning and the Law*.

He, Y.; Jia, J.; and Yu, B. 2015. Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 16(1): 2589–2609.

Hoyer, P. O.; Janzing, D.; Mooij, J.; Peters, J.; and Schölkopf, B. 2008. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, 689–696.

Huang, W.; Zhang, L.; and Wu, X. 2022. Achieving counterfactual fairness for causal bandit. In *AAAI Conference on Artificial Intelligence*, 6952–6959.

Jin, J.; Li, H.; and Feng, F. 2024. On the maximal local disparity of fairness-aware classifiers. In *International Conference on Machine Learning*, 22115–22144.

Jin, J.; Li, H.; Feng, F.; Ding, S.; Wu, P.; and He, X. 2023. Fairly recommending with social attributes: a flexible and controllable optimization approach. In *Advances in Neural Information Processing Systems*, 21454–21465.

Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, 4069–4079.

Li, H.; Liu, Y.; Geng, Z.; and Zhang, K. 2024. A local method for satisfying interventional fairness with partially known causal graphs. In *Advances in Neural Information Processing Systems*, 135415–135436.

Liu, Y.; Fang, Z.; He, Y.; Geng, Z.; and Liu, C. 2020. Local causal network learning for finding pairs of total and direct effects. *Journal of Machine Learning Research*, 21(148): 1–37.

Ma, J.; Guo, R.; Zhang, A.; and Li, J. 2023. Learning for counterfactual fairness from observational data. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1620–1630.

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

Pearl, J. 2014. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*, 19(2): 3.

Perković, E.; Textor, J.; Kalisch, M.; and Maathuis, M. H. 2015. A complete generalized adjustment criterion. In *Uncertainty in Artificial Intelligence*, 682–691.

Peters, J.; Mooij, J. M.; Janzing, D.; and Schölkopf, B. 2014. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(1): 2009–2053.

Rabonato, R. T.; and Berton, L. 2025. A systematic review of fairness in machine learning. *AI and Ethics*, 5(3): 1943–1954.

Robertson, J.; Hollmann, N.; Awad, N.; and Hutter, F. 2024. FairPFN: Transformers Can do Counterfactual Fairness. In *ICML 2024 Next Generation of AI Safety Workshop*.

Shimizu, S.; Hoyer, P. O.; Hyvärinen, A.; Kerminen, A.; and Jordan, M. 2006. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7: 2003–2030.

Su, C.; Yu, G.; Wang, J.; Guo, W.; Zheng, Y.; and Domeniconi, C. 2025. Multi-dimensional Causality Fairness Learning. *IEEE Transactions on Knowledge and Data Engineering*, 37(7): 4166–4178.

Su, C.; Yu, G.; Wang, J.; Yan, Z.; and Cui, L. 2022. A review of causality-based fairness machine learning. *Intelligence and Robotics*, 2(3): 244–274.

Wang, Z.; Chu, Z.; Blanco, R.; Chen, Z.; Chen, S.-C.; and Zhang, W. 2024. Advancing graph counterfactual fairness

through fair representation learning. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 40–58.

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual fairness: unidentification, bound and algorithm. In *International Joint Conference on Artificial Intelligence*, 1438–1444.

Wu, Y.; Zhang, L.; Wu, X.; and Tong, H. 2019. PC-fairness: a unified framework for measuring causality-based fairness. In *Advances in Neural Information Processing Systems*, 3404–3414.

Xu, D.; Wu, Y.; Yuan, S.; Zhang, L.; and Wu, X. 2019. Achieving causal fairness through generative adversarial networks. In *International Joint Conference on Artificial Intelligence*, 1452–1458.

Zafar, M. B.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *The Web Conference*, 1171–1180.

Zanga, A.; Ozkirimli, E.; and Stella, F. 2022. A survey on causal discovery: Theory and practice. *International Journal of Approximate Reasoning*, 151: 101–129.

Zheng, Y.; Huang, B.; Chen, W.; Ramsey, J.; Gong, M.; Cai, R.; Shimizu, S.; Spirtes, P.; and Zhang, K. 2024. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60): 1–8.

Zuo, A.; Li, Y.; Wei, S.; and Gong, M. 2024. Interventional fairness on partially known causal graphs: a constrained optimization approach. In *International Conference on Learning Representations*, 1–35.

Zuo, A.; Wei, S.; Liu, T.; Han, B.; Zhang, K.; and Gong, M. 2022. Counterfactual fairness with partially known causal graph. In *Advances in Neural Information Processing Systems*, 1238–1252.