

Fairness Detectability: Do Explanations Recover Fairness-Relevant Causal Effects?

Sakina Hansen¹, Kateryna Koval², Anna-Alina Bondarets², Joshua R. Loftus¹

¹London School of Economics and Political Science

²Ukrainian Catholic University
s.a.hansen1@lse.ac.uk

Abstract

Explainability (XAI) methods can play an essential role in fairness-based auditing of an algorithm. In such cases, an incorrect explanation could have detrimental consequences. In this paper, we begin by categorizing three goals of XAI for fairness: enabling correct epistemic footing, algorithm critique, and enabling change through intervention or policy. We take a pragmatic goal-oriented view where we believe XAI should be evaluated in terms of how well they achieve these goals. Framed by the three fairness goals identified, we 1) use causal data simulations with popular causal and non-causal XAI to highlight limitations through our knowledge of the underlying data generating process, 2) formalize these observations as impossibility propositions for non-causal XAI and 3) propose fairness detectability measures to visualize and quantify the performance of a particular XAI method with respect to a baseline causal effect. We find that non-causal XAI methods are less able to address the goals that fairness asks of explainability, with fairness detectability acting as a useful diagnostic tool to enable informed decision-making about XAI. The paper therefore contributes both empirical and methodological findings. Overall, our work contributes to research on pragmatic XAI as well as work aimed at developing the relationship between fairness and explainability.

1 Introduction

Fairness evaluations and audits may rely on explainable AI (XAI) methods that recover associations rather than causal effects (Zhao and Hastie 2021; Aïvodji et al. 2019). However, the goal of fairness audits might often be highly complex as they are involved in high-stakes decisions, such as in hiring, health care or university admission. These are complex socio-technical situations where audits ask specific questions; non-causal XAI¹ may systematically fail for fairness auditing tasks that require specific causal effects to be measured. Such causal relationships become particularly important as researchers have argued that XAI should go beyond transparency to instead emphasize socio-technical context (Smart and Kasirzadeh 2025; Sullivan and Kasirzadeh

2024), and move from descriptive to actionable explanations (Bhattacharya 2024).

The consequences of a flawed explanation in a fairness audit can be severe. Consider an external auditor using XAI tools to probe an algorithm that is, in truth, unfair. If the explanation tool’s limitations prevent the auditor from detecting the unfairness, they reach the wrong conclusion: no action is taken, no policy change is made, and an unjust situation persists. Understanding the behavior of an algorithm is essential for an audit to succeed, and how well XAI enables that understanding depends on what we ask it to do.

In this paper, we begin by motivating our approach with a taxonomy of fairness goals for XAI which can structure how we reason about the appropriateness of the use of an XAI method. We then formalize the limitations of non-causal XAI and demonstrate it empirically through simulation experiments and impossibility propositions, and propose fairness detectability as a diagnostic tool which measures the alignment between an explainer and a desired effect. We argue that this is a useful practical tool for adjudicating decisions about which XAI tool is appropriate for answering specific fairness questions, and improves transparency about assumptions involved when using XAI. The article proceeds as follows: in Section 2 we review the relevant literature and define the three fairness goals for XAI; in Section 3 we cover the necessary technical background; in Section 4 we interpret simulation experiments and formalize our observations through impossibility propositions; Section 5 introduces the fairness detectability methodology, validated on synthetic and real data experiments; Section 6 discusses the paper’s results in the context of our three fairness goals.² The paper makes the following contributions:

1. Empirically, the paper contributes simulation experiments that show that causal XAI methods have a stronger ability to identify fairness. We formalize these failures through impossibility propositions which apply causal results to XAI. We show that non-causal XAI is (1) blind to indirect unfairness when the predictor is invariant and (2) incapable of identifying a specific causal effect even when a non-zero relationship is visible.
2. Methodologically, this paper contributes fairness de-

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹We use “non-causal” to refer to methods which do not use a interventional reference distribution, instead using a marginal or conditional reference distribution.

²Supplementary material and code can be found at <https://github.com/sakinahansen/detectability>

tectability to quantify and visualize the discrepancy of an XAI method against a selected baseline effect, serving as a diagnostic tool for using XAI appropriately. Using heatmaps of FD scores, we show further evidence of how non-causal methods like PDP degrade with increasing mediation and confounding effects. Through experiments we show that FD can enable hypothesis driven decision-making about XAI choice that helps us to reason about which method is appropriate.

The contributions are sequential: the simulations show non-causal XAI failing where causal methods succeed, the propositions show these failures are structural, and fairness detectability turns that limitation into a practical diagnostic. To the best of our knowledge, this is the first paper to propose a methodology for evaluating XAI through alignment with the effect desired by a user. The paper’s contributions are both normative and technical in order to speak meaningfully across disciplinary boundaries at AIES; the paper sits within a tradition of XAI research at AIES that look at how XAI affects fairness outcomes (Wan, Belo, and Zejnilovic 2022; Sharma, Henderson, and Ghosh 2020), interdisciplinary approaches to XAI (Ayonrinde and Jaburi 2025), and the importance of factoring in causal effects (Mhasawade and Chunara 2021).

2 Background & Related Work

2.1 The Fairness-Explainability Relationship

Fairness and explainability are increasingly understood as intertwined and socio-technical rather than purely technical concerns. Deck et al. (Deck et al. 2024) argue that XAI for fairness must account for normative and societal context, while Skirzynski et al. (Skirzynski, Danks, and Ustun 2025) show that human experts cannot reliably use explanations to identify discriminatory predictions. Related work on socio-structural explanations (Smart and Kasirzadeh 2025), algorithmic remediation (West 2020), and the right to explanation (Vredenburg 2022) similarly frames explanation as a mechanism for understanding and addressing unfairness beyond the model itself. Our work complements this literature by experimentally showing where non-causal explanations fail to recover fairness-relevant effects and developing a method of fairness detectability for reasoning about these failures.

The relationship between fairness and explainability also depends on the type of unfairness: direct, where harm is attributable to a protected attribute, or indirect, where discrimination emerges through mediating or correlated factors (Sargeant and Magnusson 2025; Weerts et al. 2024; Poe and El Mestari 2024). As a result, the appropriateness of a given explanation method depends on the fairness question being asked: associational explanations may be sufficient for some forms of direct discrimination, while identifying indirect or systemic unfairness often requires explanations capable of recovering causal pathways and socio-technical relationships (Kim et al. 2023; Bhattacharya 2024).

We identify fairness goals for XAI:

1. *Placing a user or auditor on correct epistemic footing.* This mean being in a “good epistemic position to engage

in informed self-advocacy” (Vredenburg 2022) by having all the necessary information for a query that is not misleading or deceptive; the explanation has to match the specific goals and context of the question. For fairness, this will likely involve socio-structural explanations so that the entire socio-technical system can be understood, not just the algorithm (Smart and Kasirzadeh 2025).

2. *Critiquing the algorithm.* Once on the correct epistemic footing regarding an algorithm’s socio-technical behavior, an auditor may desire to make changes to or challenge the algorithm, with audit recommendations such as: it needs to be retrained, it creates unfairness through performativity (Zezulka and Genin 2023) so may require post-deployment testing, it should be taken down, etc.
3. *Intervening in the world.* Once on correct epistemic footing, a user or auditor may want to take action. This could mean wider societal policy changes such as “before deploying this hiring algorithm, state policy should ensure that women and men have equal access to university courses”. Action could also be on the instance level such as recourse, where an individual wants to change something to receive a better outcome, or if the algorithm is discriminatory the user could take legal action.

We want to know how well XAI methods can actually address these fairness goals. We hypothesize that causal vs non-causal XAI models address these goals to varying extent, where non-causal methods may fail. Our key assumption is that a meaningful and actionable explanation that lets you take action in the world should accurately summarize the important causal relationships relevant to a user’s query. As (Janzing, Minorics, and Blöbaum 2020) argues, feature relevance is fundamentally a causal question, a view supported by philosophical work on causal explanation (Woodward 2005). Indeed, the limitations of some marginalization methods such as the partial dependence plot have been examined (Zhao and Hastie 2021), but not in the context of the goals of XAI in a fairness context. This paper therefore takes a pragmatic view of XAI, viewing it as a tool for facilitation of these fairness goals. Where (Janzing, Minorics, and Blöbaum 2020) argues that XAI should be grounded in causality and (Skirzynski, Danks, and Ustun 2025) shows that XAI often fails to enable discrimination detection, this paper offers results that strengthen and merge these two findings, with a quantitative measure of detection through fairness detectability.

2.2 Causality, Fairness and Explainability

Humans naturally reason in causal terms, often leading to causal interpretations even when only correlational evidence is available (Loftus 2024). Causal explanations are valuable because they support intervention and manipulation (Woodward 2005), enabling actions that can reliably change outcomes. A substantial literature connects causality and algorithmic fairness (Mossé et al. 2025; Kilbertus et al. 2017; Kusner et al. 2017; Nabi and Shpitser 2018; Loftus et al. 2018; Zhang and Bareinboim 2018; Chiappa 2019), including work on causal explanation methods for fairness analysis (Heskes et al. 2020; König, Freiesleben, and Grosse-

Wentrup 2021; von Kügelgen et al. 2022; Karimi et al. 2022). Together, this literature suggests that understanding and mitigating discrimination often requires reasoning about the causal social processes underlying model predictions.

2.3 Evaluating XAI

As XAI has become more widely used, evaluating explanation quality has become increasingly important. Prior work has shown that explanations do not necessarily improve user understanding or decision-making (Bertrand, Eagan, and Maxwell 2023), while other work argues that explanations should be assessed relative to stakeholder needs, values, and pragmatic goals (Kasirzadeh 2021; Sullivan 2024). Existing evaluation research has examined fidelity (Lipton 2018; Arrieta et al. 2020), robustness to manipulation (Slack et al. 2020, 2021; Dombrowski et al. 2019; Alvarez-Melis and Jaakkola 2018; Hwang et al. 2025), and the limitations of associational explanations under causal dependence (Zhao and Hastie 2021; Plumb, Molitor, and Talwalkar 2018; Altmann et al. 2020; Molnar et al. 2020; Aas, Jullum, and Løland 2021). However, comparatively little work evaluates whether an explanation method actually recovers the *fairness-relevant effect of interest*. Our work addresses this gap by framing XAI evaluation around fairness goals and introducing fairness detectability as a pragmatic measure of how well an explanation aligns with a target causal effect.

3 Technical Background

3.1 Supervised Machine Learning

Supervised machine learning (ML) uses a dataset including predictor variables, denoted $X = (x_1, x_2, \dots, x_p)$, and an outcome variable y , and fits a function $\hat{f}$ to this dataset with the goal of accurately predicting $\hat{y} = \hat{f}(X)$. We use random forest (RF) and linear regression (LR) models.

3.2 Structural Causal Models

Fairness and Discrimination. We treat discrimination to be a causal effect of a protected attribute A on a prediction $\hat{Y} = \hat{f}(A, X)$ transmitted along pathways. Let the effect along a path set $\mathcal{P}$ as $\Delta_{\mathcal{P}}(\hat{f})$, then direct discrimination is the effect along $A \rightarrow \hat{Y}$ alone, indirect discrimination is the effect through mediators, and the total effect $\Delta_{TE\hat{f}}$ aggregates all paths. A *proxy* is a feature $x_i \neq A$ that has causal relationship to A and thereby carries influence into $\hat{f}$; proxy discrimination is indirect discrimination through an x_i and can occur even when A is not an input to $\hat{f}$. We deliberately do not commit to a single fairness criterion: which $\mathcal{P}$ is illegitimate is an audit-specific normative choice, and fairness detectability (Section 5) is defined relative to the auditor’s choice. Causal reasoning therefore provides a framework for distinguishing direct from indirect discrimination.

Our simulation scenarios and some of the explanation methods will make use of structural causal models (SCMs). SCMs represent variables and their causal relationships using directed acyclic graphs (DAG). Each variable is generated as a function of its parents and noise, and interventions can be modeled using do-operations that modify variable

values and propagate effects through the graph (Pearl 2009). SCMs allow estimation of causal effects relevant to fairness auditing.

3.3 Model-agnostic Explanation Methods

Model explanation methods useful for fairness audits are often model-agnostic feature attribution methods. To understand how $\hat{f}$ depends on feature x_j we can vary x_j and observe how $\hat{f}$ varies. Since $\hat{f}$ requires the other features x_{-j} as well, we must decide what values they should take when we vary x_j . *Different answers to this important and unavoidable choice lead to the variety of different explanation methods.* We focus on a selection of popular existing methods: partial dependence plots (PDP) (Friedman 2001), accumulated local effects plots (ALE) (Apley and Zhu 2020), causal dependence plots (CDP) (Loftus, Bynum, and Hansen 2024), Shapley values (SHAP) and asymmetric Shapley values (ASV). These tools have great potential for both deception (Slack et al. 2020) and for increasing transparency, broadening human understanding, and assessing fairness of opaque models (Ingram, Gursoy, and Kakadiaris 2022). Their simplicity comes with limitations regarding their correct interpretation, especially given that the choice of explanation can significantly alter perceptions of fairness (Van Berkel et al. 2021). We briefly describe the methods we use here; see the supplementary material for more details.

PDP (Friedman 2001) estimates the marginal relationship between a feature and model predictions by averaging over other features. PDP assumes independence between the feature of interest and remaining inputs, which can lead to misleading interpretations when features are causally related; **ALE** (Apley and Zhu 2020) estimates feature effects using conditional distributions, reducing unrealistic feature combinations compared to PDP. However, ALE remains observational and does not capture causal effects; **CDP** (Loftus, Bynum, and Hansen 2024) use a causal model to replace the assumption of independence between x_j and x_{-j} with an assumption of dependence. Varying x_j can now be modeled as an intervention that may also cause changes to x_{-j} . Having assumed or learned a SCM, we are able to estimate the total impact of a feature of interest on the outcome variable using the Total Dependence Plot (TDP), one of the methods from the family of CDPs; **SHAP** (Lundberg and Lee 2017) decomposes predictions into additive feature contributions using marginal reference distributions; **ASV** (Frye, Rowat, and Feige 2020) incorporates causal ordering constraints through asymmetric feature weighting, but still relies on conditional data distributions.

3.4 Simulation Scenarios

We generate synthetic data across prediction contexts — interview success (Figure 1), hiring decisions (Figure 2), and university admission (Figure 3) — with different causal features such as mediation and confounding. Figures 1-3 are causal DAGs representing the relationships between variables in the data-generating SCMs, not the fitted predictors. We sample data from each SCM and then train predictors $\hat{f}$ which may use only a subset of these variables. The SCM

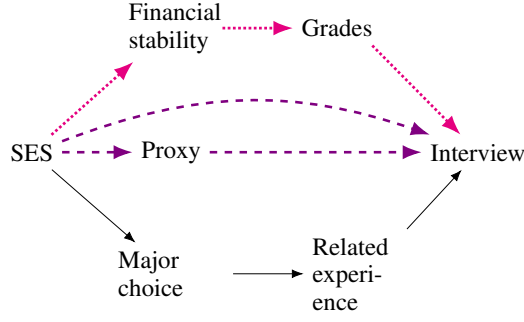


Figure 1: In this example model $\mathcal{M}_p$, SES has three causal pathways to influence the outcome of being interviewed.

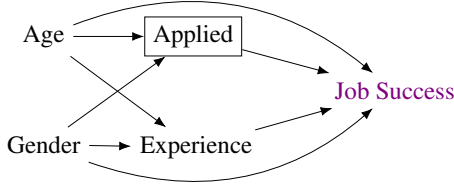


Figure 2: Causal model $\mathcal{M}_h$.

is used to generate data and to compute the causal baseline; the predictor is the object being explained. Since we know the true data generating process (DGP), we can compare the discrimination shown by an explanation method to the true discrimination. In each example we fit ML models and apply a range of XAI methods, allowing us to assess variability across these key factors.

Interview Simulation (Fairness Pathway Analysis). We fit an RF model $\hat{f}_{pRF}$ using data generated based on the SCM represented by $\mathcal{M}_p$ in Figure 1. “SES” represents a feature that acts as proxy for indicators of socioeconomic status. There are three different causal pathways that influence the outcome, which generate different versions of $\mathcal{M}_p$; one graph will have only the dashed violet arrows (the proxy pathway), one graph will have only the dotted magenta arrows (the financial pathway) and one graph will only have the solid black line (the experience pathway). One graph will have the entire model $\mathcal{M}_p$.

Hiring Simulation (Mediating Variable). Suppose the SCM $\mathcal{M}_h$ in Figure 2 generates training data for a hiring algorithm predicting applicant success using age and gender as sensitive attributes, with experience and applied acting as mediators, where applied is a selection variable. We fit predictive models with varying specifications from this SCM to illustrate different interpretive scenarios: 1) Model $\hat{f}_{hE}$ is a reduced model that includes only experience as a predictor, reflecting a designer choice to not include all information, for motivations such as “fairness through unawareness”; 2) Model $\hat{f}_{hEAG}$ is a full model that includes all variables.

University Admissions Simulation (Confounding Variable). Suppose the SCM $\mathcal{M}_u$ in Figure 3 generates train-

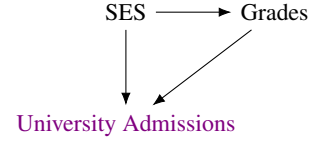


Figure 3: Causal model $\mathcal{M}_u$.

ing data for a university admissions algorithm predicting acceptance using grades and SES-related proxy features (e.g., school location, type, and demographics), where SES acts as a confounder of grades and admission due to institutional bias. We fit two models: 1) Model $\hat{f}_{uLR}$ is a LR model; 2) Model $\hat{f}_{uRF}$ is a RF regressor.

4 Limitations of Non-Causal XAI

4.1 Simulation Results

Our hypothesis when conducting these experiments was that there will be particular data scenarios where causal XAI methods are able to uncover discriminatory behavior where non-causal XAI methods fail to. For each simulation example and supervised learning model described in the previous section, we generate a set of XAI outputs: PDP, conditional PDP, TDP, ALE, SHAP and ASV plots. We assume in these examples that the auditor is interested in understanding both direct and indirect unfairness, so is interested in the total effect. This results section highlights three main takeaways we found from observing this range of plots: 1) causal XAI is better for fairness pathway analysis; 2) non-causal XAI methods may reach incorrect conclusions about fairness; 3) non-causal XAI may completely miss discrimination.

Causal XAI is better for fairness pathway analysis Figure 4 shows XAI results for data generated from $\mathcal{M}_p$ across the three causal pathways. Non-causal methods (PDP bottom left, SHAP bottom right) show no relationship between SES and the outcome, whereas TDP reveals the contribution of each pathway — enabling a hypothesis-driven auditor to test sub-graphs under uncertainty about the causal structure. An audit that uses these TDPs could produce better policy recommendations or insights into the fairness of using models in such scenarios. Without the ability to distinguish unfairness through different causal pathways, an auditor cannot be on correct epistemic footing (Goal 1) and lacks the granularity needed to recommend targeted policy interventions (Goal 3).

Non-causal XAI may reach incorrect conclusions about fairness For the university scenario, different XAI methods yield different conclusions about disparate treatment. Only TDP (Figure 5(c)) shows discrimination towards low SES, whereas PDP (a), ALE (d), SHAP (e), and ASV (f) do not, implying fundamentally different audit recommendations. Moreover, some methods show the opposite direction of discrimination. PDP (Figure 5(a,b)) and ALE (d) suggest that increasing SES decreases admission chances, while TDP (c) correctly shows that lower SES reduces admission chances, as confirmed by the DGP. Notably, ASV

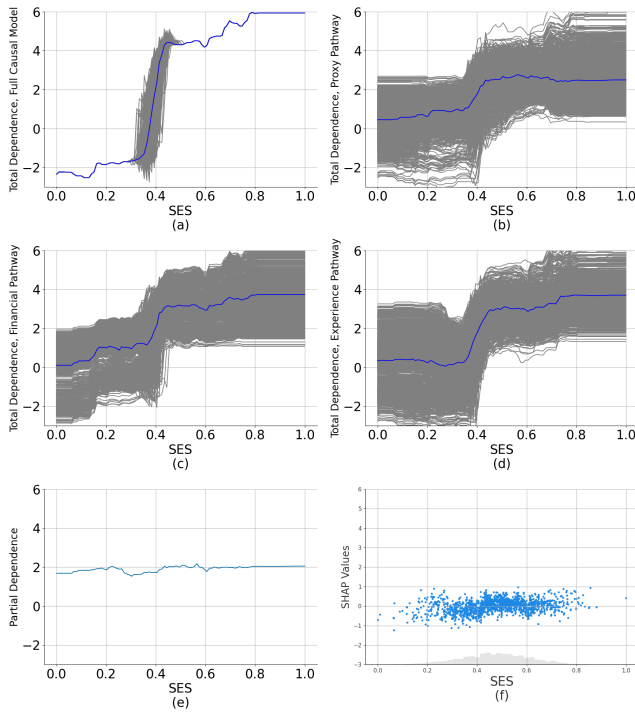


Figure 4: Total dependence plots for social category variable “SES” (socioeconomic status) for $\mathcal{M}_p$. From left to right, up to down, we have (a) TDP with the full causal model, (b) TDP with only the proxy pathway, (c) TDP with only the financial pathway, (d) TDP with only the experience pathway, (e) PDP, (f) SHAP summary plot. TDP is able to reveal the total effect through isolated pathways, where PDP and SHAP don’t show a relationship.

(f) does not improve upon SHAP (e) despite incorporating causal structure, because the causal graph informs feature weighting rather than the reference distribution, which remains conditional rather than interventional. The choice of XAI method could therefore produce opposite audit conclusions, with potential harm to already disadvantaged groups; this is consistent with work on infra-marginality producing wrong-sign conclusions about unfairness (Simoiu, Corbett-Davies, and Goel 2017). Opposite or misleading fairness conclusions undermine all three goals: the auditor is on incorrect epistemic footing (Goal 1), critique of the algorithm based on these results would be incorrect (Goal 2), and interventions derived from them could harm the groups they are intended to help (Goal 3).

Non-causal XAI may completely miss discrimination

Figure 6 shows the explanations of a full model $\hat{f}_{hEAG}$ and a reduced model $\hat{f}_{hE}$ which omits sensitive features as may occur through “fairness through unawareness”. For $\hat{f}_{hE}$, PDP (Figure 6(b)) and ALE (d) fail to detect indirect age discrimination, whereas TDP (f) reveals it and is unchanged between the full and reduced models (e,f). This highlights that marginal and conditional XAI method only reflect relationships present in the model inputs and are therefore vul-

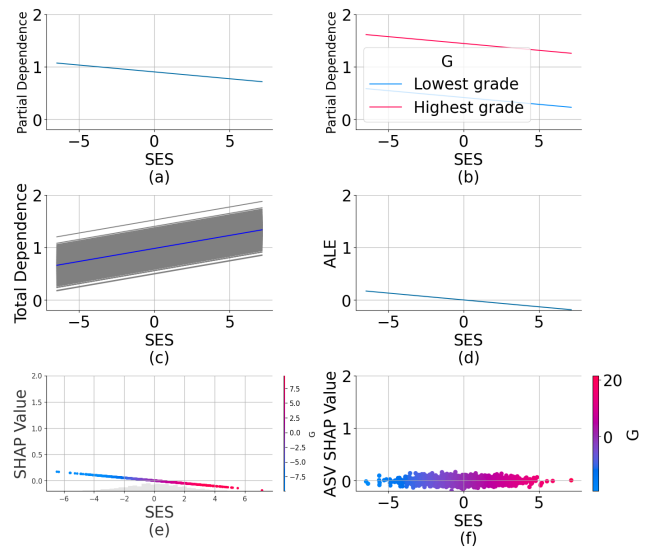


Figure 5: Explanations for variable ‘SES’ in our university scenario $\mathcal{M}_u$ for the model $\hat{f}_{uLR}$. (a) PDP (b) Conditional PDP on grades (c) TDP (d) ALE plot (e) SHAP and (f) ASV. TDP is the only method correctly showing discrimination against low SES.

nerable to **proxy discrimination** through mediators. Similarly, ASV does not improve over SHAP because it incorporates causal structure only in feature ordering, not in the data distribution. Even when trained with full information ($\hat{f}_{hEAG}$), PDP attenuates discrimination by averaging over correlated mediators. In contrast, TDP identifies the causal impact of age and gender even when these variables are excluded from the predictor, revealing discrimination mediated through omitted variables. When discrimination is entirely invisible to the explanation method, the auditor has no basis for critique (Goal 2) or intervention (Goal 3) — not because unfairness is absent, but because the tool has failed to surface it.

4.2 Impossibility Propositions

In the simulations, we observed that non-causal XAI sometimes showed no apparent dependence or displayed signals that did not correspond to the true total dependence. This raises a natural question: is this failure a feature of non-causal XAI methods? To answer this question we develop impossibility propositions, applying causal inference results to the XAI for fairness setting.

We distinguish two notions of effect. A causal effect (CE), defined in terms of interventions on a protected attribute in an SCM. An explainer-defined effect (EDE) is obtained by evaluating a trained predictor under feature manipulations according to a non-causal XAI method. These quantities are often incorrectly conflated.

Let the data set $A \in \mathcal{A}$ denote the protected attribute and assume binary $\mathcal{A} = \{0, 1\}$. Let $X \in \mathcal{X}$ be the other features, including mediators, proxies, confounders or any other variable with a causal relationship with A . Let $\hat{f}(A, X)$ be a

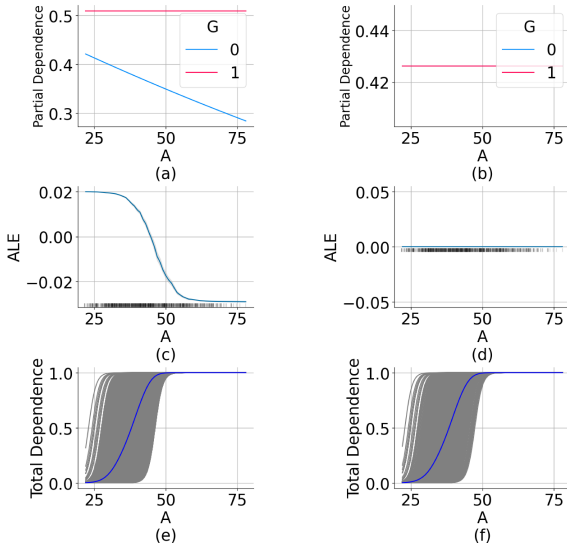


Figure 6: Explanations for the variable Age ‘A’ for the hiring scenario M_h , with some methods conditioning on gender, for the full model $\hat{f}_{hEAG}$ and reduced model $\hat{f}_{hE}$. The left column are the explanations for $\hat{f}_{hEAG}$, and the right column are the explanations for $\hat{f}_{hE}$. TDP is able to observe an effect even in the absence of the sensitive feature in the model.

predictor trained on the data. A general class of marginalization based explanation methods can be written as:

$$E_{\hat{f}}(a) := \mathbb{E}_{X \sim q(X)}[\hat{f}(a, X)] \quad (1)$$

where $q(X)$ is a typically a distribution over features derived from observational data used by the EDE. For a marginal distribution $q(X) = p(X)$ and for a conditional distribution $p(x|a)$. The EDE then reports an effect for each value of A . When these methods are used to assess the unfairness of the influence of feature A , this is done by contrasting different values of $E_{\hat{f}}(a)$. The EDE looking at a binary feature is:

$$\Delta_{E_{\hat{f}}} = E_{\hat{f}}(1) - E_{\hat{f}}(0) \quad (2)$$

On the other hand, in an SCM, the interventional mean prediction under $do(A = a)$ is:

$$\begin{aligned} \mu_{\hat{f}}(a) &= \mathbb{E}[\hat{f}(A, X)|do(A = a)] \\ &= \mathbb{E}_{X \sim p(X|do(A=a))}[\hat{f}(a, X)] \end{aligned} \quad (3)$$

with, for example, the causal total effect for a binary feature:

$$\Delta_{TE_{\hat{f}}} := \mu_{\hat{f}}(1) - \mu_{\hat{f}}(0) \quad (4)$$

Invariant Models

Proposition 1. Marginal blindness Assume $\hat{f}(1, x) = \hat{f}(0, x) \forall x$. For any marginal XAI with $q(X) = p(X)$,

$$\Delta_{E_{\hat{f}}} = \mathbb{E}_{q(X)}[\hat{f}(1, X)] - \mathbb{E}_{q(X)}[\hat{f}(0, X)] = 0 \quad (5)$$

This follows from pointwise invariance $\hat{f}(1, x) = \hat{f}(0, x) \forall x$ so $\mathbb{E}_{q(X)}[\hat{f}(1, X)] = \mathbb{E}_{q(X)}[\hat{f}(0, X)]$

For conditional explainers under invariance, $\Delta_{E_{\hat{f}}}$ is generally non-zero whenever $p(X|A = 1) \neq p(X|A = 0)$, but the resulting signal is not identifiable as any specific causal effect; this is a special case of Proposition 2. A model may be invariant because of $\hat{f}$ being trained without A for “fairness through unawareness”, if all predictive power of A flows through a mediator, or if the model is incorrectly specified. This effect was shown in the simulation results in Section 4 where we discussed how non-causal XAI may completely miss discrimination. While this result follows directly from the definition of invariance, its significance lies in formalizing a failure mode that is routinely overlooked in practice: models trained without the protected attribute, or where the protected attribute’s influence is fully mediated, will produce zero signal from a non-causal XAI.

Misleading Effects

Proposition 2. Spurious signal Let $\Delta_{E_{\hat{f}}} = \mathbb{E}_{X \sim q}[\hat{f}(1, X) - \hat{f}(0, X)]$ for some distribution q that doesn’t depend on A . The total causal effect of the model will be:

$$\begin{aligned} \Delta_{TE_{\hat{f}}} &= \mathbb{E}_{X \sim p(X|do(A=1))}[\hat{f}(1, X)] \\ &\quad - \mathbb{E}_{X \sim p(X|do(A=0))}[\hat{f}(0, X)] \end{aligned} \quad (6)$$

As long as $p(X|do(A = 1)) \neq p(X|do(A = 0))$, meaning A causally effects some component of X , then there exist models $\hat{f}$ where $\Delta_{E_{\hat{f}}} \neq \Delta_{TE_{\hat{f}}}$. Equality for all $\hat{f}$ is only possible if there is genuinely no effect, i.e. $p(X|do(A = 1)) = p(X|do(A = 0)) = q(X)$, or if $\hat{f}$ ignores the distributional changes in X under $do(A)$. Therefore, even when there is an EDE, it can’t be equated to the CE. For a conditional distribution, causal validity cannot be assumed unless you assume no unobserved confounding or causal sufficiency.

Proposition 3. Non-identifiability of causal unfairness from marginal or conditional distribution

Even if $E_{\hat{f}}$ shows an effect, for example if $\hat{f}$ is not invariant to A , the effect can’t be described as causal because the effect is *non-identifiable*. If we consider a trained predictor $\hat{f}$ and the reference distribution derived from the observational data q . There exists two distinct SCMs, $\mathcal{M}_1$ and $\mathcal{M}_2$ such that:

1. They induce the same observational distribution $p(A, X)$, and so the same marginal or conditional distribution
2. The trained predictor $\hat{f}$ and $E_{\hat{f}}$ are the same under both SCMs
3. But the actual causal effect differs: $\Delta_{TE_{\hat{f}}}^{\mathcal{M}_1} \neq \Delta_{TE_{\hat{f}}}^{\mathcal{M}_2}$

Proposition 3 applies the well-established result that distinct causal structures can be observationally equivalent (Pearl et al. 2000; Peters, Janzing, and Schölkopf 2017) to the explainability setting, showing that non-identifiability directly undermines the causal interpretation of EDEs. A non-zero EDE might reflect direct discrimination, proxy discrimination, or confounding — each warranting different audit recommendations — but the EDE provides no basis for

choosing between them. Formally, EDEs are neither necessary ($\Delta_{E_{\hat{f}}} = 0$ despite indirect unfairness, per Proposition 1) nor sufficient ($\Delta_{E_{\hat{f}}} \neq 0$ does not imply any particular type of unfairness). Non-causal XAI fail silently when the model is invariant, and fail misleadingly when it is not. In summary, for a broad class of non-causal explanations, these methods are (i) blind to indirect unfairness when the predictor is invariant to the protected attribute and (ii) incapable of identifying a causal effect even when a non-zero relationship is detected. This significantly limits their ability to satisfy the fairness goals outlined in Section 2.

5 Fairness Detectability for Explainability

Knowing that non-causal XAI fails in principle does not tell a practitioner whether their method, applied to their data, diverges from the relevant causal effect enough to undermine the audit. This motivates fairness detectability (FD): a context-specific diagnostic that quantifies the alignment between an explainer-defined effect and a chosen causal baseline, revealing how badly and where a given method fails.³ Since correct epistemic footing is a precondition for both critiquing an algorithm and intervening on the world, this diagnostic has implications for all three fairness goals identified in Section 2.

5.1 Fairness Detectability Score

Motivated by the limits we have identified for non-causal XAI, in this section we develop FD to measure the alignment between an XAI method and a chosen baseline effect. We can do this through the following steps:

Selecting a causal effect target The first step is to define one fairness-relevant CE which for example could be the total causal effect: $\Delta_{TE_{\hat{f}}} = \mathbb{E}[\hat{f}|do(A = 1)] - \mathbb{E}[\hat{f}|do(A = 0)]$. Depending on the problem and audit context, other causal effects may be of interest, such as partially controlled dependence or natural indirect dependence (Loftus, Bynum, and Hansen 2024). The causal baseline doesn't require a fully specified SCM, just a hypothesis about the causal relationships.

Defining the explainer effect EDE function Most explainability methods produce a function of sensitive feature A (such as PDP), per-instance attributions (such as SHAP) or single importance scores (such as permutation importance). We can define $\Delta_{E_{\hat{f}}}$ as the EDE. Using PDP as an example, $\Delta_{E_{\hat{f}}} = \Delta_{PDP} = \mathbb{E}_{X \sim q}[\hat{f}(1, X)] - \mathbb{E}_{X \sim q}[\hat{f}(0, X)]$.

Defining the fairness detectability metric We define FD, a method-specific evaluation measure, as a relative alignment ratio comparing an EDE to a specified CE baseline. FD is unbounded and may be negative (a guide for interpretation is given in Table 1). The CE is **any desired effect that the auditor would ideally like to measure**.

$$FD = \frac{\Delta_{E_{\hat{f}}}}{\Delta_{CE}} \quad (7)$$

³Note it is not a statistical test for fairness, but a diagnostic tool to aid decision making.

FD measures whether an EDE aligns with the target CE. FD scores can be calculated for particular values of the target variable if two values are of particular interest for a question, if A is binary, or to compare all values of $A = a_0, a_1, a_2, \dots, a_i$ for a continuous A .

$$FD_i = \frac{E_{\hat{f}}(a_i) - E_{\hat{f}}(a_0)}{CE(a_i) - CE(a_0)} \quad (8)$$

This calculates fairness detectability scores for all values of A , and a plot of FD against A can visualize a) if detectability changes for different values of A and b) how poorly it performs compared to a baseline. Computing FD requires that the baseline CE and $E_{\hat{f}}$ are on the same scale and evaluated for the same amount of values. Computing FD across values of A requires selecting a reference point a_0 from which all deltas are measured. This choice should be guided by the audit question: for instance, the population mean if the question concerns deviation from typical individual, or the youngest age if the question concerns whether older applicants are treated differently from the youngest cohort. FD is inherently less stable near the reference value.

In practice, an auditor asking whether an algorithm discriminates is implicitly asking a causal question, but methods like PDP answer an associational one. FD makes this gap visible: even without a fully specified SCM, computing FD under a plausible causal hypothesis reveals whether the explainer is tracking the quantity of interest.

Interpreting Fairness Detectability Fairness detectability measures relative alignment with a chosen causal effect that may be relevant to various fairness questions, rather than a bounded performance metric. It shouldn't be confused with the existence of unfairness itself or the unseen effect (i.e. the difference between the EDE and CE). The unseen effect is the absolute difference which can be relatively small, which would make two curves look similar, but could still be misaligned with the causal effect which is what FD measures. In particular if the causal effect is small, absolute differences appear negligible even when the EDE fails to capture the CE.⁴ Table 1 below explains the different interpretations of FD scores. FD may exceed 1 because non-causal explanation methods aggregate correlated downstream variation rather than isolating a single causal pathway, consistent with our impossibility results.

5.2 Detectability Curve Examples

In this section we generate several experiments to validate fairness detectability: 1) a synthetic experiment with a total effect baseline; 2) a synthetic experiment with a natural indirect effect baseline; 3) a synthetic grid experiment which varies mediation and confounding strength; 4) a real data

⁴When CE is near zero, FD is undefined as a ratio; this reflects the absence of a causal disparity rather than a failure of detectability. We define FD only when $|\Delta_{CE}| > \epsilon$ and mark such points as undefined in our plots. In practice, ϵ should be set small enough to avoid excluding regions where a genuine causal disparity exists, but large enough to avoid computing FD where the baseline effect is so close to zero that the FD becomes undefined.

SCORE	INTERPRETATION	ACCEPTABLE?
$FD < 0$	Opposite unfairness	×
$FD \approx 0$	Blind to unfairness	×
$0 < FD < 1$	Attenuation	×
$FD \approx 1$	Exact alignment	✓
$FD > 1$	Over attribution	×

Table 1: Interpreting FD scores with respect to the baseline effect. These cases reflect alignment properties of the EDE.

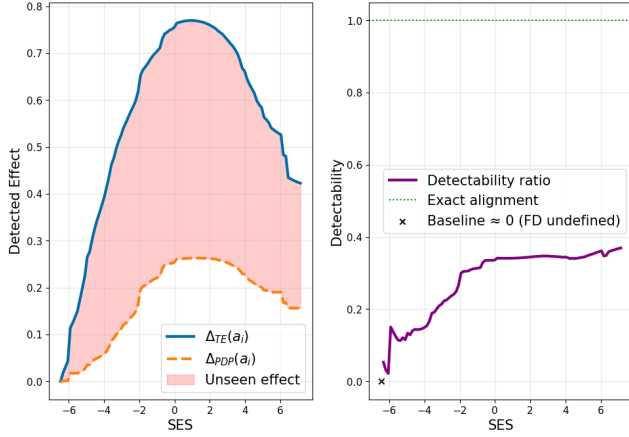


Figure 7: Unseen effect and detectability plots. On the left panel we observe Δ_{TE} as blue line and Δ_{PDP} as an orange line for $\hat{f}_{uRF}$, with a shaded orange region to show the unseen effect when using PDP with total effect as the baseline (the “unseen effect” can be interpreted as $\Delta_{TE} - \Delta_{PDP}$). The right shows the plotted FD scores for all values of SES, with a green dashed line to show what would be exact alignment with the baseline. X marks points where FD is undefined as the baseline is 0. We can see the detectability score is low for PDP, but also improves for higher SES.

experiment on the German Credit dataset (Hofmann 1994), where we train $\hat{f}_{GCLR}$ and $\hat{f}_{GCRF}$ to classify credit risk, with two DAGs from the literature to fit to the observed data to compute Δ_{TE} ; confidence intervals come from bootstrap resamples. We report FD curves alongside unseen effect curves. The unseen effect shows the absolute gap between the explainer and the causal baseline, while FD shows relative alignment; both are useful because a small absolute gap can still reflect poor alignment when the causal effect itself is small.

Total Effect Baseline We compute FD plots with a total effect baseline for PDP for $\hat{f}_{uRF}$ in Figure 7, with an FD score for maximum and minimum SES of 0.39. From Figure 7 we observe on the right that the detectability of PDP improves as SES increases. This is an additional finding that FD provides; PDP actually performs better for those with higher SES and so if used in an audit would hide the discrimination against lower SES specifically.

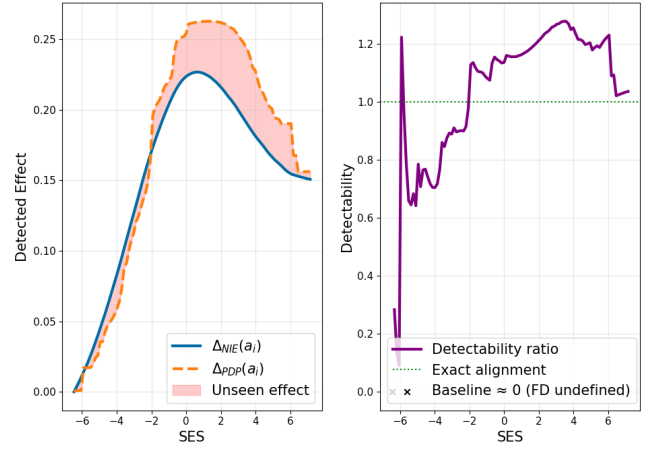


Figure 8: Unseen effect and detectability plots. On the left panel we observe Δ_{NIE} as blue line and Δ_{PDP} as an orange line for $\hat{f}_{uRF}$, with a shaded orange region to show the unseen effect when using PDP with natural indirect effect as the baseline (the “unseen effect” can be interpreted as $\Delta_{NIE} - \Delta_{PDP}$). The right shows the plotted detectability scores for all values of SES, with a green dashed line to show what would be exact alignment with the baseline. X marks points where FD is undefined as the baseline is 0. For some values of SES PDP performs well (close to 1 FD), whereas there is over and under-attribution for other values.

Natural Indirect Effect Baseline An auditor may be interested in specific path effects, in which case they may want to use a different CE baseline. For this example, we use the natural indirect effect (NIE) as a baseline. For the university simulation, this would be how much effect only goes through the mediator “grades”. In line with (Lofus, Bynum, and Hansen 2024), NIE would be defined as $NIE = \hat{f} \left(P\mathcal{M}_X^{do(I)} | X=x; do(K) \right)$, where K is an additional intervention removing outgoing edges from any nodes intervened upon by intervention I. We compute an FD plot for PDP for $\hat{f}_{uRF}$ in Figure 8, with FD score comparing maximum and minimum SES of 1.04. In this example, we can interpret the FD scores to show that PDP would overestimate the effect at some points as the NIE is focused on a particular pathway. Using PDP would lead to over attribution of the desired causal effect at various values of SES.⁵ This behavior is consistent with our impossibility results; non-causal XAI can be misleading even when it shows an effect because the EDE isn’t constrained to isolate a specific causal effect.

Varying Confounding and Mediation Using FD we diagnose how an EDE will perform or degrade with respect to an effect by varying confounding or mediation in the data. To characterise when the PDP fails as a fairness audit tool, we construct a simulation grid varying mediation

⁵The rapid variation at very low SES reflects the choice of reference point: as $a_i \rightarrow a_0$ both $\Delta_{E_{\hat{f}}}$ and Δ_{CE} tend to zero where it is less stable as a ratio measure. Measuring uncertainty therefore becomes important which we explore in later experiments.

DAG	FD SCORE [95% BOOTSTRAP CI]
$\mathcal{G}_1$	0.76 [0.36, 1.55](LR), 0.73 [0.04, 1.74](RF)
$\mathcal{G}_2$	0.90 [0.56, 1.82](LR), 0.89 [0.23, 1.60] (RF)

Table 2: Scalar FD scores comparing max and minimum age for the German Credit dataset, with 95% bootstrap confidence intervals.

strength α and confounding strength γ on the data scenario shown in Figure 9. For each cell of (α, γ) we compute $FD = \Delta_{PDP} / \Delta_{TE}$, where Δ_{TE} is the total causal effect of the protected attribute A on the outcome. Figure 10 reports the mean FD score for a LR and RF model trained on data generated from the simulation grid. The heatmap reveals that when α is small (weak or absent mediation), PDP achieves FD scores approaching 1, suggesting it is a reasonable proxy for the causal effect in these limited settings. However, as mediation strength increases, FD scores drop significantly: more of the total effect flows through pathways PDP cannot isolate, consistent with the findings in Section 4, making PDP increasingly unreliable for high mediation.

Fairness Detectability Under Causal Model Uncertainty

To assess robustness to causal model specification and how an auditor may use FD with competing causal hypotheses, we experiment on the German Credit dataset (Hofmann 1994), where applicants are classified as good or bad credit risks. We explore two DAG structures: $\mathcal{G}_1$ from Karimi et al. (Karimi et al. 2020) (Figure 11 (a)) and $\mathcal{G}_2$ from Chiappa (Chiappa 2019) (Figure 11 (b)). These encode fundamentally different fairness hypotheses: $\mathcal{G}_1$ assumes age and gender causally shape education, income, and loan amount, so effects on credit risk are heavily mediated; $\mathcal{G}_2$ has no mediation so the effect of age is more direct. In Table 2, we report scalar FD scores comparing max and minimum age for an LR model $\hat{f}_{GCLR}$ and an RF model $\hat{f}_{GCRF}$ with bootstrap confidence intervals (CI) which involves repeatedly re-sampling observations with replacement from the original dataset to estimate sampling variability in FD. Under $\mathcal{G}_2$, PDP achieves reasonably good alignment with TE. Under $\mathcal{G}_1$, PDP no longer aligns well with the total effect because PDP is unable to account for high mediation, supporting the findings from the heatmap.

Detectability in Audit Practice FD does not require certainty about the causal model, only a hypothesis. An auditor can sketch a plausible DAG based on domain knowledge and use it to inform causal baseline. FD then answers a conditional question: *if the world is structured this way, can my explanation method detect the resulting unfairness I’m interested in measuring?* Computing FD under several plausible DAGs treats disagreement across specifications as a signal that causal assumptions matter. If FD is sufficiently close to 1, the non-causal tool may be adequate; if not, the auditor knows not to trust it. Figure 12 shows a feasible process for using FD in an audit pipeline. The choice of baseline should be driven by the audit question, for example, total effect for assessing equal treatment, or a path-specific effects for in-

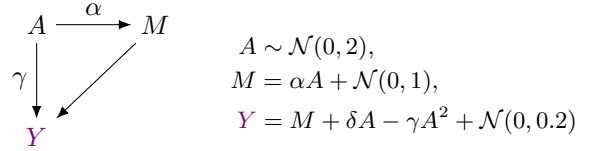


Figure 9: Causal model $\mathcal{M}_v$ with structural equations used to generate the simulation grid. Varying ‘ α ’ controls mediation, ‘ γ ’ controls confounding, and ‘ δ ’ toggles direct effect.

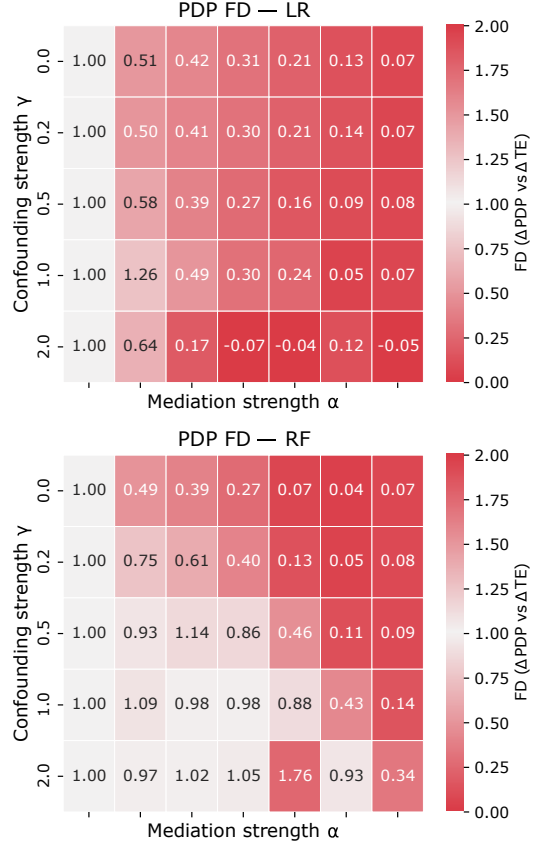


Figure 10: The heatmaps display mean FD scores for PDP against the TE, computed across data generated from $\mathcal{M}_v$ with varying mediation strength α (x -axis) and confounding strength γ (y -axis). A score of $FD = 1$ (white) indicates exact alignment between the PDP signal and the causal baseline; red cells indicate attenuation or mis-direction of the effect, with darker shading corresponding to greater divergence from Δ_{TE} . FD degrades with mediation strength.

terventions targeting a particular causal pathway. FD with different baselines can therefore determine appropriateness for different fairness goals, rather than assuming appropriateness a priori. FD does not certify fairness; it answers a conditional question based on the user’s causal assumptions.

6 Discussion

Reflection on fairness goals Our results have implications for all three fairness goals. For epistemic footing (Goal 1), audit frameworks — such as those emerging from the EU

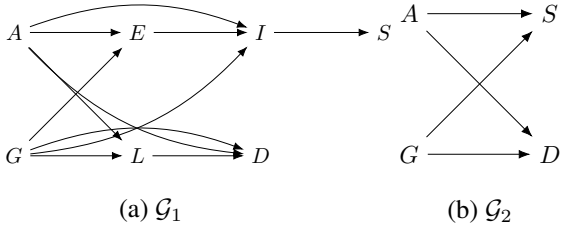


Figure 11: DAGs for the German Credit dataset from (Karimi et al. 2020) (a) and (Chiappa 2019) (b). A is age, G is gender, E is education, L is loan amount, D is loan duration, I is employment tenure, S is savings.

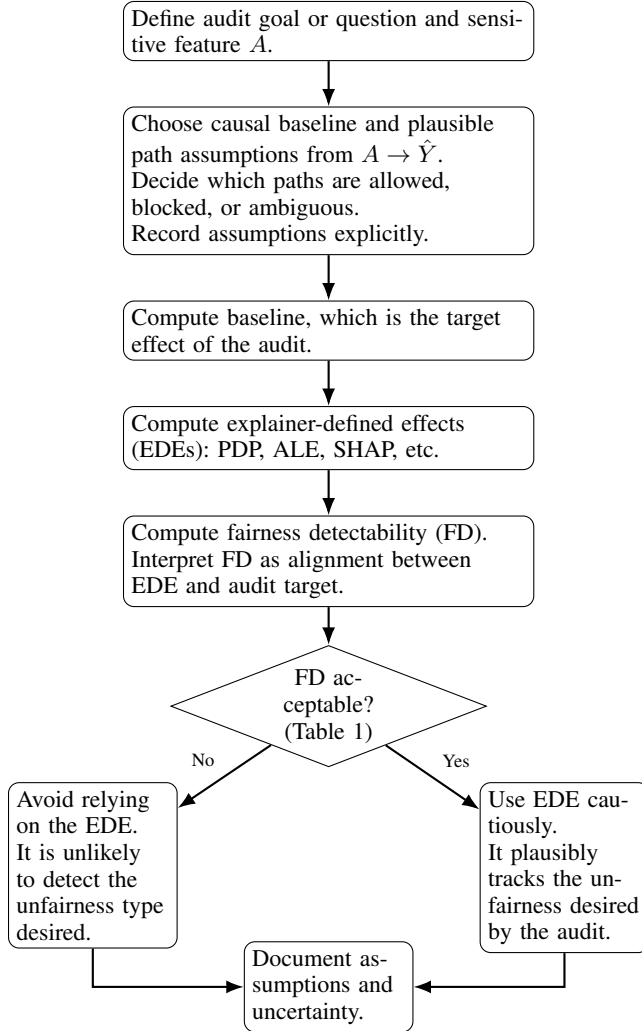


Figure 12: Operational audit flow for using fairness detectability scores in an audit.

AI Act — should not treat XAI methods as interchangeable; specifying which method was used and what causal assumptions it encodes should be a reporting requirement, and FD can support audit transparency and contestability (Alfrink et al. 2023). For algorithm critique (Goal 2), the choice of

XAI method constrains what critiques are possible: an auditor using PDP alone cannot make claims about proxy discrimination or mediated unfairness. Audit reports should be explicit not just about what was found, but what the chosen method was capable of finding. For intervention (Goal 3), a policy recommendation based on a PDP that shows the wrong sign of discrimination could actively harm the group it intends to help. Causal XAI paired with FD offers a more defensible basis for intervention; if FD is close to 1, the non-causal tool may be adequate, but if not, the auditor knows the intervention risks resting on a misleading foundation.

Limitations A key limitation of causal modeling concerns identifying a single “true” causal model (Greenland and Mansournia 2015). However, FD supports a hypothesis-driven approach (Miller 2023): *if* the world is like this, how unfair is the model? Our simulations also use simplified causal models to represent complex social phenomena, such as binary gender, and we use a selected set of experimental results, though these adequately serve the purpose of illustrating the limitations of non-causal XAI methods. FD has some limitations; FD is not user-tested to understand if the diagnostic translates into better audit practices. This is left for future work. FD itself is a conditional diagnostic; if an auditor uses FD with plausible but wrong causal assumptions, $FD \approx 1$ does not mean the explainer gives correct fairness conclusions, only that it agrees with the assumed hypothesis. FD does not validate the hypothesis itself. FD therefore could be misused by picking a DAG that makes the XAI appear aligned; this is avoided by clearly documenting assumptions. FD is unstable close to the reference point and so the reference point should be selected thoughtfully. Lastly, our experiments focus on a representative set of XAI methods; empirical evaluation of functionally different methods, such as counterfactuals or gradient-based explanations, is left to future work.

7 Conclusion

This paper considered fairness as an objective for XAI, where we investigated how different XAI methods influence the conclusions of a fairness audit through fairness simulation experiments. Based on these experiments, we formalized impossibility propositions for non-causal XAI that shows that for a broad class of non-causal XAI methods, they are (1) blind to indirect unfairness when the predictor is invariant to the protected characteristic and (2) incapable of having a causal interpretation even when a non-zero relationship is identified. This motivated our development of fairness detectability, a methodological tool that can be used to understand the performance of an XAI method with respect to an effect that is of interest. This has implications for the development of XAI methods and procedures; they should be goal-oriented and matched to a desired effect.

References

Aas, K.; Jullum, M.; and Løland, A. 2021. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *Artificial Intelligence*, 298: 103502.

- Aïvodji, U.; Arai, H.; Fortineau, O.; Gambs, S.; Hara, S.; and Tapp, A. 2019. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, 161–170. PMLR.
- Alfrink, K.; Keller, I.; Kortuem, G.; and Doorn, N. 2023. Contestable AI by design: Towards a framework. *Minds and Machines*, 33(4): 613–639.
- Altmann, T.; Bodensteiner, J.; Dankers, C.; Dassen, T.; Fritz, N.; Gruber, S.; et al. 2020. Limitations of interpretable machine learning methods.
- Alvarez-Melis, D.; and Jaakkola, T. S. 2018. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*.
- Apley, D. W.; and Zhu, J. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4): 1059–1086.
- Arrieta, A. B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R.; et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82–115.
- Ayonrinde, K.; and Jaburi, L. 2025. A Mathematical Philosophy of Explanations in Mechanistic Interpretability. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 265–278.
- Bertrand, A.; Egan, J. R.; and Maxwell, W. 2023. Questioning the ability of feature-based explanations to empower non-experts in robo-advised financial decision-making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, 943–958. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.
- Bhattacharya, A. 2024. Towards directive explanations: Crafting explainable ai systems for actionable human-ai interactions. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–6.
- Chiappa, S. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7801–7808.
- Deck, L.; Schoeffer, J.; De-Arteaga, M.; and Kühl, N. 2024. A critical survey on fairness benefits of explainable AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1579–1595.
- Dombrowski, A.-K.; Alber, M.; Anders, C.; Ackermann, M.; Müller, K.-R.; and Kessel, P. 2019. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32.
- Friedman, J. H. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.
- Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33: 1229–1239.
- Greenland, S.; and Mansournia, M. A. 2015. Limitations of individual causal models, causal graphs, and ignorability assumptions, as illustrated by random confounding and design unfaithfulness. *European journal of epidemiology*, 30: 1101–1110.
- Heskes, T.; Sijben, E.; Bucur, I. G.; and Claassen, T. 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in neural information processing systems*, 33: 4778–4789.
- Hofmann, H. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5NC77>.
- Hwang, H.; Bell, A.; Fonseca, J.; Pliatsika, V.; Stoyanovich, J.; and Whang, S. E. 2025. SHAP-based Explanations are Sensitive to Feature Representation. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1588–1601.
- Ingram, E.; Gursoy, F.; and Kakadiaris, I. A. 2022. Accuracy, Fairness, and Interpretability of Machine Learning Criminal Recidivism Models. *arXiv preprint arXiv:2209.14237*.
- Janzing, D.; Minorics, L.; and Blöbaum, P. 2020. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on artificial intelligence and statistics*, 2907–2916. PMLR.
- Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2020. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In *Advances in Neural Information Processing Systems*, volume 33, 265–277. Curran Associates, Inc.
- Karimi, A.-H.; von Kügelgen, J.; Schölkopf, B.; and Valera, I. 2022. Towards Causal Algorithmic Recourse. In Holzinger, A.; Goebel, R.; Fong, R.; Moon, T.; Müller, K.-R.; and Samek, W., eds., *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, Lecture Notes in Computer Science, 139–166. Cham: Springer International Publishing. ISBN 978-3-031-04083-2.
- Kasirzadeh, A. 2021. Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 14–14.
- 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.
- Kim, S. S.; Watkins, E. A.; Russakovsky, O.; Fong, R.; and Monroy-Hernández, A. 2023. ” help me help the ai”: Understanding how explainability can support human-ai interaction. In *proceedings of the 2023 CHI conference on human factors in computing systems*, 1–17.
- Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual Fairness. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

- König, G.; Freiesleben, T.; and Grosse-Wentrup, M. 2021. A Causal Perspective on Meaningful and Robust Algorithmic Recourse. *arXiv:2107.07853 [cs, stat]*. ArXiv: 2107.07853.
- Lipton, Z. C. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3): 31–57.
- Loftus, J.; Bynum, L.; and Hansen, S. 2024. Causal dependence plots. *Advances in Neural Information Processing Systems*, 37: 112656–112683.
- Loftus, J. R. 2024. Position: The Causal Revolution Needs Scientific Pragmatism. In *Forty-first International Conference on Machine Learning*.
- Loftus, J. R.; Russell, C.; Kusner, M. J.; and Silva, R. 2018. Causal Reasoning for Algorithmic Fairness. *arXiv:1805.05859 [cs]*. ArXiv: 1805.05859.
- Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Mhasawade, V.; and Chunara, R. 2021. Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 784–794.
- Miller, T. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, 333–342. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701924.
- Molnar, C.; König, G.; Herbringer, J.; Freiesleben, T.; Dandl, S.; Scholbeck, C. A.; Casalicchio, G.; Grosse-Wentrup, M.; and Bischl, B. 2020. Pitfalls to avoid when interpreting machine learning models.
- Mossé, M.; Schechtman, K.; Eberhardt, F.; and Icard, T. 2025. Modeling discrimination with causal abstraction. *arXiv preprint arXiv:2501.08429*.
- Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
- Pearl, J. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none): 96 – 146.
- Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2).
- Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press.
- Plumb, G.; Molitor, D.; and Talwalkar, A. S. 2018. Model agnostic supervised local explanations. *Advances in neural information processing systems*, 31.
- Poe, R. L.; and El Mestari, S. Z. 2024. The conflict between algorithmic fairness and non-discrimination: An analysis of fair automated hiring. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1907–1916.
- Sargeant, H.; and Magnusson, M. 2025. Formalising Anti-Discrimination Law in Automated Decision Systems. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 181–194.
- Sharma, S.; Henderson, J.; and Ghosh, J. 2020. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 166–172.
- Simoiu, C.; Corbett-Davies, S.; and Goel, S. 2017. The problem of infra-marginality in outcome tests for discrimination.
- Skirzynski, J.; Danks, D.; and Ustun, B. 2025. Discrimination Exposed? On the Reliability of Explanations for Discrimination Detection. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 2554–2569.
- Slack, D.; Hilgard, A.; Lakkaraju, H.; and Singh, S. 2021. Counterfactual explanations can be manipulated. *Advances in neural information processing systems*, 34: 62–75.
- Slack, D.; Hilgard, S.; Jia, E.; Singh, S.; and Lakkaraju, H. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.
- Smart, A.; and Kasirzadeh, A. 2025. Beyond model interpretability: Socio-structural explanations in machine learning. *AI & SOCIETY*, 40(4): 2045–2053.
- Sullivan, E. 2024. SIDEs: Separating Idealization from Deceptive Explanations in xAI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1714–1724.
- Sullivan, E.; and Kasirzadeh, A. 2024. Explanation Hacking: The perils of algorithmic recourse.
- Van Berkel, N.; Goncalves, J.; Russo, D.; Hosio, S.; and Skov, M. B. 2021. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.
- von Kügelgen, J.; Karimi, A.-H.; Bhatt, U.; Valera, I.; Weller, A.; and Schölkopf, B. 2022. On the fairness of causal algorithmic recourse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9584–9594.
- Vredenburg, K. 2022. The right to explanation. *Journal of Political Philosophy*, 30(2): 209–229.
- Wan, C.; Belo, R.; and Zejnilovic, L. 2022. Explainability’s gain is optimality’s loss? how explanations bias decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 778–787.
- Weerts, H.; Kelly-Lyth, A.; Binns, R.; and Adams-Prassl, J. 2024. Unlawful Proxy Discrimination: A Framework for Challenging Inherently Discriminatory Algorithms. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1850–1860.
- West, S. M. 2020. Redistribution and recognition: A feminist critique of algorithmic fairness. *Catalyst: Feminism, Theory, Technoscience*, 6(2).
- Woodward, J. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.
- Zezulka, S.; and Genin, K. 2023. Performativity and prospective fairness. *arXiv preprint arXiv:2310.08349*.

Zhang, J.; and Bareinboim, E. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhao, Q.; and Hastie, T. 2021. Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, 39(1): 272–281.