A Causal Framework to Measure and Mitigate Non-binary Treatment Discrimination

Ayan Majumdar[*1,2], **Deborah D. Kanubala**[*2], **Kavya Gupta**[2], **Isabel Valera**[2]

[1]Max Planck Institute for Software Systems, Saarbrücken, Germany
[2]Saarland University, Saarbrücken, Germany
ayanm@mpi-sws.org, {kanubala, gupta, ivalera}@cs.uni-saarland.de

Abstract

Fairness studies of algorithmic decision-making systems often simplify complex decision processes, such as bail or lending decisions, into binary classification tasks (e.g., approve or not approve). However, these approaches overlook that such decisions are not inherently binary; they also involve non-binary treatment decisions (e.g., loan or bail terms) that can influence the downstream outcomes (e.g., loan repayment or reoffending). We argue that treatment decisions are integral to the decision-making process and, therefore, should be central to fairness analyses. Consequently, we propose a causal framework that extends and complements existing fairness notions by explicitly distinguishing between decision-subjects' covariates and the treatment decisions. Our framework leverages path-specific counterfactual reasoning to: (i) measure treatment disparity and its downstream effects in historical data; and (ii) mitigate the impact of past unfair treatment decisions when automating decision-making. We use our framework to empirically analyze four widely used loan approval datasets to reveal potential disparity in non-binary treatment decisions and their discriminatory impact on outcomes, highlighting the need to incorporate treatment decisions in fairness assessments. Finally, by intervening in treatment decisions, we show that our framework effectively mitigates treatment discrimination from historical loan approval data to ensure fair risk score estimation and (non-binary) decision-making processes that benefit all stakeholders.

Code — https://github.com/ayanmaj92/fair-nonbin-treat
Extended version — https://arxiv.org/abs/2503.22454

1 Introduction

Data-driven systems are increasingly used to automate decisions in domains such as finance, healthcare, and criminal justice (Almheiri 2023; Moscato, Picariello, and Sperlí 2021; Habehh and Gohel 2021; Dieterich, Mendoza, and Brennan 2016). These systems typically reduce complex decisions to binary classification tasks, for example, simply predicting for loan repayment or recidivism (Mhasawade, D'Amour, and Pfohl 2024). However, *real-world decisions are rarely as simple as a binary choice, they involve multiple, non-binary treatment decisions* which in turn affect outcomes. Fig 1 illustrates the data-driven decision-making pipeline, which involves both (a) treatment decisions, (b) binary decisions, and (c) their downstream effects.

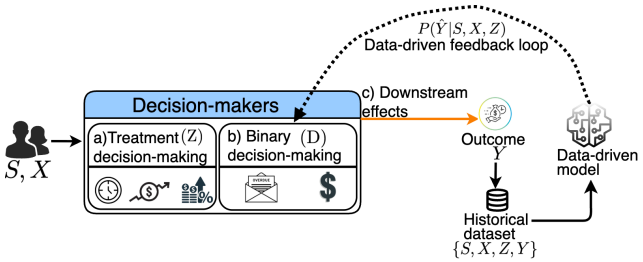

Figure 1: Illustration of data-driven decision-making pipelines. Parts (a)–(c) highlight the three common simplifications in existing fairness analyses.

Despite growing research on fairness in algorithmic decision-making, existing analyses simplify complex decision-making mechanisms in three key ways. First, many studies (Zafar et al. 2017; Hardt, Price, and Srebro 2016; Corbett-Davies et al. 2023) include non-binary treatments as covariates of the decision-subject (see covariate and treatment examples from different deployment contexts in Appendix B.1), restricting fairness analyses to binary decisions (Fig. 1b). This obscures the fact that such decisions are in control of the decision-makers. As a result, potential sources of discrimination may be masked, limiting the scope of fairness assessments. Second, some works (Madras et al. 2019; Coston et al. 2020; Plecko and Bareinboim 2023) account for treatment decisions (Fig. 1a) but assume they are binary, ignoring their often non-binary nature (e.g., loan or bail amounts). Finally, while a few studies do focus on disparities in non-binary treatment decisions, e.g., in lending (Agier and Szafarz 2013; Escalante et al. 2018; Alesina, Lotti, and Mistrulli 2013), they only provide empirical insights into disparities in treatment but overlook the downstream effects these treatment disparities may have on outcomes (Fig. 1c).

In this paper, we bridge the gaps in the fairness literature by extending fairness analyses to multiple, non-binary treatment decisions and their downstream outcome effects. We propose a causal framework that explicitly distinguishes between a decision-subject's covariates and the treatment decisions made by a decision-maker. Our framework enables us

[*]These authors contributed equally.

to: (i) measure disparities in non-binary treatment decisions induced by the sensitive attributes, and (ii) quantify how these disparities propagate to outcomes, thereby inducing *treatment discrimination*. Furthermore, we discuss how the proposed measures of treatment disparities and their downstream effects relate and complement existing fairness notions in binary decision-making.

To make our framework actionable, we introduce a practical approach for generating fairness-aware counterfactual data, leveraging recent advances in causal inference (Javaloy, Sánchez-Martín, and Valera 2024). Unlike prior methods, which often relied on restrictive assumptions or computationally intensive procedures (Nabi and Shpitser 2018; Chiappa 2019), our approach is broadly applicable and scalable. Moreover, it supports *treatment discrimination mitigation* via *pre-processing* techniques. These techniques generate *treatment-fair counterfactual datasets*, enabling fairer non-binary automated decisions and more equitable risk score estimation to inform future decision-making.

We validate our framework using four real-world lending datasets across two decision-making scenarios and empirically show that: (i) historical data encode treatment disparities that fair binary predictions fail to correct for owing to downstream effects of treatment decisions; and (ii) fairer treatment decisions yield more optimal outcomes, benefiting all stakeholders. Our findings thus highlight the need to address treatment decision disparities to achieve fairer outcomes, bridging a key gap in fairness research, and *laying the foundation for more holistic fairness assessments and mitigation efforts.* Thus, our analysis opens up avenues for future work, such as developing fair data-driven approaches for jointly making treatment and binary decisions.

2 Background

This section introduces the key causal concepts we leverage in our framework for studying treatment decision disparities.

Structural causal models (SCMs). An SCM (Pearl et al. 2000) encodes the *cause-effect relations* among observed features V as $\mathcal{M} = (V, U, F)$, where U is a set of mutually independent, unobserved exogenous variables, and F is a set of functions such that each feature $V_i \in V$ is given by $V_i = F_i(\mathrm{pa}(V_i), U_i)$, with $\mathrm{pa}(V_i) \subseteq V$ denoting the causal parents of V_i. The structure of these equations induces a directed acyclic graph (DAG) $\mathcal{G}$ over V, as seen in Fig. 2.

Performing interventions. Interventions are external actions that forcibly fix a feature's value, breaking its usual dependence on causal parents (Pearl et al. 2000). In SCMs, such *hypothetical interventions* are denoted by the do-operator. For instance, the intervention $\mathcal{I} = \mathrm{do}(V_i = \alpha)$ replaces V_i's structural equation with the constant assignment $V_i = \alpha$, producing a modified model $\mathcal{M}^{\mathcal{I}}$. This *intervened* SCM captures the changed causal structure and enables reasoning about the intervention's downstream effects.

Counterfactuals. In the absence of hidden confounders, the intervened SCM $\mathcal{M}^{\mathcal{I}}$ enables *counterfactual analysis* for a specific individual with observed features ϑ^F. This analysis answers the question: "What would the counterfactual features ϑ^{CF} be if we intervened with $\mathrm{do}(V_i \to \alpha)$, all else remaining the same?" Computing counterfactuals involves

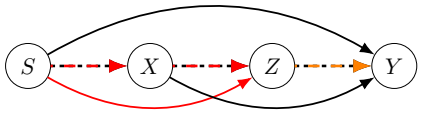

Figure 2: Causal graph of the data generation process, distinguishing decision-subject covariates X and treatment decisions Z, assuming positive binary decisions. Red paths (solid: direct, dotted: indirect) show the potential disparate influence of sensitive S on Z. The orange dashed path from Z to Y reflects possible discriminatory effects on outcomes.

three steps (Pearl et al. 2000): (i) *abduction* that infers the exogenous u^F from ϑ^F using the original $\mathcal{M}$; (ii) *action* that intervenes using $\mathrm{do}(V_i \to \alpha)$ to produce the intervened SCM $\mathcal{M}^{\mathcal{I}}$; and (iii) *prediction* that computes the counterfactual features ϑ^{CF} from $\mathcal{M}^{\mathcal{I}}$ using u^F and $V_i = \alpha$. Counterfactuals capture the *total effect* of an intervention on *downstream* causal features for the *individual*. This framework is widely used in fairness research (Kusner et al. 2017) to quantify the total effect of sensitive attributes (e.g., gender) on binary decisions (e.g., granting loans).

Path-specific counterfactuals. In certain scenarios, instead of focusing on total effects, we may want to understand how one feature influences another through *specific causal pathways*. This can be done using *path-specific counterfactuals*, which apply multiple do operations to isolate effects along selected paths. Consider a causal structure involving variables $\langle V_i, V_j, V_k \rangle$ with two paths: a direct path $V_i \to V_k$ and an indirect path $V_i \to V_j \to V_k$. For a factual instance $\vartheta^F = \langle v_i^F, v_j^F, v_k^F \rangle$, we can compute the counterfactual effect of $\mathrm{do}(V_i \to \alpha)$ along each path separately. To isolate the direct path $V_i \to V_k$, we block the indirect path by fixing V_j to its factual value, yielding $v_k^{CF}(\mathrm{do}(V_i \to \alpha), \mathrm{do}(V_j \to v_j^F))$. To isolate the indirect path $V_i \to V_j \to V_k$, we fix $V_i = v_i^F$ for the direct path, but update V_j based on the intervention using $v_k^{CF}(v_i^F, \mathrm{do}(V_j \to v_j^{CF}(\mathrm{do}(V_i \to \alpha))))$. This approach has been effective in fairness studies (Chiappa 2019) to determine whether sensitive features influence binary decisions *through potentially problematic causal paths*.

3 Measuring Treatment Disparities

In this section, we introduce our causal framework to measure non-binary treatment disparities and their downstream effects. Our framework considers historical data $\mathcal{D}$ generated by the causal graph $\mathcal{G}$ (Fig. 2) , where:

1. *Treatment decisions Z are assigned by decision-makers* (e.g., banks setting loan terms) and are causally influenced by decision-subject covariates X (e.g., income, savings) and possibly the sensitive attribute S (e.g., gender, race). This mechanism *may exhibit* disparities via *direct* (solid red) and *indirect* (dashed red) effects of S.
2. *Treatment decisions Z causally affect outcomes Y* (e.g., loan terms affect repayments). Thus, S may influence Y indirectly through Z (orange dashed line in Fig. 2), po-

tentially propagating treatment disparities to outcomes.

Remarks. Although Fig. 2 presents a single S for simplicity, our framework supports modeling multiple as independent root nodes. We treat X and Z as *multivariate blocks*, focusing on causal relations between blocks without making assumptions about causal structures within them. The directed edges in $\mathcal{G}$ indicate the *causal ordering* (Javaloy, Sánchez-Martín, and Valera 2024) among feature groups, although some of these relations may have zero effect. To enable general analyses, we retain $S \to Y$ to account for direct effects in scenarios like healthcare.

3.1 Disparities in Non-Binary Treatment Decisions

Based on the causal graph in Fig. 2, we use the SCM framework to measure how S influences treatment decisions Z, capturing both the total and the direct disparities. For simplicity and without loss of generality, we assume a binary sensitive attribute, $S \in \{0, 1\}$. [1] Given a dataset $\mathcal{D} = \{s^F, x^F, z^F, y^F\}$ consisting of instances that received a positive decision under a historical binary predictive policy, we define disparity measures using a disparity function Δ, which depends on the type of treatment variable, e.g., using value differences for continuous treatments.

Definition 1. *Total Treatment Disparity* (TTD) against decision-subjects with factual sensitive feature s^F, measures the expected disparity between the treatment z^F they received and the treatment z^{SCF} they would have received if they belonged to a different sensitive group, i.e.,

$$\begin{aligned} \mathrm{TTD}(s^F) &= \mathbb{E}_{\mathcal{D}}\left(\Delta(z^{\mathrm{SCF}}, z^F)\right), \text{ where} \\ z^{\mathrm{SCF}} &= z^{CF}(\mathrm{do}(S \to (1 - s^F))) \end{aligned} \tag{1}$$

Following Kusner et al. (2017), this notion captures the *total effect* of S on the treatment decisions Z. The *sensitive counterfactual* treatment z^{SCF} reflects, e.g., the loan terms a female applicant would receive if they were male with counterfactually adjusted covariates like income.

Definition 2. *Direct Treatment Disparity* (DTD), against decision-subjects with factual sensitive feature s^F, measures the expected disparity between the treatment they received and the sensitive *direct-path counterfactual treatment*, i.e.,

$$\begin{aligned} \mathrm{DTD}(s^F) &= \mathbb{E}_{\mathcal{D}}\left(\Delta(z^{\mathrm{SPCF}}, z^F)\right), \text{ where} \\ z^{\mathrm{SPCF}} &= z^{CF}(\mathrm{do}(S \to (1 - s^F)), \mathrm{do}(X \to x^F)) \end{aligned} \tag{2}$$

Following Nabi and Shpitser (2018); Chiappa (2019), this captures the *direct effect* of S on Z. The *sensitive direct-path counterfactual* z^{SPCF} reflects, e.g., the loan terms a female applicant would have received if they identified as male while maintaining their original covariates like income.

Relation to fair binary predictions. Ensuring fairness in binary decisions determines who receives a positive decision (e.g., loan granted) but does not directly address disparities in more complex, non-binary treatment assignments. We illustrate this empirically in Sec. 5. Still, drawing parallels with predictive fairness concepts can help interpret the implications of treatment disparities. For example, DTD captures disparities caused by the *explicit use* of sensitive attributes S in assigning Z. While this may be acceptable in settings like healthcare, it is generally inappropriate in domains like lending, echoing the notion of *disparate treatment* (Dwork et al. 2012). Similarly, zero TTD requires Z to be independent of S, analogous to *demographic parity* (Dwork et al. 2012), which requires binary predictions to be independent of S. As with predictive fairness, the appropriateness of treatment parity depends on context. When the true outcome Y is unobserved (e.g., hiring), treatment disparities may be considered discriminatory. In contrast, when Y is observable and the ground-truth (e.g., loan repayment), separation-based fairness frameworks (Hardt, Price, and Srebro 2016; Zafar et al. 2017) suggest that *outcome differences may justify some disparities*. Hence, analyzing the downstream impact on Y is essential for assessing the unfairness of treatment disparities.

3.2 Downstream Effects of Treatment Disparities

While Def. 1 and 2 help measure treatment disparities, it is crucial to assess the downstream impact of these disparities on the outcome Y (e.g., when Y is a ground truth as in lending) to determine if they are discriminatory. Next, we introduce two novel metrics to quantify the downstream outcome effects of total and direct treatment disparities.

Definition 3. *Total Treatment Disparity Effect* (TTD-E) measures the downstream effect of the sensitive counterfactual treatment z^{SCF}, computed in Eq. 1, as the probability of a *label change* from the factual value y^F, i.e.,

$$\mathrm{TTD\text{-}E}(s^F, y^F) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}\left(y^{CF}(\mathrm{do}(Z \to z^{\mathrm{SCF}})) \neq y^F\right)\right], \tag{3}$$

where $\mathbb{I}$ is an indicator function. This measures how the outcome would change if the decision-maker applied the *sensitive counterfactual* treatment z^{SCF}, e.g., how a female applicant's repayment would differ if her covariates stayed the same but treatment was adjusted by the *total effect of* S.

Definition 4. *Direct Treatment Disparity Effect* (DTD-E) measures the downstream effect of the *sensitive path* counterfactual treatment z^{SPCF}, computed in Eq. 2, as the probability of a *label change* from the factual value y^F, i.e.,

$$\mathrm{DTD\text{-}E}(s^F, y^F) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{I}\left(y^{CF}(\mathrm{do}(Z \to z^{\mathrm{SPCF}})) \neq y^F\right)\right]. \tag{4}$$

This measures how the outcome would change if the decision-maker applied the *sensitive direct-path counterfactual treatment* z^{SPCF}, e.g., how a female applicant's repayment would differ if her covariates stayed the same but only the treatment was adjusted by the direct effect of S.

Relation to fair binary predictions. Ensuring parity in treatment decisions *complements* parity in binary decisions, and fairness in predictions alone cannot eliminate unfair effects of treatment disparities on outcomes. However, connecting to predictive fairness concepts helps identify when treatment disparities lead to *discrimination through unfair*

[1] Discussions on possible extensions to multi-valued and multiple sensitive features in Appendix B.3.

decision outcomes. For example, under the notion of *disparate treatment* (Dwork et al. 2012), any direct use of sensitive attribute S in assigning treatment Z is problematic, and its impact on outcome Y is considered discriminatory. However, in some contexts like healthcare, DTD effects may be justified if they benefit Y. Similarly, disparities between counterfactual and factual treatments (Z^{SCF} and Z^F) can cause unfairness in Y. In predictive fairness, separation (Zafar et al. 2017) allows decision disparities if they do not harm ground-truth outcomes. Likewise, treatment disparities (e.g., TTD) may be deemed discriminatory if they fail to improve Y, such as when factual treatment z^F leads to worse outcomes than counterfactual treatment z^{SCF}.

3.3 Practical Implementation

Prior work (Nabi and Shpitser 2018; Chiappa 2019) on path-specific counterfactuals for predictive fairness was either limited to linear settings or relied on imprecise approximations and regularization. To allow for a *practical yet theoretically grounded* method for computing path-specific counterfactuals, we leverage *causal normalizing flows* (CNF) (Javaloy, Sánchez-Martín, and Valera 2024), which have shown accurate causal estimates in societal decision-making settings (Majumdar and Valera 2024).

Background on CNF. Given factual data $\mathcal{D} = \{s^F, x^F, z^F, y^F\}$ and causal graph $\mathcal{G}$, CNF approximates the unknown SCM $\mathcal{M}$ using a single invertible normalizing flow model T_ψ. CNFs support *partial causal graphs*, where some features are grouped into *multivariate blocks*, requiring only a *causal ordering* across blocks. CNFs can also flexibly learn from data if certain causal relations have *zero effect*. Once trained, CNF estimates counterfactuals for a given factual instance $\vartheta^F = \{v_1^F, \dots, v_K^F\}$ under an intervention $\text{do}(V_j \to \alpha)$ through three steps: (i) *Abduction*; infer exogenous variables as $u^F = T_\psi(\vartheta^F)$; (ii) *Action*, modify u^F to reflect the intervention, yielding $u^{CF} = \{u^F_{1:j-1}, T_\psi(\text{do}(V_j \to \alpha)), u^F_{j+1:K}\}$; and, (iii) *Prediction*, estimate counterfactual $\vartheta^{CF} = T_\psi^{-1}(u^{CF})$. See Appendix B.2 for extended details. Next, we show how to adapt CNF to estimate treatment disparities.

Estimating counterfactual treatments and outcomes. We can use a causal normalizing flows model trained using factual $\mathcal{D}$ to compute counterfactual quantities for some factual $\vartheta^F \in \mathcal{D}$. Importantly, we can readily utilize the *existing* counterfactual estimation method shown above to compute the *sensitive total treatment counterfactual* z^{SCF} as:

$$z^{\text{SCF}} = T_\psi^{-1}(\langle T_\psi(\text{do}(s \to (1 - s^F))), u_X^F, u_Z^F, u_Y^F\rangle) \quad (5)$$

We can also directly use the existing CNF approximation process to compute the *counterfactual labels* in Def. 3 and 4 resulting from a treatment intervention $\text{do}(Z \to \hat{z})$, as:

$$y^{CF}(\text{do}(Z \to \hat{z})) = T_\psi^{-1}(\langle u_S^F, u_X^F, T_\psi(\text{do}(z \to \hat{z})), u_Y^F\rangle), \quad (6)$$

where $\hat{z}$ denotes respectively z^{SCF} from Def. 1, and z^{SPCF} from Def. 2. However, note that to compute the effect of the counterfactual treatments on the outcome, we need to perform counterfactual approximations *in two steps*, where the first step measures the counterfactual treatment and the second step measures the effect on the outcome.

Extension for path-specific disparity measures. To analyze our path-specific measures in Def. 2 and 4, we cannot directly apply the existing CNF counterfactual estimation process since it does not work for multiple sequential interventions. Moreover, since we want to compute path-specific measures in non-linear settings, we cannot directly apply the existing linear approach in (Chiappa 2019) either. Hence, we provide a *novel procedure* that *extends* the existing non-linear causal normalizing flows counterfactual approach to approximate *path-specific counterfactuals*.

For computing z^{SPCF} in Def. 2 and 4, we need to perform *sequential interventions* in the CNF *following the causal ordering* of the features to obtain:

$$\begin{aligned} z^{\text{SPCF}} &= T_\psi^{-1}(\langle u_S^{CF}, \hat{u}_X^F, u_Z^F, u_Y^F\rangle), \text{ where} \\ u_S^{CF} &= T_\psi(\text{do}(S \to (1 - s^F))), \text{ and} \\ \hat{u}_X^F &= T_\psi((1 - s^F), \text{do}(X \to x^F)) \end{aligned} \quad (7)$$

Hence, to compute z^{SPCF}, we perform two interventions following the causal ordering. First we use $\text{do}(S \to (1 - s^F))$ on ϑ^F to get ϑ^{SCF}. Then, we use $\text{do}(X \to x^F)$ on ϑ^{SCF}, combine the different exogenous u from the separate intervention steps and generate z^{SPCF}. Further details and pseudocode regarding the different counterfactual computations using CNF can be found in Appendix B.6.

Assumptions and considerations. Our causal framework is designed to incorporate advances in causal generative modeling flexibly, but its current CNF-based implementation relies on key assumptions. Our framework assumes *overlap*, i.e., all relevant treatment-covariate combinations have non-zero support, a condition that generally holds in practice (Appendix B.4). Consistent with prior work (Khemakhem et al. 2021; Sánchez-Martin, Rateike, and Valera 2022; Kusner et al. 2017; Chiappa 2019), we also assume *no hidden confounders* and *full observability* of treatments, covariates, and sensitive attributes from the *decision-maker's perspective*. However, in real-world settings, decision-makers *may not observe* all covariates impacting the outcome, introducing potential confounding. However, this *would only impact* the direct disparity measures (Definitions 2 and 4) since they need interventions on X. Importantly, since decision-makers only intervene on treatment decisions in our framework, robustness can be practically tested through small-scale interventional studies on subpopulations. Future work can also explore integrating recent advancements for tackling confounding (Almodóvar et al. 2025). We provide a more detailed discussion in Appendix B.4. Appendix D presents synthetic data experiments *validating the framework's effectiveness against an oracle*.

4 Mitigating Historical Treatment Unfairness

Our framework identifies when historical data reflects discriminatory treatment policies, i.e., disparities in treatments

Z that are not justified by corresponding benefits in ground-truth outcomes Y. To prevent such biases from being perpetuated in data-driven decision-making, we propose an automated pre-processing procedure. Although pre-processing approaches may lack formal guarantees, they remain a practical and effective mitigation strategy (Mutlu, Yousefi, and Ozmen Garibay 2022). Our method augments the biased dataset $\mathcal{D}$ to produce a fairer version, $\mathcal{D}^{\text{fair}}$, enabling more equitable non-binary decisions (Sec. 4.1). We further show that using $\mathcal{D}^{\text{fair}}$ can mitigate the effects of treatment disparities on risk score estimation, a central component in many decision-making pipelines (Sec. 4.2).

4.1 Fairness in Non-Binary Decision-Making

Automated decision-making systems often rely on historical data, including sensitive attributes, individual covariates, and past treatments to train classifiers that predict outcomes: $h_y : \{S, X, Z\} \mapsto Y$. These predictions are used to make binary decisions, such as granting loans (Corbett-Davies et al. 2017). However, biases in the decision-maker's *non-binary past treatment policy* (Def. 1–4) can propagate into future automated decisions. As we show later in Sec. 5 and Appendix E, optimizing for predictive fairness (Dwork et al. 2012; Hardt, Price, and Srebro 2016) *does not eliminate treatment disparities or their downstream effects*.

To prevent automated systems from reinforcing such disparities, we propose a pre-processing method that uses our causal framework to transform biased historical data $\mathcal{D}$ into a *treatment-fair* dataset $\mathcal{D}^{\text{fair}}$. We define a treatment policy $\pi(Z)$ that adjusts treatment assignments for disadvantaged individuals $S^F = s^-$, those who previously received worse treatment *without benefiting in outcome* (Def. 3). For these cases, $\pi(Z)$ assigns the *sensitive counterfactual* treatment z^{SCF} (Def. 1). The fair dataset is then constructed as:

$$\mathcal{D}^{\text{fair}} = \left\{s^F, x^F, z^F, y^F\right\}_{s=s^+} \bigcup \left\{s^F, x^F, z^{\text{SCF}}, y^{CF}(\text{do}(Z \to z^{\text{SCF}}))\right\}_{s=s^-} \tag{8}$$

For individuals in group $S^F = s^-$, we replace the historically assigned treatment with z^{SCF} and simulate the counterfactual outcome under this intervention (Def. 1 and 3). Despite its simplicity, $\pi(Z)$ corrects disparities without harming outcomes, ensuring $P\left[Y^{CF}(\text{do}(Z \to z^{\text{SCF}})) = 1 \mid S^F = s^-\right] \geq P\left[Y^F = 1 \mid S^F = s^-\right]$. The resulting $\mathcal{D}^{\text{fair}}$ *removes past disparate treatment and its effects*, enabling the training of predictive models that do not perpetuate such disparities.

4.2 Fair Risk Score Estimation

Risk scores are widely used in algorithmic decision-making to inform lending decisions (Hurley and Adebayo 2016; Obermeyer et al. 2019). For example, banks often rely on risk scores from third-party institutions such as FICO or SCHUFA to assess creditworthiness and justify loan terms (Doroghazi 2020; Toh 2023; Loqbox 2023). These scores are typically estimated from historical data $\mathcal{D}$ and past decision outcomes (Maiden 2024; MyFico 2018) as $R(X, Z) = P(Y = 0 \mid X, Z)$. However, disparities in prior non-binary treatment decisions can bias both outcomes and risk scores. Crucially, this bias may persist even if treatment decisions are excluded from the estimation, due to the causal effect of treatment Z on outcome Y.

Data	SCF	Measure	Disparity Amount (K$) Hist.	Disparity Amount (K$) EOD	Effect(%) $Y^F = 0/1$ Hist.	Effect(%) $Y^F = 0/1$ EOD
NY	F→M	TTD(-E)	+30	+28	2.2/0.4	4.8/0.1
		DTD(-E)	+7	+7	2.5/0.2	3.6/0.1
	M→F	TTD(-E)	−30	−28	3.9/0.2	2.6/0.0
		DTD(-E)	−6	−7	1.8/0.1	2.6/0.0
TX	F→M	TTD(-E)	+19	+21	1.6/0.3	0.0/0.0
		DTD(-E)	+3	+3	1.6/0.1	0.0/0.0
	M→F	TTD(-E)	−20	−21	3.0/0.2	0.0/0.0
		DTD(-E)	−4	−3	2.4/0.1	0.0/0.0

Table 1: Treatment disparities and outcome effects in S.1 (HMDA). Disparity in median U.S.D. ($) across sensitive counterfactuals (SCF), female (F) and male (M). $Y^F = 0$: label changed from negative to positive agreement; $Y^F = 1$: reverse. Hist: all test-set loans granted (past policy); EOD: use loan grants predicted by equalized odds predictor.

Since the treatment decisions are beyond an applicant's control (Kleinberg, Mullainathan, and Raghavan 2016; Coston et al. 2020), we argue for mitigating the unfair effects of past treatment policies on risk score estimation. To this end, we formulate a fair risk score estimate as $R_{\text{fair}}(x, s) = \mathbb{E}_{z' \sim \pi(Z)}\left[P\left(y^{CF}(s^F, x^F, \text{do}(Z \to z')) = 0\right)\right]$, where we use $\pi(Z)$ to marginalize out disparate effects of Z. To ensure the marginalization does not systematically disadvantage any group, we set $\pi(Z) = P_{\mathcal{D}^{\text{fair}}}(Z \mid S = s^+)$, using treatments from the advantaged group. Treatment distribution overlap across groups (Appendix B.4) ensures our interventions remain realistic while removing disparate effects.

5 Use Case: Lending Decisions

In this section, we evaluate our framework by (i) auditing historical data to identify treatment disparities and their *potentially discriminatory* effects on outcomes, and (ii) assessing the effectiveness of our mitigation strategies. We evaluate using four real-world lending datasets that span diverse geographic regions, loan structures, and two representative decision-making scenarios (experimental details in Appendix C and additional results in Appendix E):

- **Decision agreement as label (S.1):** The *ground-truth* Y captures a decision's *immediate outcome*, where $Y = 1$ shows *agreement on treatment terms*, e.g., a loan approved by the bank and terms accepted by the applicant. We evaluate this scenario using the 2017 HMDA dataset from New York and Texas (FFIEC 2017).
- **Decision outcome as label (S.2)**: The *ground-truth* Y captures a decision's *downstream outcome*, e.g., $Y = 1$ for loan repaid and $Y = 0$ for default. We analyze this scenario using Home Credit (Montoya, Odintsov, and Kotek 2018) and German Credit (Hofmann 1994) data.

Data	SCF	Measure	Disparity				Effect (%)			
			Annuity (INR)		Amount (INR)		$Y^F = 0$		$Y^F = 1$	
			Hist.	EOD	Hist.	EOD	Hist.	EOD	Hist.	EOD
Home Credit	F → M	TTD(-E)	+1005.26	+1031.86	+7834.47	+8606.90	0.22	0.18	0.25	0.13
		DTD(-E)	+708.24	+694.17	−797.40	−972.88	0.22	0.00	0.08	0.05
	M → F	TTD(-E)	−957.37	−924.41	−6778.88	−7310.10	2.46	3.08	0.05	0.04
		DTD(-E)	−604.94	−529.54	+857.12	+1069.00	1.36	1.54	0.08	0.06
			Duration (months)		Amount (DM)		$Y^F = 0$		$Y^F = 1$	
German Credit	F → M	TTD(-E)	+1.15	0.94	+272.15	+189.89	4.25	0.00	2.70	3.64
		DTD(-E)	+0.42	+0.36	+133.39	+116.74	4.25	0.00	1.35	1.82
	M → F	TTD(-E)	−0.87	−0.67	−237.97	−212.44	12.33	9.10	0.97	0.81
		DTD(-E)	−0.58	−0.42	−189.68	−189.68	12.33	4.54	0.48	0.81

Table 2: Treatment disparities and outcome effects in S2. For sensitive counterfactuals (SCF), female (F) and male (M), $Y^F = 0$: outcome changed defaulted → repaid, $Y^F = 1$ the reverse. Reporting median INR (Indian Rupees) and DM (Deutsche Marks). Hist: all test-loans were granted (past policy), EOD only uses test-loans granted by the equalized odds predictor.

5.1 Measuring Treatment Disparities

We begin by auditing disparities in historical treatment assignments in test data. First, we examine disparities under historical binary decisions using the full test set (Hist.), where all individuals received positive decisions. We also evaluate the impact of predictive fairness by analyzing treatment disparities for positively predicted individuals under an equalized odds (EOD) predictor. Results for other predictors in Appendix E.2 , analysis of unsplit data in Appendix E.1.

S.1: Agreement of the decision-making process as label. In both HMDA datasets, we follow the identification and pre-processing procedures for sensitive attributes, covariates, and treatment decisions outlined in (Cooper et al. 2024). We consider loan amount and preapproval status as treatment decisions within the bank's authority. Table 1 reports the treatment disparity analysis results.

Takeaway. Although loan agreement rates do not differ significantly by gender, our analysis (ref. Hist) shows that banks assign more conservative loan terms to female applicants, suggesting *potential treatment unfairness*. Offering higher loan amounts to women would *not* reduce agreement rates, challenging the justification for such disparities. We also observe that *predictive fairness* (e.g., EOD) *does not eliminate treatment disparities*. For example, Table 1 shows that in the NY dataset, an EOD-compliant predictor still yields disparities: *females, if treated as males, would have received $27K higher loan amounts* without any negative impact on Y, indicating *potential treatment unfairness*. These differences are largely explained by applicant covariates, with minimal direct disparity observed.

Our observations raise important questions about the rationale behind these disparate treatment decisions. Specifically, we ask: *Are disparities in loan terms for female applicants justified by downstream outcomes, such as repayment behavior, and thus defensible as a business necessity?* This motivates our second scenario examining treatment disparities, repayment outcomes, and potential discrimination.

S.2: Decision outcome as label. In the Home Credit dataset, we analyze two treatment decisions: loan amount and annuity amount, while in the German Credit dataset, we examine three treatment decisions: loan amount, loan duration, and installment rate with results summarized in Table 2.

Takeaway. In Scenario S.2, our results show female applicants experience negative treatment disparities *not explained* by downstream outcomes. Conversely, while males receive *perceived positive* treatment, they are ultimately *negatively impacted* by its downstream effects. Notably, our findings align with similar insights observed for German Credit in a simpler linear setting (Kanubala, Valera, and Gupta 2024), reinforcing the validity of our results. Despite appearing favorable, loan terms offered to certain males may increase their default risks and lower their creditworthiness over time. Applying the treatment decisions of their female counterparts to male applicants could mitigate this discrimination and improve repayment outcomes. Importantly, this scenario also further illustrates that *predictive fairness alone does not mitigate treatment disparities or their effects on outcomes*. In German Credit, 9.1% of male applicants who received loans from the EOD predictor but later defaulted *would have repaid under counterfactual treatments*.

5.2 Mitigating Treatment Discrimination

Our results in Table 2 show that banks exhibit negative treatment disparity against female applicants. However, males are *potentially discriminated against* since they experience a negative downstream effect on repayment ability, and treating them as females improves repayment performance. Hence, *treating all applicants conservatively as females* may enhance the overall repayment ability of borrowers, benefiting both the banks and the applicants.

Fair non-binary decision-making. Following Sec. 4.1, we examine the creation of a *treatment-fair* dataset aimed at automating future fair non-binary decision-making tasks. Based on our Scenario S.2 analyses, we identify males as the group needing adjusted treatment to generate our fair dataset, $\mathcal{D}^{\text{fair}}$. To understand the impact of $\mathcal{D}^{\text{fair}}$, we

Dataset	Group	LGD (INR/DM)	ESI (INR/DM)
Home Credit	Female	586692.67	290208.73
	Male ($\mathcal{D}$)	576590.35	422114.31
	Male ($\mathcal{D}^{\text{fair}}$)	**568157.43**	**391110.73**
German Credit	Female	1220.51	312.13
	Male ($\mathcal{D}$)	1163.88	495.65
	Male ($\mathcal{D}^{\text{fair}}$)	**940.03**	**393.86**

Table 3: Fairness analysis in S.2, comparing bank's LGD and applicant's ESI (both: lower better) for various policies. $\mathcal{D}$ is factual data, in $\mathcal{D}^{\text{fair}}$ all applicants treated as females.

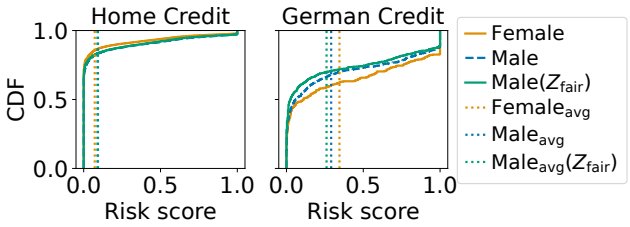

Figure 3: Fair risk score estimation in S.2 for females and males under different treatment distributions $\pi(Z)$. Unfair estimates use the factual distribution; fair estimates apply an interventional distribution, treating all applicants as female.

aim to compare the fair dataset with the original (unfair) dataset on two key metrics: the bank's Loss Given Default (LGD) (Schuermann 2004) and the applicant's Expected Simple Interest (ESI) (Ross, Westerfield, and Jordan 2014). Appendix B.5 contains further details on stakeholder losses.

Takeaway. Table 3 shows at the *data-level* how adjusting male treatments reduces bias and benefits stakeholders by lowering LGD for banks *and* ESI for male borrowers, highlighting the value of fairer treatment. Appendix E.3 further confirms that training binary predictors on $\mathcal{D}^{\text{fair}}$ *improves* outcome utilities while preserving predictive performance.

Fair risk score estimation. Following Sec. 4.2, we now analyze *treatment-fair* risk score estimates. Based on insights from scenario S.2, we focus on male applicants and apply different treatment distributions $\pi(Z)$ to generate risk estimates. We approximate $\pi(Z)$ for males under two conditions: (i) a fair interventional distribution, derived from the empirical distribution of Z conditioned on female applicants to produce "fair" risk scores, and (ii) the factual "unfair" distribution, based on the empirical distribution of Z conditioned on male applicants. For female applicants, $\pi(Z)$ is approximated using the empirical distribution of Z conditioned on female applicants. We compare the empirical cumulative distribution functions (CDFs) of the resulting risk scores for male and female applicants in Fig. 3.

Takeaway. While males generally exhibit lower risk scores than females for German Credit, the pattern reverses for Home Credit. However, historical treatment practices have led to an *unfair overestimation* of males' risk scores, especially in German Credit. This bias is corrected when the risk scores are recalculated under the fair interventional distribution, adjusting for potentially unfair treatments. Integrating these fair risk scores into decision-making pipelines mitigates the residual impacts of historical treatment biases, paving the way for more fair decision-making processes.

6 Discussion

Through a novel causal framework, our work highlighted the overlooked role of non-binary treatment decisions in fairness analyses. This section notes several open challenges (additional discussion in Appendix B.4).

Robust causal frameworks. Our approach assumes access to ground-truth outcomes and no hidden confounders, assumptions that often fail in practice. Hidden confounders (Kilbertus et al. 2020) and proxy labels for unmeasured outcomes (Mhasawade, D'Amour, and Pfohl 2024) can bias fairness measures, particularly when different stakeholders (e.g., decision-makers vs. auditors) apply the framework. Incorporating emerging methods (Almodóvar et al. 2025) is a key step toward addressing these limitations.

Considering diverse stakeholders. Although our preprocessing approach mitigates treatment discrimination, it lacks theoretical guarantees of optimality. Future work should benchmark our method against existing interventions (Coston et al. 2020). Moreover, such approaches may not satisfy differing stakeholder objectives, raising a central question: *What defines a fair treatment policy that accounts for all stakeholders?* For instance, banks may prioritize repayment, while applicants seek lower interest rates (O'Neil and Gunn 2020). Achieving fairness thus requires a *holistic* approach to learning policies that balance these utilities.

Collecting treatment data. While our framework generalizes to various domains (Appendix B.1), we evaluated only in lending owing to major data challenges in other domains. Many datasets conflate covariates with treatments, omit treatment details (e.g., COMPAS (Angwin et al. 2022) lacks bail terms; lending datasets exclude interest rates), or oversimplify complex treatments into binary variables (e.g., IHDP (Madras et al. 2019) reduces care to a single binary indicator). Moreover, most datasets are subject to selective labeling (Lakkaraju et al. 2017), where outcomes are observed only when both decision-makers and individuals agree on terms, ignoring potential *agreement-phase discrimination*. These limitations underscore the need for transparent data collection across domains that includes full treatment assignments (Centre for Public Data 2023) and agreement-phase decisions to enable more comprehensive analyses.

Conclusion. Our work underscores the importance of analyzing the fairness of non-binary treatment decisions given positive binary decisions, e.g., loan approvals. Our novel treatment disparity measures extend existing fairness frameworks toward a more comprehensive evaluation of algorithmic decision-making. Our findings show that fairness cannot rely on technical solutions alone; it requires *socio-technical* approaches that account for diverse stakeholder utilities. By capturing the nuanced impacts of treatment decisions, our framework advances aligning algorithmic systems with societal values, working toward *jointly addressing* fairness in binary predictions and non-binary treatment decisions.

Acknowledgements

We thank Adrián Javaloy for providing invaluable guidance and feedback regarding the causal generative models. This work has been funded by the European Union (ERC-2021-STG, SAML, 101040177). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible.

References

Agier, I.; and Szafarz, A. 2013. Microfinance and gender: Is there a glass ceiling on loan size? *World development*.

Alesina, A. F.; Lotti, F.; and Mistrulli, P. E. 2013. Do women pay more for credit? Evidence from Italy. *Journal of the European Economic Association*.

Almheiri, A. S. 2023. *Automated Loan Approval System for Banks*. Master's thesis, Rochester Institute of Technology.

Almodóvar, A.; Javaloy, A.; Parras, J.; Zazo, S.; and Valera, I. 2025. DeCaFlow: A Deconfounding Causal Generative Model. *arXiv preprint arXiv:2503.15114*.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications.

Centre for Public Data. 2023. Data and Statistical Gaps in Criminal Justice. Available at: https://static1.squarespace.com/static/5ee7a7d964aeed7e5c507900/t/64230b79130cdc7e4b83930f/1680018298114/CFPD+justice+data+gaps+report.pdf, Accessed on 29 July 2025, page 10.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *AAAI conference on artificial intelligence*.

Cooper, A. F.; Lee, K.; Choksi, M. Z.; Barocas, S.; De Sa, C.; Grimmelmann, J.; Kleinberg, J.; Sen, S.; and Zhang, B. 2024. Arbitrariness and social prediction: The confounding role of variance in fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(312): 1–117.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Coston, A.; Mishler, A.; Kennedy, E. H.; and Chouldechova, A. 2020. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 582–593.

Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*.

Doroghazi, R. M. 2020. Fico scores. *American Journal of Cardiology*, 130: 157–158.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Innovations in theoretical computer science conference*.

Escalante, C. L.; Osinubi, A.; Dodson, C.; and Taylor, C. E. 2018. Looking beyond farm loan approval decisions: loan pricing and nonpricing terms for socially disadvantaged farm borrowers. *Journal of Agricultural and Applied Economics*.

FFIEC. 2017. HMDA Data Publication. https://ffiec.cfpb.gov/data-publication/. Accessed: 2025-01-24.

Habehh, H.; and Gohel, S. 2021. Machine learning in healthcare. *Current genomics*, 22(4): 291–300.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*.

Hofmann, H. 1994. Statlog (german credit data) data set. *UCI Repository of Machine Learning Databases*.

Hurley, M.; and Adebayo, J. 2016. Credit scoring in the era of big data. *Yale JL & Tech.*, 18: 148.

Javaloy, A.; Sánchez-Martín, P.; and Valera, I. 2024. Causal normalizing flows: from theory to practice. *Advances in Neural Information Processing Systems*.

Kanubala, D. D.; Valera, I.; and Gupta, K. 2024. Fairness Beyond Binary Decisions: A Case Study on German Credit. *Proceedings of the 3rd European Workshop on Algorithmic Fairness*.

Khemakhem, I.; Monti, R.; Leech, R.; and Hyvarinen, A. 2021. Causal autoregressive flows. In *International conference on artificial intelligence and statistics*, 3520–3528. PMLR.

Kilbertus, N.; Ball, P. J.; Kusner, M. J.; Weller, A.; and Silva, R. 2020. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in artificial intelligence*, 616–626. PMLR.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv*.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*.

Lakkaraju, H.; Kleinberg, J.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Loqbox. 2023. Link Between Credit Score and Interest Rate. https://www.loqbox.com/en-gb/blog/the-link-between-credit-score-and-interest-rates. Accessed: 2025-01-24.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2019. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Conference on fairness, accountability, and transparency*.

Maiden, S. E. 2024. FICO Score. *Darden Business Publishing Cases*, 1–11.

Majumdar, A.; and Valera, I. 2024. CARMA: A practical framework to generate recommendations for causal algorithmic recourse at scale. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Mhasawade, V.; D'Amour, A.; and Pfohl, S. R. 2024. A Causal Perspective on Label Bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Montoya, A.; Odintsov, K.; and Kotek, M. 2018. Home Credit Default Risk. Available at= https://kaggle.com/competitions/home-credit-default-risk. Accessed: 2025-01-21.

Moscato, V.; Picariello, A.; and Sperlí, G. 2021. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications*.

Mutlu, E. Ç.; Yousefi, N.; and Ozmen Garibay, O. 2022. Contrastive counterfactual fairness in algorithmic decision-making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 499–507.

MyFico. 2018. What is in my FICO Scores? Available at=https://www.myfico.com/credit-education/credit-scores/what-is-in-your-score. Accessed: 2025-01-21.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

O'Neil, C.; and Gunn, H. 2020. Near-term artificial intelligence and the ethical matrix. *Ethics of Artificial Intelligence*.

Pearl, J.; et al. 2000. Models, reasoning and inference. *Cambridge, UK: Cambridge University Press*.

Plecko, D.; and Bareinboim, E. 2023. Causal Fairness for Outcome Control. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 47575–47597. Curran Associates, Inc.

Ross, S. A.; Westerfield, R.; and Jordan, B. D. 2014. *Fundamentals of corporate finance*. Irwin New York, USA.

Sánchez-Martin, P.; Rateike, M.; and Valera, I. 2022. VACA: Designing variational graph autoencoders for causal queries. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Schuermann, T. 2004. What Do We Know About Loss Given Default? Working Paper 04-01, Wharton Financial Institutions Center.

Toh, Y. L. 2023. Addressing Traditional Credit Scores as a Barrier to Accessing Affordable Credit. *Economic Review (01612387)*, 108(3).

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 *International conference on world wide web*.