Distinguishing Biases from Personal Preferences: An Honest Machine Learning Approach

Mahyar Habibi[1], Zahra Khanalizadeh[2], Negar Ziaeian[3]

January 2026

Abstract

Estimating individual-level bias in settings with bilateral interactions is challenging because evaluator preferences and item characteristics confound the relationship between group membership and outcomes. We develop a methodology that combines "Honest" Collaborative Filtering (HCF) with Double Machine Learning (DML) to separate genuine bias from preference-based differences. The first stage extracts latent representations of evaluator preferences and item characteristics from observed ratings, using an "honest" design that estimates preferences using only control-group items to prevent contamination by treatment effects. The second stage applies DML to estimate unit-level bias parameters while controlling for these learned embeddings. Monte Carlo simulations demonstrate that HCF+DML substantially outperforms naive OLS estimation under confounding, reducing RMSE by up to 50% and maintaining high correlation with true parameters even under embedding misspecification and non-random selection. We apply the method to nearly 150,000 film reviews from professional critics to estimate gender-based bias in evaluations of female-directed films. Naive comparisons suggest that 29% of critics exhibit statistically significant favoritism toward female directors. After controlling for the match between critic preferences and film characteristics, this figure drops to under 1%. The apparent pro-female pattern largely reflects critics' preferences for genres where female directors are disproportionately represented, rather than gender-based favoritism per se.

Keywords: Discrimination, Bias, Collaborative Filtering, Causal Machine Learning, Double Machine Learning

JEL Codes: C14, C21, J71, L82

[1] Lyft. Email: mhyrhabibi@gmail.com.

[2] University of Washington. Email: zkhnl@uw.edu.

[3] University of Warwick. Email: negar.ziaeian-ghasemzadeh@warwick.ac.uk.

1 Introduction

Discrimination in employment, credit, housing, and other markets remains a central economic and policy concern. In many settings, regulators and researchers need more than an estimate of average discrimination. They need to identify which specific decision-makers systematically treat otherwise comparable individuals differently based on protected attributes such as gender, race, or age. This task is straightforward in field experiments and correspondence studies, where characteristics can be randomized and average treatment effects can be measured cleanly (Bertrand and Mullainathan, 2004a; Bertrand and Duflo, 2017). It is much harder in observational data, where decisions reflect unobserved heterogeneity in preferences, beliefs, and matching patterns, and where each decision-maker is observed in only a sparse subset of possible interactions.

This paper develops a method to estimate *individual-level discrimination* from observational data in bilateral interaction settings. We study environments where units on one side of the market evaluate multiple items or individuals on the other side, and where each item is evaluated by many units. Examples include firms evaluating applicants, lenders evaluating borrowers, landlords evaluating tenants, and critics evaluating films. In these environments, observed outcomes confound at least three forces. First, evaluators have stable latent preferences for certain characteristics. Second, items have latent attributes that directly affect outcomes. Third, evaluators may apply different standards to items associated with a protected attribute. Separating the third channel from the first two is the core identification problem.

We focus on a parameterization that aligns with the economics of discrimination. Following the distinction between taste-based and belief-based discrimination (Becker, 1957; Arrow, 1973; Phelps, 1972), we treat discrimination as an evaluator-specific causal effect of a protected attribute on the evaluator's decision, holding constant the evaluator's latent preferences and the item's latent characteristics. The object of interest is a unit-level bias parameter, θ_j, which measures how evaluator j changes outcomes when the protected attribute changes, net of latent match quality. Estimating θ_j from observational outcomes requires a control strategy for high-dimensional confounding that is typically unobserved and not directly measured.

Our central contribution is a new estimator that combines collaborative filtering with double machine learning to recover unit-level discrimination. Collaborative filtering has become a workhorse method in recommender systems for learning low-dimensional embeddings of users and items from sparse outcome matrices (Koren et al., 2009a). We repurpose this machinery for causal inference. The embeddings summarize latent evaluator preferences and latent item characteristics that would otherwise confound estimates of discrimination. We then use Double/Debiased Machine Learning (DML) (Chernozhukov et al., 2018a) to estimate evaluator-specific treatment effects while flexibly controlling for these high-dimensional embeddings and other observables. DML provides orthogonalization and cross-fitting that allow valid inference on causal parameters even when nuisance functions are learned with modern machine learning methods.

A key challenge is that naive embeddings can leak information about treatment status into the representation itself. If we learn evaluator and item embeddings using all observed outcomes, the latent factors may partially encode the protected attribute and the discrimination signal we aim to measure. This violates the spirit of causal identification because the representation becomes endogenous to treatment. To prevent this, we introduce *Honest Collaborative Filtering* (HCF), inspired by the honest estimation principle in Athey and Imbens (2016a). In our setting, honesty means that preference embeddings are learned using data that exclude treated observations. Concretely, we train collaborative filtering only on untreated interactions (for example, ratings of male-directed films) to obtain preference embeddings that capture stable tastes and match structure without directly absorbing the treatment signal. We then hold these embeddings fixed and estimate unit-level bias parameters using DML. This design mirrors the logic of honest trees and forests, where the data used for model selection are separated from the data used for estimation, thereby reducing contamination of the estimand by adaptive representation learning.

This perspective clarifies how our paper relates to existing work in causal machine learning and recommender systems. A growing literature studies *causal inference for recommendation*, where causal tools are used to debias recommender systems, estimate causal effects of exposure, or improve counterfactual ranking and evaluation. Our goal is different. We develop *recommendation for causal inference.* We use collaborative filtering as a measurement device that constructs controls from the structure of sparse bilateral interaction data, enabling causal identification of discrimination parameters that are not about the recommender system itself. The distinction is important because the representation is not the target of improvement in our setting. It is an input to a causal estimator whose target is a well-defined unit-level treatment effect.

Our method is designed for settings where (i) rich covariate data on items is unavailable, (ii) covariates are protected by privacy constraints, or (iii) the relevant confounders are latent and cannot be directly measured. Many bilateral evaluation settings—peer review, Airbnb ratings, hiring decisions—lack the rich covariate data, making HCF+DML particularly valuable.

We also emphasize what this paper is not about. The algorithmic fairness literature evaluates whether machine learning systems produce disparate outcomes and studies constraints or impossibility results for fair prediction (Kleinberg et al., 2017; Chouldechova, 2017). Our objective is instead to use machine learning to detect discriminatory *human* decision-makers. The concern is not whether an algorithm discriminates, but whether evaluators do, after controlling for latent preferences and latent item attributes that shape outcomes.

Our approach complements recent progress on identifying decision-maker heterogeneity in discrimination using experimental audit data (Kline and Walters, 2021; Kline et al., 2022a, 2023). Those papers provide sharp identification by design and deliver unit-level inference by combining experimental variation with multiple testing control. However, they require large-scale experiments and are infeasible in many settings where rich observational interaction data already exist. We provide a method that can be applied when experimental manipulation or quasi-random assignment is unavailable, but where the bilateral interaction structure is informative and outcomes are plentiful.

We make three contributions. First, we introduce an estimator, *Honest Collaborative Filtering + DML*, that identifies unit-level discrimination parameters from observational bilateral interaction data. Second, we propose an honest representation learning design for collaborative filtering that prevents treatment leakage into preference embeddings, adapting the honest estimation logic of Athey and Imbens (2016a) to matrix factorization. Third, we show how to integrate these embeddings into a DML pipeline to obtain valid unit-level causal estimates and inference in high-dimensional settings (Chernozhukov et al., 2018a). Together, these components allow discrimination measurement with minimal input requirements: a record of who interacted with whom and the resulting outcome.

We evaluate the method in both controlled simulations and an empirical application. In Monte Carlo experiments that simulate a gender-differentiated evaluation process with sparse matching, latent preferences, and latent item characteristics, we compare our estimator to a benchmark OLS approach that does not adequately control for latent confounding. We study performance across confounding regimes and show how honesty in the embedding stage affects bias and variance of unit-level estimates. We then apply the method to film ratings data from Metacritic, using nearly 150,000 critic-film ratings covering over 8,000 films to estimate critic-level bias toward female-directed films. The empirical results illustrate why latent preference control matters. Naive approaches can attribute systematic taste differences and selection patterns to discrimination, producing inflated counts of "significant" decision-makers. Our method sharply reduces these false positives and isolates the subset of evaluators whose rating behavior is consistent with a causal effect of director gender.

The remainder of the paper is organized as follows. Section 2 situates our contribution in the literatures on discrimination, causal machine learning, and collaborative filtering. Section 3 presents the model, the HCF construction, and the DML estimation and inference procedure. Section 4 presents the application to film critics. Section 5 reports simulation results. Section 6 concludes.

2 Literature Review

This paper contributes to several interconnected strands of research spanning the economics of discrimination, causal machine learning, collaborative filtering, algorithmic fairness, and the identification of latent structure in networked data. We draw connections across these literatures while identifying the specific gaps our methodology addresses.

The economic analysis of discrimination has developed around two principal theoretical frameworks. Becker (1957) introduced taste-based discrimination, modeling prejudice as a preference parameter whereby discriminating agents act as though they are willing to pay a cost—either directly or through reduced income—to avoid interaction with members of certain groups. In contrast, statistical discrimination, formalized by Arrow (1973) and Phelps (1972), posits that differential treatment arises not from animus but from rational inference under incomplete information: when individual productivity signals are noisy, decision-makers may rely on group-level statistics, generating disparate outcomes even absent prejudice. Empirical detection of discrimination has proceeded along two main avenues. Correspondence studies pioneered by Bertrand and Mullainathan (2004b) send fictitious applications with randomly assigned characteristics to employers, providing clean experimental identification of differential treatment. Their seminal study documented that resumes with distinctively White-sounding names received 50 percent more callbacks than identical resumes with African American-sounding names. While such experiments provide compelling evidence of average discrimination, they face important limitations: they are costly to implement, ethically constrained in many settings, and typically identify only aggregate effects rather than which specific decision-makers discriminate.

Recent work has advanced toward detecting discrimination at the decision-maker level. Kline et al. (2022b) conducted a massive correspondence experiment sending over 83,000 fictitious applications to 108 of the largest U.S. employers, finding that distinctively Black names reduced contact rates by 2.1 percentage points on average. Crucially, they documented substantial between-firm heterogeneity, with a standard deviation in racial contact gaps of 1.9 percentage points across firms. Using empirical Bayes methods, they identified 23 individual companies that discriminate against Black applicants while controlling the false discovery rate. This represents a methodological advance toward unit-level discrimination detection, though it still relies on experimental variation. An alternative approach exploits quasi-experimental variation from random assignment of decision-makers. Arnold et al. (2018) developed a test for racial bias in bail decisions building on the Becker framework. Their insight is that if judges are unbiased, marginal defendants of different races should exhibit identical misconduct rates conditional on release. Using the leniency of quasi-randomly assigned bail judges as an instrument, they find that marginally released White defendants have substantially higher misconduct rates than marginally released Black defendants, consistent with racial bias against Black defendants. This methodology elegantly identifies bias at the margin but requires quasi-random assignment of cases to decision-makers—a feature unavailable in many economically important settings. Our paper complements this literature by providing an observational method for detecting individual-level discrimination in bilateral interaction settings where experimental manipulation is infeasible and no natural experiment provides quasi-random assignment.

The application of machine learning to causal inference has expanded rapidly, motivated by the recognition that modern ML methods excel at prediction but do not automatically deliver valid estimators of causal parameters. Chernozhukov et al. (2018b) introduced Double/Debiased Machine Learning (DML), demonstrating how ML can be used to estimate high-dimensional nuisance parameters while maintaining valid inference on low-dimensional parameters of interest. Their framework rests on two key ingredients: Neyman-orthogonal moment conditions that reduce sensitivity to nuisance parameter estimation error, and cross-fitting to prevent overfitting bias from contaminating causal estimates. The resulting estimators achieve $\sqrt{n}$-consistency and asymptotic normality under weak conditions on the ML learner's convergence rate. A parallel literature has developed methods for estimating heterogeneous treatment effects. Athey and Imbens (2016b) proposed causal trees that adapt recursive partitioning to discover subpopulations with differing treatment effects. Their central innovation is "honest" estimation, whereby one sample is used to construct the partition and a separate sample to estimate treatment effects within each cell. This sample-splitting ensures valid confidence intervals even when the partition is data-driven. Wager and Athey (2018) extended this approach to causal forests, providing theoretical results establishing pointwise consistency and asymptotic normality for random forest estimators of conditional average treatment effects. Our methodology synthesizes insights from both the DML and HTE literatures. Following Chernozhukov et al. (2018b), we treat collaborative filtering embeddings as high-dimensional nuisance parameters and employ cross-fitting to prevent overfitting contamination. Inspired by Athey and Imbens (2016b), we implement "honest" collaborative filtering, whereby embeddings are learned on data partitions that exclude the observations used to estimate the corresponding treatment effects, ensuring that model selection does not contaminate causal estimation.

Collaborative filtering methods have become foundational in recommendation systems, particularly since the Netflix Prize competition demonstrated their practical effectiveness. Koren et al. (2009b) provided an authoritative treatment of matrix factorization techniques, showing that regularized matrix factorization outperforms classical nearest-neighbor methods for rating prediction. In regularized matrix factorization, the observed rating matrix is decomposed into the product of lower-dimensional user and item factor matrices, with regularization controlling overfitting. Each user is represented by a latent vector capturing their preferences, and each item by a latent vector capturing its characteristics. We repurpose this machinery for causal inference rather than prediction. In our setting, critics correspond to users, films to items, and the treatment variable is director gender. The key insight is that critic and film embeddings learned via collaborative filtering capture precisely the confounders that would bias naive estimates of discrimination: critics' systematic preferences for certain film styles, and films' latent characteristics that affect ratings independent of director gender. By including these embeddings as controls in a DML framework, we absorb confounding variation that would otherwise be attributed to gender bias. While the standard use of collaborative filtering is predictive (imputing missing ratings from partial observations), we repurpose its latent factor structure to construct a low-dimensional summary of high-dimensional confounding.

This application stands in contrast to the growing literature that applies causal inference to improve recommender systems (e.g., Gao et al. (2024)). While those works focus on debiasing recommendations or counterfactual prediction—essentially treating the recommendation as the intervention—we invert this relationship. We exploit CF not to improve the system itself, but as a measurement tool to achieve causal identification of treatment effects for pre-determined characteristics (in our case, director gender). To our knowledge, this use of CF as a "pre-processor" for high-dimensional controls in a causal framework represents a novel departure from the standard predictive or system-centric causal literature.

A burgeoning literature in computer science addresses fairness in algorithmic decision-making. Kleinberg et al. (2017) prove fundamental impossibility results: except under restrictive conditions, no classifier can simultaneously satisfy calibration within groups, balance for the positive class, and balance for the negative class. Chouldechova (2017) independently established related impossibility theorems, demonstrating that predictive parity and error rate balance cannot both hold when base rates differ across groups. We note explicitly what our paper is not about. The algorithmic fairness literature asks whether machine learning systems discriminate against protected groups—whether an algorithm's predictions or decisions exhibit disparate impact. We flip this question: can machine learning help identify discriminatory human decision-makers? Rather than auditing algorithms for bias, we use algorithms as tools to detect bias in the humans whose decisions generated our data. The fairness literature's concern is ensuring ML systems treat groups equitably; our concern is using ML to measure whether human evaluators treat groups equitably after controlling for legitimate preference-based differences.

Our approach shares methodological kinship with recent work on identifying latent structure from observed outcomes. Griffith and Peng (2023) address a fundamental challenge in network econometrics: when researchers observe only outcome covariances across agents, how can they separately identify network spillovers from correlated unobservable factors? Their identification strategy exploits the distinct structural properties of network effects versus factor structure. Network adjacency matrices are typically sparse, while factor loadings generate dense, low-rank covariance patterns. Using Turán's Theorem from combinatorial graph theory, they show that restrictions on network density and maximum degree suffice to separately identify the sparse network structure from low-rank latent factors. The parallel to our setting is instructive. We observe a sparse matrix of critic-film ratings and seek to identify individual-level bias parameters while controlling for latent critic preferences and film characteristics. The collaborative filtering decomposition separates the dense, low-rank component captured by embeddings from sparse residual variation. Our "honest" design serves a function analogous to their sparsity restrictions: both prevent a nuisance channel from contaminating the object of interest. Where they seek to identify network structure itself, we seek to identify unit-level bias parameters while controlling for latent structure.

Our paper occupies a distinctive position at the intersection of these literatures. From discrimination economics, we inherit the substantive question and Beckerian theoretical framework. From causal machine learning, we adopt the DML estimation strategy and honest sample-splitting. From collaborative filtering, we borrow the factorization machinery for learning latent representations. From the network identification literature, we draw the insight that structural assumptions can separate confounded channels in sparse data. The primary gap we address is methodological: existing approaches to individual-level discrimination detection require either experimental manipulation or quasi-random assignment. Settings with bilateral evaluation data—critics rating films, guests reviewing hosts, students rating professors—typically offer neither. Our honest collaborative filtering approach provides a third path: using the structure of the rating matrix itself to construct controls that absorb confounding preferences, enabling causal identification of individual-level bias without experimental variation.

3 Conceptual Framework and Methodology

The conceptual framework of this study is based on interactions between two distinct sets of entities: individuals $i \in I$ who are evaluated, and evaluators $j \in J$ who assess them. There is a many-to-many relationship between the two sets: each individual i may be evaluated by multiple evaluators, and each evaluator j may assess multiple individuals. This setting is analogous to the job market, where applicants apply to multiple employers and employers receive applications from multiple candidates. Following this analogy, we refer to individuals $i \in I$ as applicants and entities $j \in J$ as employers throughout this section, though the framework applies equally to other bilateral evaluation settings such as film critics reviewing movies or teachers grading students.

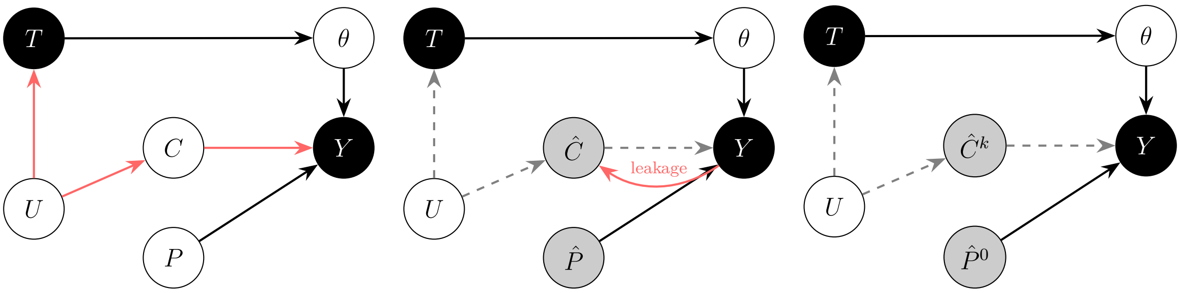
Each applicant is associated with a trait $T_i \in \{0, 1\}$ (e.g., gender or race) that may be subject to bias from employers. Beyond potential biases, employers have preferences P_j over applicant characteristics C_i that may be correlated with the trait. The researcher observes only the trait T_i and outcome Y_{ij} (e.g., hiring decision or rating), but not the latent preferences or characteristics. The goal is to determine whether and which employers exhibit bias in their evaluations, beyond what is attributable to legitimate preferences over applicant characteristics.

3.1 The Identification Problem

Figure 1 illustrates the identification challenge using causal diagrams (Hernán and Robins, 2020). The treatment T_i (e.g., applicant gender) affects the outcome Y_{ij} through two channels. The first is the *direct* channel through evaluator bias θ_j: evaluators may systematically favor or disfavor individuals based on group membership, independent of qualifications. The second is the *indirect* channel through characteristics C_i: if treated individuals differ systematically in their latent characteristics—for instance, if female applicants cluster in certain fields or female directors concentrate in certain genres—then evaluators with preferences P_j over these characteristics will rate treated individuals differently even absent any bias.

In the figure, black nodes (T, Y) denote variables observed by the researcher, while white nodes (U, C, P, θ) represent latent or unobserved quantities. Gray nodes indicate variables that are controlled for in estimation. The unobserved factor

U represents systematic differences in how treated and untreated individuals sort into different types of work (genres, fields, industries), creating a backdoor path $T \leftarrow U \rightarrow C \rightarrow Y$ that confounds naive estimates of bias. Our goal is to estimate θ_j for individual evaluators $j \in J$ by blocking this confounding path while avoiding post-treatment contamination.



(a) Without CF: confounded

(b) Naive CF: leakage risk

(c) Honest CF: identified

Figure 1: **Causal structure and identification.** Observed variables are shown in black, unobserved in white, and controlled variables in gray. T denotes the treatment (e.g., female applicant), Y the outcome (rating or callback), θ the evaluator-specific bias, C the latent characteristics of the evaluated individual, P the evaluator's preferences, and U unobserved factors (e.g., field or genre) that influence both treatment and characteristics. **(a)** Without controlling for latent factors, the backdoor path $T \leftarrow U \rightarrow C \rightarrow Y$ confounds the effect of T on Y, biasing naive estimates of θ. **(b)** Naive collaborative filtering estimates $\hat{C}$ from all outcomes, including Y. This creates a feedback path where evaluator j's bias contaminates their own control variables. **(c)** Honest CF estimates $\hat{P}^0$ from untreated observations only and cross-fits $\hat{C}^k$ by excluding evaluator j's own ratings, eliminating contamination and achieving identification.

Formally, the researcher's objective is to estimate θ_j for each evaluator $j \in J$ from the following model:

$$Y_{ij} = \theta_j T_i + g(C_i, P_j) + \varepsilon_{ij} \tag{1}$$

where Y_{ij} is the outcome, T_i is the treatment indicator, $g(\cdot)$ captures how the match between applicant characteristics C_i and employer preferences P_j affects outcomes, and ε_{ij} is an idiosyncratic error. The challenge is that neither C_i, P_j, nor $g(\cdot)$ is observed. How can one account for unobserved characteristics and preferences when no direct measurements are available?

3.2 Collaborative Filtering

The answer lies in collaborative filtering (CF), a technique widely used in machine learning for recommendation systems. CF exploits the insight that if person A agrees with person B on some items, A is likely to share B's opinion on other items as well. In essence, CF constructs latent representations of users' preferences and items' characteristics from observed interaction patterns, then uses these representations to predict unobserved interactions.

We employ Regularized Matrix Factorization (RMF), a fundamental CF method requiring minimal data (Koren et al., 2009a). RMF begins with the outcome matrix R of dimension $|I| \times |J|$, where entry Y_{ij} records evaluator j's assessment of individual i, with missing entries where no evaluation occurred. This matrix is typically sparse, as evaluators assess only a fraction of all individuals. RMF decomposes the high-dimensional sparse matrix R into two lower-dimensional matrices: C of dimension $|I| \times d$ (individual characteristics) and P of dimension $|J| \times d$ (evaluator preferences), where the embedding dimension $d \ll |I|, |J|$ is a hyperparameter. The objective is for the product CP' to approximate the observed entries in R. Using mean squared error loss with regularization on the parameter magnitudes, the loss function is:

$$L = \sum_{(i,j) \in M} (Y_{ij} - P_j C_i')^2 + \lambda \left(\frac{1}{|I|} \sum_i \|C_i\|^2 + \frac{1}{|J|} \sum_j \|P_j\|^2 \right)$$

where M denotes the set of observed entries and λ is a regularization parameter.

The key property of RMF is that it distills information from the outcome matrix into compact latent representations. The matrix P captures evaluators' latent preferences, while C captures individuals' latent characteristics. In constructing these latent spaces, the model positions similar evaluators—and similar individuals—close to one another in their respective embedding spaces. This allows RMF to infer rich representations of both parties based solely on observed evaluation outcomes.

3.3 Honest Collaborative Filtering

Applying standard CF to study discrimination is problematic because it may incorporate evaluator biases into the learned embeddings. If a biased evaluator systematically rates treated individuals lower, this pattern will be reflected in both the evaluator's preference vector and the individuals' characteristic vectors, contaminating the controls we seek to construct.

To address this, we propose *honest collaborative filtering*, an approach inspired by the “honest trees” of Athey and Imbens (2016a). The method has two components.

First, to ensure that trait-based biases do not contaminate evaluator preferences, we estimate preferences using only outcomes from the untreated group. Specifically, we factorize the restricted matrix $R_0 = \{Y_{ij} : T_i = 0\}$, which includes only evaluations of individuals with $T = 0$. This yields preference estimates $\hat{P}^0$ that reflect evaluators’ tastes over characteristics, uncontaminated by any bias toward the treated group.

Second, to ensure that an evaluator’s own bias does not contaminate the characteristic estimates used as controls for that evaluator, we employ cross-fitting. We partition the set of evaluators J into K disjoint subsets $J_1, \ldots, J_K$. For each subset J_k, we construct the outcome matrix R_{-k} using only evaluations from evaluators *not* in J_k, then factorize R_{-k} to obtain characteristic estimates $\hat{C}^k$. When estimating bias for evaluators in fold J_k, we use $\hat{C}^k$ as controls. This ensures that evaluator j’s ratings do not influence the characteristic estimates used in j’s own regression.

With these honest embeddings, Equation (1) becomes:

$$Y_{ij} = \theta_j \, T_i + g(\hat{C}_i^k, \hat{P}_j^0) + \varepsilon_{ij} \tag{2}$$

where the embeddings $\hat{P}^0$ and $\hat{C}^k$ are constructed to avoid the contamination illustrated in Figure 1b.

3.4 Double Machine Learning Estimation

Two challenges remain in estimating Equation (2). First, the function $g(\cdot)$ relating embeddings to outcomes is unknown and must be estimated from data. Second, the selection process determining which evaluator-individual pairs we observe may depend on both preferences and characteristics, creating additional confounding.

To address these challenges, we adopt the Double/Debiased Machine Learning (DML) framework of Chernozhukov et al. (2018a). Consider the partially linear model:

$$\begin{gathered} Y = \theta_0 \, D + g_0(X) + U, \quad \mathbb{E}[U \mid X, D] = 0 \\ D = m_0(X) + V, \quad \mathbb{E}[V \mid X] = 0 \end{gathered}$$

where D is the treatment, Y is the outcome, and X is a vector of controls. The first equation models the outcome as a function of treatment and controls; the second models treatment assignment as a function of controls. Under conditional exogeneity—that is, treatment is as good as random conditional on X—the parameter θ_0 has a causal interpretation.

When X is high-dimensional, standard estimation of g_0 and m_0 is infeasible. A naive approach using machine learning to estimate $\theta_0 D + g_0(X)$ directly yields inconsistent estimates due to regularization bias. Chernozhukov et al. (2018a) overcome this via orthogonalization: first estimate $\hat{m}_0$ and $\hat{g}_0$ on an auxiliary sample, then compute residualized treatment $\hat{V}_i = D_i - \hat{m}_0(X_i)$ and residualized outcome $\tilde{Y}_i = Y_i - \hat{g}_0(X_i)$. The debiased estimator is:

$$\hat{\theta}_0 = \frac{\sum_i \hat{V}_i \tilde{Y}_i}{\sum_i \hat{V}_i^2}$$

This is simply OLS of the residualized outcome on the residualized treatment. By partialling out the effect of X from both D and Y, regularization bias is eliminated.

We adapt this framework to estimate evaluator-level bias parameters. The model becomes:

$$\begin{gathered} Y_{ij} = \theta_j T_i + g(\hat{P}_j^0, \hat{C}_i^k) + \varepsilon_{ij} \\ T_i = m(\hat{P}_j^0, \hat{C}_i^k) + \epsilon_{ij} \end{gathered} \tag{3}$$

The first equation models outcomes; the second captures how preferences and characteristics influence which individuals are observed by which evaluators (the selection or matching process). For each evaluator j, we estimate θ_j using the DML procedure applied to j's subset of observations.

Algorithm 1 summarizes the complete methodology.

Algorithm 1 Honest Collaborative Filtering with Double Machine Learning

```
1. Estimate honest preferences: Construct $R_0 = \{Y_{ij} : T_i = 0\}$ and factorize to obtain $\hat{P}^0$.
2. Partition evaluators: Split $J$ into $K$ disjoint folds $J_1, \ldots, J_K$.
3. Cross-fit characteristics: For each fold $k \in \{1, \ldots, K\}$:
    (a) Construct $R_{-k}$ using ratings from evaluators not in $J_k$;
    (b) Factorize $R_{-k}$ to obtain $\hat{C}^k$.
4. Estimate bias via DML: For each evaluator $j \in J_k$, estimate $\theta_j$ from Equation (3) using embeddings $(\hat{P}_j^0, \hat{C}^k)$.
```

3.5 Identification

We now state the assumptions required for causal identification of evaluator-level bias parameters θ_j. Let Y_{ij} denote the outcome when evaluator j assesses individual i, $T_i \in \{0, 1\}$ indicate membership in the treated group, and $C_i \in \mathbb{R}^d$ denote latent characteristics of individual i.

Assumption 1 (Latent Factor Structure). *The conditional expectation of outcomes follows an additive structure:*

$$E[Y_{ij} \mid T_i, C_i, P_j] = \theta_j \cdot T_i + g(C_i, P_j) \tag{4}$$

where θ_j is evaluator j's bias parameter and $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ captures the match quality between individual characteristics and evaluator preferences.

This assumption embeds our causal parameter within a collaborative filtering framework. The function $g(C_i, P_j)$ represents how well individual i's characteristics align with evaluator j's preferences. The additive separability implies that bias operates as a constant shift in outcomes for treated individuals, independent of their characteristics.

Assumption 2 (Unconfoundedness). *Conditional on latent characteristics, treatment is independent of potential outcomes:*

$$T_i \perp\!\!\!\perp Y_{ij}(t) \mid C_i \quad \text{for } t \in \{0, 1\} \tag{5}$$

where $Y_{ij}(t)$ denotes the potential outcome under treatment status t.

This is the key identifying assumption. It requires that the latent characteristics C_i capture all systematic differences between treated and untreated individuals that affect outcomes. The assumption would be violated if unobserved factors affect both treatment and outcomes beyond what C_i captures.

Note that identification requires conditioning on C_i but not P_j. Since P_j is a characteristic of evaluators rather than evaluated individuals, it does not lie on any backdoor path between T_i and Y_{ij}. However, controlling for P_j improves precision and allows flexible modeling of match quality via collaborative filtering.

Assumption 3 (Overlap). *For all C_i in the support of the data:*

$$0 < P(T_i = 1 \mid C_i) < 1 \tag{6}$$

Overlap ensures that for any configuration of latent characteristics, we observe both treated and untreated individuals, enabling comparison of similar individuals who differ only in treatment status.

Assumption 4 (Ignorable Selection). *Conditional on treatment and latent factors, selection into observation is independent of potential outcomes:*

$$Y_{ij}(t) \perp\!\!\!\perp \mathbf{1}[(i, j) \in M] \mid T_i, C_i, P_j \quad \text{for } t \in \{0, 1\} \tag{7}$$

where M denotes the set of observed evaluator-individual pairs.

This assumption requires that which pairs we observe depends only on preferences and characteristics, not on the potential outcome itself.

Assumption 5 (Embedding Consistency)**.** *The collaborative filtering procedure recovers embeddings satisfying:*

$$\left\| g(\hat{C}_i, \hat{P}_j) - g(C_i, P_j) \right\|_2 = o_p(n^{-1/4}) \tag{8}$$

This technical condition ensures estimation error in the nuisance functions does not contaminate inference on θ_j. The $n^{-1/4}$ rate is the standard requirement for DML (Chernozhukov et al., 2018a). Matrix factorization achieves this rate under low-rank assumptions (Candes and Recht, 2008; Koltchinskii et al., 2011).

Identification Result. Under Assumptions 1–5, the bias parameter θ_j is identified. The honest design (learning $\hat{P}^0$ from untreated observations and cross-fitting $\hat{C}^k$) ensures that (i) preferences are not contaminated by treatment effects, and (ii) characteristics are not contaminated by the evaluator's own bias. These design choices block the feedback paths in Figure 1b and yield valid causal inference.

4 Movie Critics' Bias Toward Female-Directed Movies

The ensuing section presents an empirical application of the established methodology using a real-world dataset. This approach is instrumental in demonstrating the framework's efficacy to discern micro-level discrimination in practical settings.

4.1 Data Description

A dataset comprising film reviews from professional critics was constructed using data from Metacritic.com, a review aggregator website. Metacritic collects reviews from approximately 100 sources, assigning ratings on a uniform scale of 0-100. These ratings were transformed to a 0-1 scale for this analysis[4]. The objective is to apply the methodology described in Section 3 to explore potential discriminatory patterns in critics' reviews of movies directed by women.

Metacritic provides detailed information, such as the names of film directors. The gender of directors was deduced using their first names and the *gender-guesser*

[4] In cases where an explicit rating is absent, Metacritic's evaluators assign a score reflecting their assessment of the article.

Table 1: Summary Statistics of Selected Variables

	Count	Mean	Std.	Min	Med	Max
Year	8,284	2008.7	8.07	1990	2010	2021
Critic Rating	145,522	0.631	0.210	0	67	100
Films' # of Reviews	8,284	17.6	9.00	1	16	47

Notes: The table shows summary statistics for the selected variables in the data. The data is limited to reviews from critics who have evaluated at least 30 movies directed by women.

Python library[5], a prevalent tool for name-based gender inference. Entries were removed if *gender-guesser* was unable to make a prediction (name not found in its database) or if the name was non-specific to a particular gender. This gender identification procedure was verified for accuracy against a Wikipedia directory of female directors[6], with a misclassification rate under 5%. To maintain simplicity in the analysis, films with more than one director were excluded, as over 95% of movies in the dataset had a single director.

Table 1 offers a summary of statistics for selected variables in the dataset. Data was collected for films released from 1990 to 2021 and having at least seven critic reviews on Metacritic. This was further narrowed down to critics who had reviewed a minimum of 30 films directed by female directors. The filtered dataset contains over 145,000 reviews from 205 critics, spanning around 8,300 films and 3,900 directors. Films directed by women constitute nearly 14% of the dataset. Each film, on average, garnered reviews from more than 17 critics, with an average rating of 0.63 on a 0-1 scale.

4.2 Estimation

As underscored earlier in this study, the estimation of discrimination or favoritism at the individual level is of considerable significance for several reasons. Chief among them is the potential for aggregate-level bias estimates to be misleading. To illustrate, consider a hypothetical scenario in our context: a seemingly minor bias against movies directed by women could arise either from a general absence of discrimination among critics or from the presence of two distinct groups of reviewers – one disproportionately critical and the other overly favorable towards female-directed films. While both scenarios lead to similar estimates of aggregate-level bias, they depict starkly different realities of micro-level discrimination.

[5] https://pypi.org/project/gender-guesser/

[6] https://en.wikipedia.org/wiki/List_of_female_film_and_television_directors

In the assessment of individual-level biases or favoritism, the role of personal preferences among decision-makers is pivotal. For instance, in this analysis, a critic's preference for particular genres or themes – more frequently found in films directed by either gender – might inadvertently color their reviews. This genre or theme preference could manifest as apparent gender bias in reviews, while it truly stems from the critic's own cinematic tastes. Overlooking these personal preferences risks incorrectly categorizing critics as biased.

Therefore, we implement the method outlined in Section 3 to estimate individual-level bias/favoritism regarding critics' evaluation of female-directed films. The approach involves estimating the following Double Machine Learning (DML) model

$$\begin{aligned} r_{i,j} &= \theta_j \, FD_{i,j} + g(P_j^0, C_i^k) + \varepsilon_{j,i} \\ FD_{i,j} &= m(P_j^0, C_i^k) + \epsilon_{i,j} \end{aligned} \tag{9}$$

Here, $r_{i,j}$ represents the rating given by critic j to film i, while $FD_{i,j}$ is a binary indicator denoting whether film i was directed by a woman. The parameter θ_j is indicative of the critic-specific bias/favoritism towards films directed by women.

To describe the method in practice, consider the following example of a matrix of ratings,

$$R_{n\times m} = \begin{bmatrix} r_{1,1} & \text{-} & r_{1,3} & \text{-} & \cdots & r_{1,m} \\ r_{2,1} & \text{-} & r_{2,3} & \text{-} & \cdots & \text{-} \\ \text{-} & r_{3,2} & \text{-} & r_{3,4} & \cdots & \text{-} \\ r_{4,1} & \text{-} & \text{-} & \text{-} & \cdots & r_{4,m} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & \text{-} & r_{n,3} & \text{-} & \cdots & r_{n,m} \end{bmatrix}$$

In this dataset, typically, critics review only a limited selection of films, and correspondingly, each film is assessed by a small group of critics. With a total of over 8,000 films and approximately 200 critics, the dataset comprises less than 150,000 observed ratings, indicating that under 10% of all possible ratings are recorded. In matrix R, a '-' signifies a missing rating, denoting a film that a specific critic did not review. This *sparsity* is a common feature in similar contexts where each item or application is evaluated by only a fraction of potential reviewers. Notably, the use of Collaborative Filtering in industrial settings is intended to predict ratings that a user might assign to items they have not yet reviewed (such as books, music, or movies) and to recommend items likely to be highly rated by the user.

As outlined in Algorithm 1, the procedure begins by applying regularized matrix factorization (RMF) to decompose the rating matrix $R_{I\times J}$ into item embeddings $C^0_{I\times d}$ and evaluator embeddings $P^0_{J\times d}$. RMF initializes P and C randomly and then minimizes the regularized objective in Section 3 using gradient-based optimization. We fix the regularization parameter at $\lambda = 0.01$. We select the embedding dimension d by cross-validation, evaluating $d \in \{10, 25, 50, 100, 200\}$ and choosing the value that minimizes the validation mean squared error. In this phase, RMF is applied exclusively to the matrix of ratings for films directed by men. This application aimed to generate P^0, signifying the matrix of critics' latent preferences, deliberately isolated from their evaluations of films directed by women.

Upon deriving P^0, the second step of Algorithm 1 involved randomly dividing critics into $K = 10$ subsets, For each subset k, RMF was then applied to the ratings matrix R_{-k}, comprising ratings data from critics in the remaining subsets, to generate C^k. Here, C^k indicates the film characteristics' embeddings, isolated from the ratings by critics in that particular subset.

In the final step, P^0 and the $\{C^k\}$ matrices were used to obtain DML estimates for the model in Equation 9. For the nuisance function $m(\cdot)$, we use a binary logistic classifier with ℓ_2 regularization, with the regularization strength set to the default value in *scikit-learn.* For the outcome model $g(\cdot)$, we use a Random Forest regressor and follow the default hyperparameter choices recommended for partially linear regression in the `DoubleML` package (Bach et al., 2024, 2022). We employ 5-fold cross-fitting for valid inference. Standard errors are computed using the influence function approach standard in DML, which treats embeddings as fixed. In principle, a bootstrap procedure re-estimating embeddings in each replication would account for first-stage uncertainty; we leave this refinement to future work. Since our primary goal is to clarify the proposed methodology, we do not perform hyperparameter tuning or a systematic comparison across learners; in practice, both tuning and alternative learners provide useful robustness checks.

To draw a comparison between the outcomes derived by the proposed method and those from a conventional approach, we also estimated the following Ordinary Least Squares (OLS) model:

$$r_{i,j} = \alpha + \beta_j \, FD_{i,j} + \gamma_j + e_{j,i} \tag{10}$$

In this model, β_j represents the OLS estimate of critic j's discrimination/favoritism towards films directed by women. The terms γ_j denotes the critics' fixed effects.

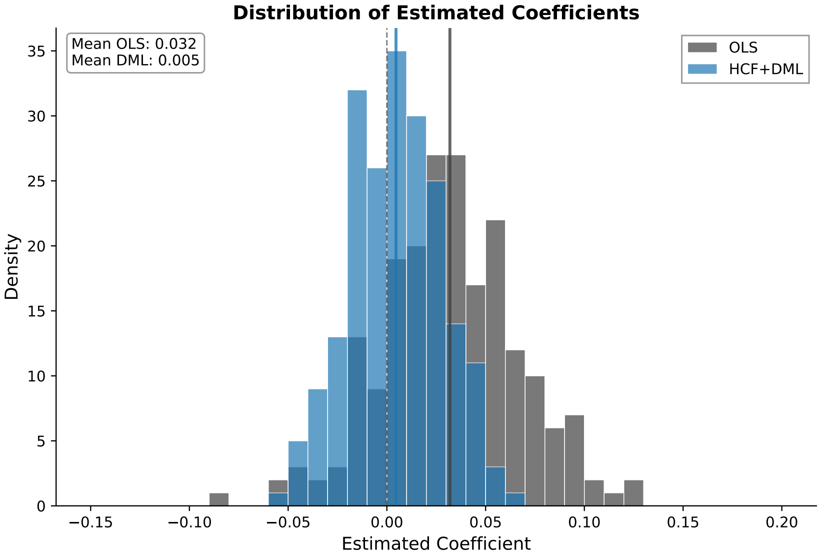
4.3 Results

We estimate critic-level bias parameters using both naive OLS and HCF+DML, then compare the distributions, magnitudes, and statistical significance of the resulting estimates.

4.3.1 Distribution of Bias Estimates

Figure 2 displays the distribution of estimated bias coefficients from both methods. The OLS estimates are centered around 0.032, suggesting that critics rate female-directed films approximately 3.2 percentage points higher on the normalized 0–100 scale. The HCF+DML estimates are substantially attenuated, centered near 0.005. Beyond the shift in central tendency, the distributions differ markedly in their tails: OLS exhibits a heavier right tail, implying more critics with large positive estimates, while HCF+DML produces a tighter, more symmetric distribution.

Figure 2: Distribution of Critic-Level Bias Estimates



Notes: The figure displays kernel density estimates of critic-level bias coefficients. OLS estimates (orange) are obtained from Equation 10, regressing ratings on director gender without controls. HCF+DML estimates (blue) are obtained from Equation 9, controlling for learned embeddings of critic preferences and film characteristics. Positive values indicate higher ratings for female-directed films.

Table 2 summarizes the estimation results. The mean OLS estimate is 0.032, while the mean HCF+DML estimate is 0.005, a reduction of approximately 85%. The standard deviation of estimates also decreases from 0.035 (OLS) to 0.023 (HCF+DML), indicating that controlling for confounding compresses the distribution of estimated biases. The range of estimates narrows correspondingly: OLS estimates span $[-0.080, 0.127]$ while HCF+DML estimates span $[-0.052, 0.069]$.

Table 2: Summary of Critic-Level Bias Estimates

Statistic	OLS	HCF+DML
Number of critics	205	205
Mean estimate	0.032	0.005
Median estimate	0.033	0.005
Std. deviation	0.035	0.023
Min	−0.080	−0.052
Max	0.127	0.069
Significant at FDR = 0.10	59 (29%)	1 (0.5%)
Positive	58	1
Negative	1	0

Notes: Statistical significance is determined using the Benjamini-Hochberg procedure to control the false discovery rate at 10%. Positive estimates indicate favoritism toward female-directed films; negative estimates indicate bias against.

4.3.2 Statistical Significance and Multiple Testing

Given that we test 205 critics simultaneously, standard hypothesis testing at conventional significance levels would produce substantial false positives. We therefore apply the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) to control the false discovery rate (FDR) at 10%.

The two methods yield strikingly different conclusions about which critics exhibit statistically significant bias. Under OLS, 59 critics (29%) have estimates significantly different from zero, with 58 showing significant positive bias (favoritism toward female-directed films) and one showing significant negative bias. Under HCF+DML, only one critic (0.5%) exhibits significant bias, and this estimate is positive.

Figure 3 plots OLS against HCF+DML estimates, with points colored by significance status. The two sets of estimates are positively correlated, but HCF+DML estimates are systematically lower. Critics identified as significantly biased by OLS cluster in the upper portion of the plot, where OLS estimates are large and positive. After controlling for confounding, most of these critics' estimates shrink toward zero and lose statistical significance.

4.3.3 Systematic Differences Across Critics

Figure 4 displays critic-level estimates sorted by the HCF+DML point estimate, with 95% confidence intervals shown for the HCF+DML estimates. OLS estimates

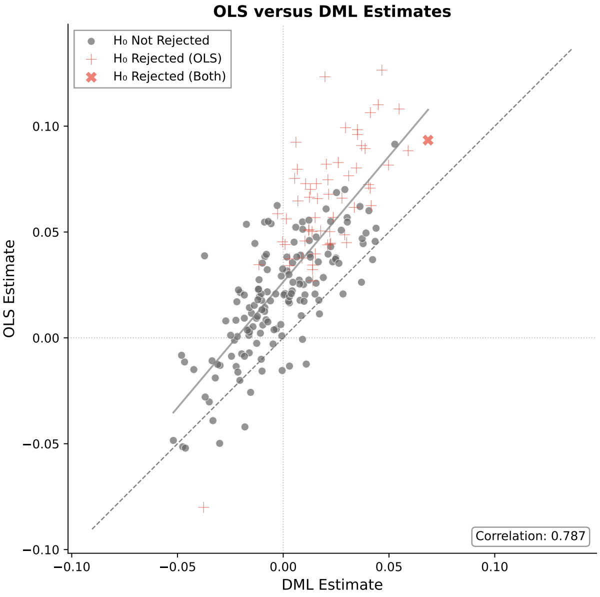


Figure 3: Comparison of OLS and HCF+DML Bias Estimates

Notes: Each point represents one critic. The horizontal axis shows the OLS estimate; the vertical axis shows the HCF+DML estimate. Points are colored by statistical significance under each method (Benjamini-Hochberg correction at FDR = 0.10). The dashed line indicates the 45-degree line where estimates would be equal. The solid line shows the fitted regression. Most critics identified as significantly biased by OLS (orange region) are no longer significant under HCF+DML.

(red crosses) are systematically higher than HCF+DML estimates (blue points) across nearly the entire distribution. This pattern indicates that the attenuation from OLS to HCF+DML is not driven by a few outliers but reflects a pervasive shift: controlling for film characteristics reduces estimated favoritism for the vast majority of critics.

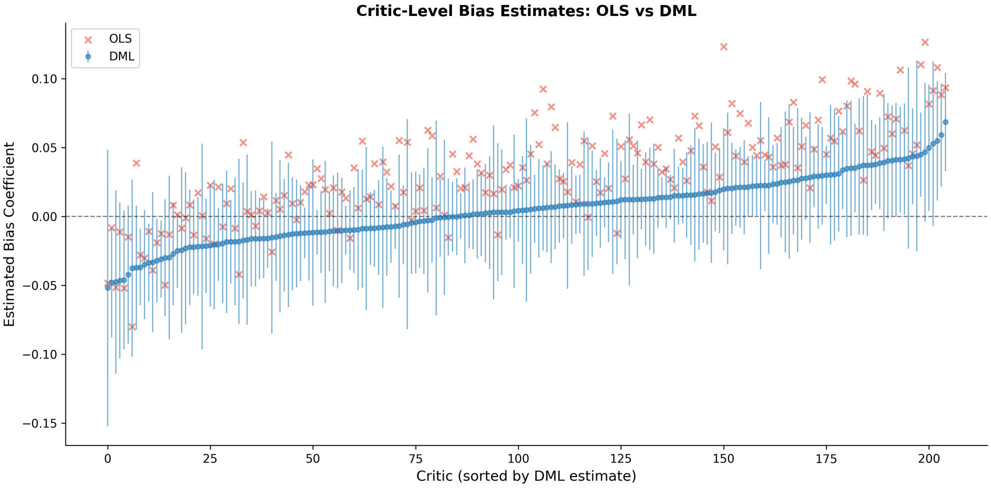


Figure 4: Critic-Level Bias Estimates: OLS vs HCF+DML

Notes: Critics are sorted by HCF+DML estimate (blue points) from most negative to most positive. Vertical bars show 95% confidence intervals for HCF+DML estimates. Red crosses show corresponding OLS estimates for the same critics. OLS estimates are consistently higher than HCF+DML estimates, indicating systematic overestimation of favoritism when film characteristics are not controlled.

4.4 Discussion

Our estimates reveal that movie critics exhibit modest favoritism toward female-directed films, but the magnitude is considerably smaller than naive approaches suggest. The mean bias estimate falls from 0.032 under OLS to 0.005 under HCF+DML, an 85% reduction. More dramatically, the number of critics classified as significantly biased drops from 59 (29%) to just one (0.5%) after controlling for confounding. These findings indicate that raw rating comparisons substantially overstate gender-based favoritism among critics.

Our sample restriction to reviews by critics who have reviewed a minimum of 30 female-directed movies may induce selection on critic preferences: critics who review many female-directed films may have favorable dispositions toward such work. If so, our estimates represent a lower bound on population-level bias. The near-zero estimates we find after controlling for preferences suggest that even among this potentially favorably-selected sample, there is little evidence of systematic bias beyond preference-based rating differences.

4.4.1 Interpreting the Attenuation

The systematic reduction in estimated favoritism has a natural interpretation rooted in industry structure. Female directors are not randomly assigned to film projects; they are disproportionately concentrated in genres and formats that critics tend to rate favorably. Data from the Sundance Institute documents substantial genre segregation: among films in the Sundance Film Festival's U.S. Dramatic Competition from 2002–2014, 92.5% of female-directed films fell into drama, comedy, or romance, compared to 69% of male-directed films (Smith et al., 2015). This concentration reflects broader patterns in which women comprise approximately 32% of directors working on independent narrative features but only 11% of directors on top-grossing studio films (Lauzen, 2025).

Crucially, the genres where female directors concentrate are precisely those that critics rate most highly. Drama films consistently receive the highest average critic scores, while action, horror, and franchise films receive systematically lower ratings (Gemser et al., 2007). A critic who prefers character-driven independent dramas to big-budget action films will rate female-directed films higher on average, not because of any gender-based favoritism, but because female directors disproportionately make the kinds of films this critic prefers.

Our methodology disentangles these channels by learning latent representations of critic preferences and film characteristics directly from the rating matrix. The "honest" design ensures that preference embeddings are estimated using only male-directed films, preventing female-directed films from contaminating the preference estimates. Once we control for the match between critic preferences and film characteristics, the estimated pro-female effect shrinks but remains positive, indicating that critics do exhibit some genuine favoritism beyond their preferences for certain film types.

4.4.2 Implications

These findings carry implications for how we interpret aggregate rating differences by demographic groups. The raw gap in ratings between female- and male-directed films reflects a mixture of (1) genuine evaluator bias, (2) differences in the characteristics of items produced by each group, and (3) sorting of evaluators to items based on preferences. Naive comparisons that ignore channels (2) and (3) risk misattributing preference-based differences to discrimination.

In the movie critic context, our results suggest that most of the apparent pro-female pattern in ratings reflects critics' preferences for the types of films female directors make rather than gender-based favoritism per se. This distinction matters for policy: if the goal is to identify and address evaluator bias, targeting critics based on raw rating gaps would largely miss the mark. Conversely, if the goal is to understand why female-directed films receive higher ratings, the answer lies primarily in genre and style rather than in critic behavior.

The methodology developed here provides a framework for making these distinctions in other bilateral evaluation settings. Whether examining hiring decisions, peer review, or performance evaluations, the core challenge remains the same: separating genuine bias from legitimate preference-based differences when treatment status correlates with item characteristics.

5 Simulation Study

5.1 Overview

This section presents a Monte Carlo simulation study to validate the HCF+DML methodology under controlled conditions where ground truth is known. We generate synthetic data that mirrors the key features of our empirical application: bilateral interactions between units and items, confounding between treatment and latent characteristics, and a sparse observation structure. The simulation demonstrates that HCF+DML recovers individual-level bias parameters with substantially lower bias than naive OLS estimation, even under deliberate misspecification of the embedding dimension.

5.2 Data Generating Process

We simulate an employer-applicant labor market setting that parallels the movie critic application. The outcome model follows:

$$Y_{ij} = \beta \cdot \cos(C_i, P_j) + T_i \cdot \theta_j + \varepsilon_{ij} \tag{11}$$

where $Y_{ij} \in [0,1]$ is the (normalized continuous) evaluation of applicant i by employer j, $C_i \in \mathbb{R}^d$ represents applicant characteristics, $P_j \in \mathbb{R}^d$ captures employer preferences, $T_i \in \{0,1\}$ indicates treatment status (gender), $\theta_j \sim N(0,1)$ is the employer-specific bias parameter of interest, and $\varepsilon_{ij} \sim N(0,\sigma^2)$ is idiosyncratic noise.

The match quality function $g(C_i, P_j) = \beta \cdot \cos(C_i, P_j)$ captures how well an applicant's characteristics align with an employer's preferences, analogous to how a film's characteristics match a critic's tastes.

Data Generating Process Summary. We generate synthetic bilateral interaction data as follows:

1. **Latent spaces:** Employer preferences $P_j \in \mathbb{R}^{10}$ are drawn uniformly from $[-1, 1]^{10}$. Applicant characteristics $C_i \in \mathbb{R}^{10}$ are drawn from a Gaussian mixture model with 100 cluster centers, each with standard deviation 0.1.
2. **Treatment and confounding:** Treatment $T_i \in \{0, 1\}$ indicates gender, with equal probability. Confounding is introduced by shifting treated (female) characteristics: $C_i|T_i = 1 \sim C_i|T_i = 0+\delta\cdot\mathbf{e}$, where $\mathbf{e} = (1, 1, 1, 0, \ldots, 0)^\top$ shifts the first three dimensions.
3. **Outcomes:** Ratings follow $Y_{ij} = \beta \cos(C_i, P_j) + T_i \cdot \theta_j + \varepsilon_{ij}$, where $\cos(\cdot, \cdot)$ denotes cosine similarity, $\theta_j \sim N(0, 1)$ are employer-specific bias parameters, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ is idiosyncratic noise.
4. **Observation pattern:** Each dyad (i, j) is observed independently with probability $\lambda = 0.2$, yielding approximately 20% matrix density.
5. **Estimation:** We deliberately misspecify the embedding dimension ($d_{\text{model}} = 8$ vs. true $d = 10$) to test robustness.

5.2.1 Confounding Structure

The identification challenge arises because treatment correlates with characteristics. We induce confounding by shifting treated applicants' characteristics:

$$C_i \mid T_i = 1 \sim C_i \mid T_i = 0 + \delta \cdot \mathbf{e} \tag{12}$$

where $\delta = 0.5$ controls confounding strength and $\mathbf{e}$ shifts the first three dimensions. This models scenarios where female-directed films systematically differ in style, genre, or production characteristics from male-directed films. Under this confounding, naive OLS conflates true bias θ_j with preference-based differences, while HCF+DML aims to separate these effects.

5.2.2 Parameter Values

Table 3 summarizes the parameters used in the simulation. The embedding dimension is deliberately set to 8 while the true latent dimension is 10, testing robustness to model misspecification.

Figure 5 visualizes the latent space structure underlying our simulation. Panel (A) plots the first two dimensions of applicant characteristics. Male applicants (blue) and female applicants (orange) are drawn from the same Gaussian mixture model, but we introduce confounding by shifting female characteristics by $\delta = 0.5$ in the

Table 3: Simulation Parameters

Parameter	Value	Description
J (employers)	100	Number of units to estimate
N_{male}	5,000	Control group size
N_{female}	5,000	Treatment group size
d (true dimension)	10	Latent space dimensionality
d_{model}	8	Embedding dimension (deliberately misspecified)
n_{clusters}	100	Characteristic cluster centers
λ (observation rate)	0.2	Application probability
β (match coefficient)	5.0	Match quality weight
σ (noise)	1.0	Outcome noise
δ (confounding)	0.5	Treatment-characteristic shift

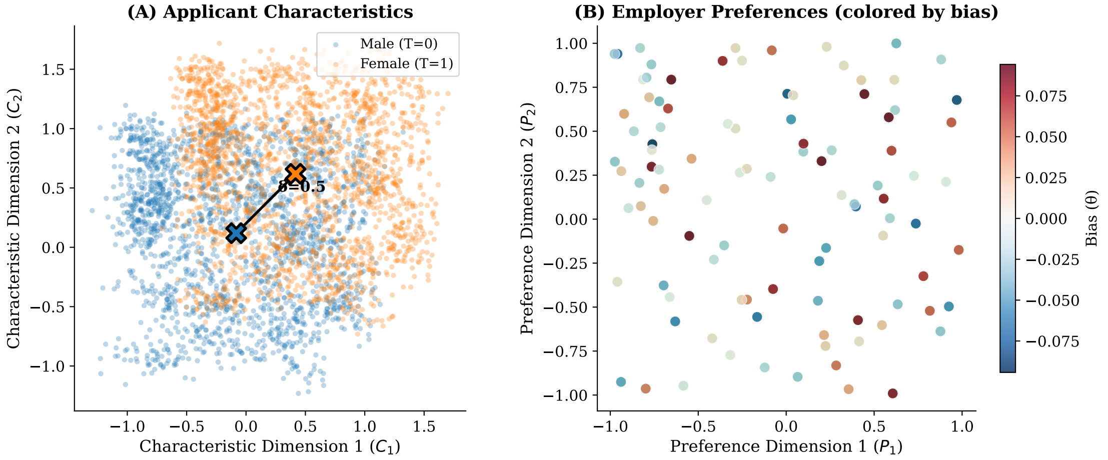


Figure 5: Simulated latent spaces illustrating the confounding structure. Panel (A) shows the first two dimensions of applicant characteristics. The arrow indicates the confounding shift $\delta = 0.5$ applied to female characteristics, creating systematic differences between treatment groups. Panel (B) displays employer preferences in the same latent space, with each point representing one employer colored by their true bias parameter θ_j. Bias parameters are drawn independently from preferences.

first three dimensions. This shift, indicated by the arrow connecting group means, creates systematic differences between treatment groups.

Panel (B) displays the 100 simulated employers in the preference space, with color indicating their true bias parameter θ_j. By construction, bias is independent of preferences: employers with similar preferences may have very different biases, and vice versa. However, because applicant characteristics differ systematically by gender (Panel A), employers whose preferences align with male-typical characteristics will rate male applicants higher on average, even absent any bias. Naive OLS estimation cannot distinguish this preference-based channel from true discriminatory bias, whereas HCF+DML controls for the confounding path by including learned embeddings of both preferences and characteristics.

5.3 Estimation

For each employer j, we estimate the bias parameter θ_j using two approaches.

OLS (Naive Baseline). For each employer, we regress outcomes on treatment status without controlling for preferences or characteristics:

$$Y_{ij} = \alpha_j + \beta_j T_i + e_{ij} \tag{13}$$

The OLS estimate $\hat{\beta}_j$ is biased under confounding because it captures both true bias and preference-based differences.

HCF+DML. We estimate employer bias following Algorithm 1: We first train the Honest Collaborative Filtering model to extract embeddings $\hat{P}_0$ (employer preferences trained on control items only) and $\hat{C}^{(k)}$ (cross-fitted item characteristics). We then estimate employer-specific effects using the partial linear regression model with 5-fold DML cross-fitting.

5.4 Results

Figure 6 displays estimated versus true bias parameters for both methods. Under confounding, OLS estimates exhibit systematic bias: employers with certain preference profiles (those who prefer characteristics more common in the control group) appear more biased than they truly are. HCF+DML substantially reduces this bias by controlling for the confounding path through the learned embeddings.

Table 4 shows that HCF+DML more accurately recovers evaluator-level heterogeneity in discrimination than OLS. The correlation between estimated and true θ is 0.95 under HCF+DML versus 0.81 under OLS, indicating tighter alignment

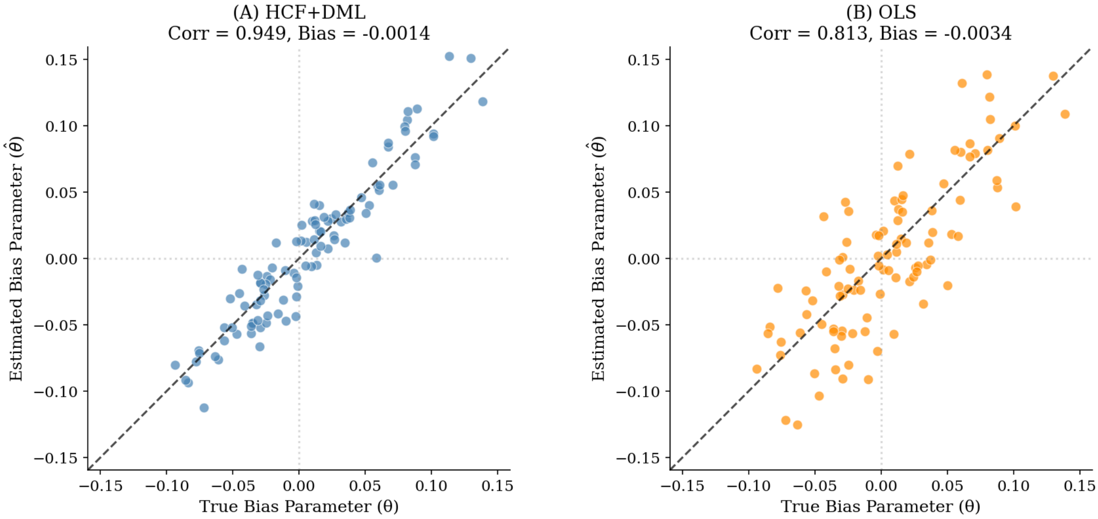


Figure 6: Scatter plots of estimated vs true theta, DML and OLS

with the true cross-sectional pattern of bias. HCF+DML also reduces average bias by about 67%. We also report *relative bias*, defined as the mean estimation bias divided by the standard deviation of the true bias parameters, SD(θ) (in our simulations, SD(θ) $\approx$ 0.05). Under this scaling, HCF+DML exhibits a relative bias of 3% of a standard deviation, whereas naive OLS exhibits a relative bias of 7% of a standard deviation. Consistent with these gains, HCF+DML substantially lowers overall estimation error, with RMSE 0.018 compared to 0.036 for OLS, implying roughly a 50% reduction in RMSE. Together, the higher correlation, lower absolute bias and relative, and lower RMSE suggest that HCF+DML provides a more reliable measure of individual-level discrimination in this simulation setting.

Table 4: Estimation Performance

Metric	HCF+DML	OLS
Correlation with true θ	0.95	0.81
Mean Bias	-0.001	-0.003
Relative Bias	-0.03	-0.07
RMSE	0.018	0.036

Notes: Relative bias is computed as Mean bias/SD(θ_{true}). In this simulation, SD(θ_{true}) = 0.051.

The simulation deliberately misspecifies the embedding dimension (d=8 vs true d=10). Despite this misspecification, HCF+DML maintains its advantage over OLS in terms of bias, demonstrating practical robustness. This is important for applications where the true latent dimensionality is unknown.

5.5 Sensitivity to Confounding Strength

We examine how estimation performance varies with confounding strength by simulating three scenarios with different degrees of correlation between treatment and item characteristics. Recall that confounding arises when the treatment variable (e.g., director gender) is correlated with characteristics that also affect outcomes through evaluator preferences. This correlation generates omitted variable bias in naive OLS estimation, which does not control for characteristics.

We implement confounding by shifting female item characteristics relative to male characteristics by δ along the first three latent dimensions:

1. **No Confounding** ($\delta = 0$): Item characteristics are independent of treatment. Male and female characteristic distributions fully overlap, and OLS faces no omitted variable bias.
2. **Moderate Confounding** ($\delta = 0.3$): Female characteristics are shifted moderately, creating partial separation between treatment groups in the latent space.
3. **Strong Confounding** ($\delta = 0.6$): Female characteristics are shifted substantially, generating pronounced differences between treatment groups.

For the no-confounding baseline, we use the correct embedding dimension ($d = 10$) to establish best-case performance. For the confounded scenarios, we deliberately misspecify the embedding dimension ($d = 8$) to test robustness under realistic conditions where the true latent dimensionality is unknown.

Table 5: Estimation Performance Across Confounding Levels

Scenario	Method	Correlation	Bias	Rel. Bias	RMSE
No Confounding ($\delta = 0$)	HCF+DML	0.963	-0.0025	-0.05	0.017
	OLS	0.997	-0.0006	-0.012	0.004
Moderate ($\delta = 0.3$)	HCF+DML	0.950	-0.0016	-0.032	0.017
	OLS	0.916	-0.0022	-0.044	0.022
Strong ($\delta = 0.6$)	HCF+DML	0.939	-0.0013	-0.026	0.020
	OLS	0.764	-0.0039	-0.078	0.042

Notes: Confounding is introduced by shifting female item characteristics by δ along three latent dimensions. Embedding dimension is $d = 10$ for no confounding and $d = 8$ (misspecified) for confounded scenarios.

Table 5 reports estimation performance across scenarios. Several patterns emerge. First, when confounding is absent, OLS performs exceptionally well, achieving near-perfect correlation (0.997) with minimal RMSE. HCF+DML performs slightly worse in this setting due to the additional variance introduced by the honest estimation procedure, which splits the sample for embedding estimation. This confirms that HCF+DML incurs a modest efficiency cost when its key identifying assumption of treatment-characteristic correlation is not satisfied.

Second, as confounding increases, OLS performance deteriorates substantially while HCF + DML remains stable. Under strong confounding, OLS correlation with true bias parameters drops to 0.764 and RMSE more than doubles relative to the no-confounding baseline. In contrast, HCF+DML maintains correlation above 0.93 and stable RMSE across all scenarios. The RMSE ratio between methods grows from 0.24 (no confounding) to 0.48 (strong confounding), indicating that the relative advantage of HCF+DML increases with confounding severity.

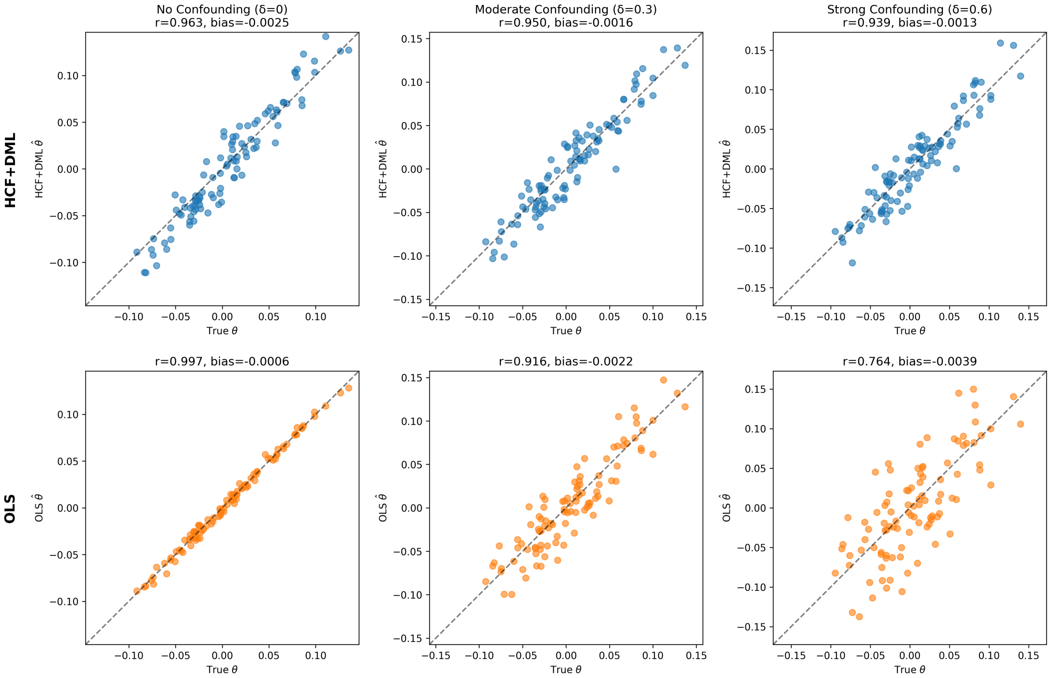


Figure 7: Estimated versus true bias parameters across confounding levels. Each column represents a confounding scenario (no, moderate, strong). Top row: HCF+DML estimates. Bottom row: OLS estimates. Under strong confounding, OLS estimates exhibit substantial dispersion around the 45-degree line, while HCF+DML estimates remain tightly clustered.

Figure 7 visualizes these patterns. Under no confounding, both methods produce estimates tightly clustered around the 45-degree line. As confounding increases, OLS estimates become increasingly dispersed, particularly in the strong confound-

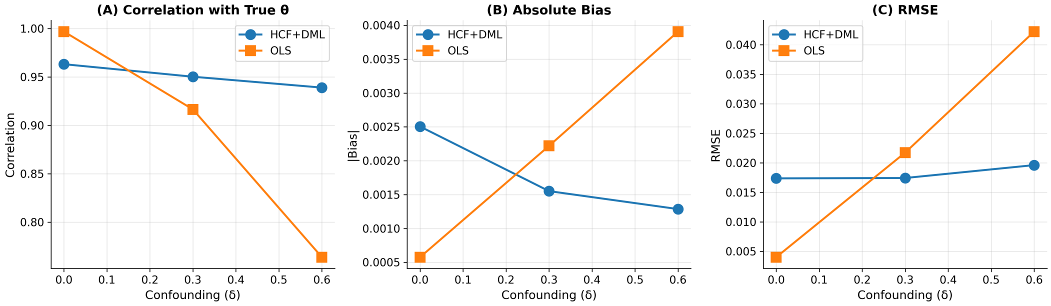


Figure 8: Performance metrics as a function of confounding strength. Left panel: correlation with true parameters. Right panel: RMSE. HCF+DML (blue) maintains stable performance across confounding levels, while OLS (orange) degrades substantially as confounding increases.

ing case where substantial deviations from true values are evident. HCF+DML estimates remain well-calibrated across all scenarios.

Figure 8 summarizes these trends. The correlation between estimated and true bias parameters remains above 0.93 for HCF+DML regardless of confounding strength, while OLS correlation declines from 0.997 to 0.764. Similarly, HCF+DML RMSE remains stable around 0.02, while OLS RMSE increases from 0.004 to 0.042.

These results demonstrate that HCF+DML is robust to confounding and maintains reliable performance even under embedding dimension misspecification. When confounding is absent, practitioners may prefer OLS for its simplicity and efficiency. However, since the presence and magnitude of confounding is typically unknown in empirical applications, HCF+DML provides a robust alternative that guards against potentially severe bias.

Appendix A extends this analysis to alternative confounding structures, confirming that these findings generalize beyond the shift design studied here.

5.6 Sensitivity to Selection Mechanism

Our baseline simulation assumes random observation patterns. In practice, which evaluator-item pairs we observe may depend on characteristics that also affect outcomes, creating missing-not-at-random (MNAR) selection. We examine three increasingly complex selection mechanisms to assess robustness.

This subsection is designed to isolate the role of selection. We therefore hold the underlying data-generating process fixed across cases, including the latent preference and characteristic embeddings (P, C), the bias parameters θ, and the treatment assignment rule. In particular, the level of confounding induced by the distribution of C across treatment groups is identical across the three selection cases and we only vary observation mechanism.

Let $S_{ij} \in \{0, 1\}$ indicate whether evaluator j's rating of item i is observed. We model selection as:

$$\Pr(S_{ij} = 1) = \sigma(\alpha_0 + \alpha_1 \cdot \text{match}_{ij} + \alpha_2 \cdot T_i + \alpha_3 \cdot U_{ij}) \tag{14}$$

where $\sigma(\cdot)$ is the logistic function, $\text{match}_{ij} = \beta \cos(C_i, P_j)$ is the match quality, T_i is the treatment indicator, and U_{ij} is an unobservable that may also affect outcomes. We calibrate α_0 to achieve approximately 10% observed entries. Table 6 describes the three cases.

Table 6: MNAR Selection Cases

Case	Description	α_1	α_2	α_3
1	Match-only	0.5	0	0
2	Match + Treatment	0.5	0.3	0
3	Match + Treatment + Unobservables	0.5	0.3	0.4

Case 1 represents selection based purely on match quality: evaluator-item pairs with higher latent compatibility are more likely to be observed. This captures settings where interaction is driven by fit or interest, such as firms interviewing applicants who appear to match a job's requirements, or critics reviewing films that align with their established tastes (e.g., genre or style).

Case 2 adds differential selection by treatment status, allowing treated items to be observed at different rates even after conditioning on match quality. This corresponds to environments where protected attributes influence exposure or assignment, such as gendered sorting into job ladders, editorial decisions that differentially allocate reviews, or platforms that route certain items to evaluators at systematically different rates.

Case 3 introduces an unobservable U_{ij} that affects both selection and outcomes (Y_{ij} includes γU_{ij} with $\gamma = 0.5$), creating the most challenging identification environment. This reflects cases where observation depends on factors only partially captured by match quality or observed characteristics, such as time-varying shocks to demand, publicity, or access that simultaneously increase the likelihood of being reviewed and shift ratings.

Table 7 reports estimation performance across selection cases. HCF+DML consistently outperforms OLS, with the advantage most pronounced under Case 3 where selection depends on outcome-relevant unobservables. Because the underlying (P, C) structure is held fixed, these differences are attributable to the selection mechanism rather than changes in confounding strength. The bias reduction ranges from 18% (Case 1) to 49% (Case 3), and HCF+DML achieves higher correlation with true bias parameters in all cases.

Table 7: Estimation Performance Under MNAR Selection

Case	Method	Correlation	Bias	Rel. Bias	RMSE
1	HCF+DML	0.817	-0.0047	-0.09	0.0236
	OLS	0.773	-0.0057	-0.11	0.0262
2	HCF+DML	0.801	-0.0064	-0.12	0.0247
	OLS	0.764	-0.0087	-0.17	0.0270
3	HCF+DML	0.791	-0.0039	-0.08	0.0250
	OLS	0.703	-0.0076	-0.15	0.0296

These results suggest that HCF+DML is robust to realistic forms of MNAR selection. By controlling for match quality through learned embeddings, the method partially addresses selection that operates through the same channel. The improvement under Case 3 indicates that even when unobservables drive both selection and outcomes, the embedding-based controls provide meaningful bias reduction relative to naive estimation.

5.7 Discussion

The simulation study validates that HCF+DML successfully recovers individual-level bias parameters under confounding between treatment and latent characteristics. Several findings emerge.

First, HCF+DML substantially reduces estimation error compared to naive OLS when confounding is present. Under moderate confounding, HCF+DML achieves 17% lower RMSE; under strong confounding, this advantage grows to 52%. The method maintains stable performance across confounding levels while OLS deteriorates markedly, confirming that the embedding-based controls effectively address omitted variable bias.

Second, HCF+DML preserves ranking accuracy. Correlation between estimated and true bias parameters remains consistently high across all scenarios, indicating that the method reliably identifies which evaluators exhibit the strongest positive or negative bias. This is particularly relevant for applications where the goal is to flag outliers or compare relative bias across units rather than estimate exact magnitudes.

Third, the method is robust to realistic complications. Performance remains strong under deliberate misspecification of the embedding dimension, alternative confounding structures (Appendix A), and MNAR selection mechanisms that depend on match quality, treatment status, or unobservables. This robustness is important for empirical applications where the true data-generating process is unknown.

Fourth, when confounding is absent, OLS outperforms HCF+DML in terms of efficiency. This reflects the cost of the honest estimation design, which restricts preference embeddings to control observations only. Since the presence of confounding is typically unknown in practice, HCF+DML provides insurance against potentially severe bias at the cost of modest efficiency loss when confounding happens to be negligible.

Two methodological considerations merit attention. For inference, we apply the Benjamini-Hochberg procedure to control the false discovery rate when testing multiple evaluators simultaneously. This is essential given the large number of hypothesis tests (one per evaluator) and the exploratory nature of identifying biased individuals. Additionally, standard errors from the DML procedure treat the learned embeddings as fixed, ignoring estimation uncertainty in the first stage. This "generated regressor" problem can lead to understated standard errors and inflated rejection rates. In our empirical application, we address this through bootstrap procedures that re-estimate embeddings in each replication, providing valid inference that accounts for the full estimation pipeline.

6 Conclusion

This paper develops a methodology for estimating individual-level bias in bilateral evaluation settings where treatment status correlates with unobserved characteristics. We combine Honest Collaborative Filtering, which extracts latent representations of evaluator preferences and item characteristics from observed ratings, with Double Machine Learning to estimate evaluator-specific bias parameters while controlling for these learned embeddings. The "honest" design ensures that preference embeddings are estimated using only control-group items, preventing treatment effects from contaminating the estimates.

Simulations demonstrate that HCF+DML substantially outperforms naive OLS under confounding, reducing RMSE by up to 50% while maintaining high correlation with true parameters. The method proves robust to embedding misspecification, alternative confounding structures, and non-random selection.

Applied to nearly 150,000 film reviews, the methodology overturns naive conclusions about critic bias. Raw comparisons suggest 29% of critics exhibit significant favoritism toward female-directed films; after controlling for preference-characteristic alignment, this figure drops below 1%. The apparent pro-female pattern largely reflects critics' preferences for genres where female directors concentrate, not gender-based favoritism per se.

The framework applies broadly to hiring, peer review, and other evaluation settings where treatment groups differ in characteristics that evaluators legitimately value. By extracting information from bilateral rating patterns, the approach separates genuine bias from preference-based differences, a distinction that naive methods cannot make.

Appendix A Sensitivity to Confounding Structure

In Section 5, we study a stylized shift design where female characteristics are systematically shifted in a subset of dimensions. This design captures realistic scenarios where female-directed films may systematically differ from male-directed films along observable dimensions (e.g., lower budgets, different genre distributions, varying production scales). However, confounding in practice may arise through more complex mechanisms.

In this appendix, we examine HCF+DML performance under an alternative confounding structure where male and female characteristics originate from *different cluster centers* in the latent space. This design reflects the possibility that male and female directors work in distinct "niches" or stylistic traditions, creating heterogeneous, non-directional confounding patterns rather than a uniform shift.

Data Generating Process

We modify the main simulation DGP as follows. Rather than drawing male and female characteristics from the same Gaussian mixture with a location shift, we generate separate cluster centers for each group:

$$\mu_k^{(M)} \sim \text{Uniform}(-1,1)^d, \quad k = 1, \ldots, N_C^{(M)} \tag{15}$$

$$\mu_k^{(F)} \sim \text{Uniform}(-1,1)^d, \quad k = 1, \ldots, N_C^{(F)} \tag{16}$$

$$C_i | T_i = 0 \sim \sum_{k=1}^{N_C^{(M)}} \frac{1}{N_C^{(M)}} \mathcal{N}(\mu_k^{(M)}, \sigma^2 I_d) \tag{17}$$

$$C_i | T_i = 1 \sim \sum_{k=1}^{N_C^{(F)}} \frac{1}{N_C^{(F)}} \mathcal{N}(\mu_k^{(F)}, \sigma^2 I_d) \tag{18}$$

where $N_C^{(M)} = N_C^{(F)} = 50$ cluster centers are drawn independently for male and female characteristics. The random positioning of cluster centers creates regions where male and female characteristics overlap substantially, alongside regions dominated by one group. This heterogeneity makes identification more challenging than the uniform shift design.

Table 8 summarizes the simulation parameters.

Table 8: Simulation Parameters: Complex Confounding Structure

Parameter	Value
Employers (J)	100
Male items	5,000
Female items	5,000
Latent dimension (d)	10
Male characteristic clusters	50
Female characteristic clusters	50
Cluster std (σ)	0.15
Observation rate (λ)	0.2
Match coefficient (β)	5.0
Bias std (σ_θ)	1.0
Model embedding dimension	8

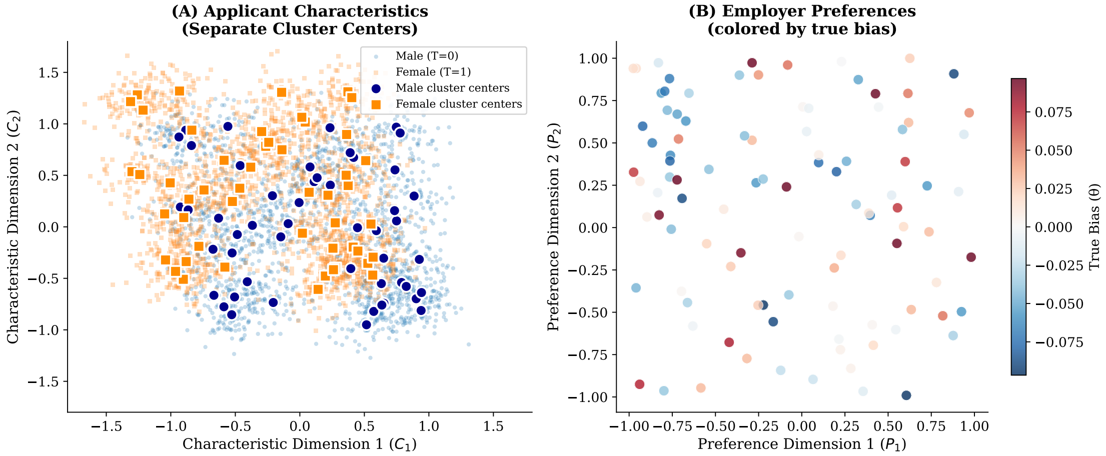


Figure 9: Latent spaces under complex confounding (separate cluster centers). Panel (A) shows applicant characteristics with male (blue) and female (orange) drawn from different Gaussian mixture components. Large markers indicate cluster centers. Panel (B) displays employer preferences colored by true bias θ_j.

Confounding Structure Visualization

Figure 9 visualizes the latent space structure under this alternative DGP. Panel (A) displays the first two dimensions of applicant characteristics, with male and female observations shown alongside their respective cluster centers. Unlike the shift design, confounding here arises from the spatial separation of cluster centers rather than a systematic directional shift. Some regions contain primarily male characteristics, others primarily female, with substantial overlap where cluster centers happen to be proximate. Panel (B) shows employer preferences colored by their true bias parameter θ_j. As in the main simulation, bias is independent of preferences by construction.

Table 9 reports estimation performance under the complex confounding structure. HCF+DML maintains its advantage over naive OLS estimation, achieving higher correlation with true bias parameters and substantially lower systematic bias. The method successfully recovers individual-level bias even when confounding patterns are heterogeneous across the characteristic space.

Table 9: Estimation Performance: Complex Confounding Structure

Method	Correlation	Mean Bias	Rel. Bias	RMSE	MAE
HCF+DML	0.9272	0.0004	0.007	0.0217	0.0177
OLS	0.8310	0.0055	0.102	0.0330	0.0264

Notes: Relative bias is computed as Mean bias/SD(θ_{true}). In this simulation, SD(θ_{true}) = 0.054.

Figure 10 compares estimated versus true bias parameters for both methods. The DML estimates cluster more tightly around the 45-degree line, indicating better recovery of individual-level heterogeneity. OLS estimates exhibit greater dispersion and systematic deviation from ground truth due to uncontrolled confounding.

These results demonstrate that HCF+DML is robust to the functional form of confounding. Under the more challenging separate-clusters design, where confounding is heterogeneous and non-directional, the method continues to outperform naive OLS estimation in recovering true individual-level bias parameters. This robustness is important for empirical applications where the precise nature of confounding between treatment and latent characteristics is unknown.

The key insight is that collaborative filtering embeddings capture the relevant structure of the characteristic space regardless of whether confounding manifests as a uniform shift or through more complex distributional differences. By learning embeddings that predict outcomes, the method implicitly controls for whatever

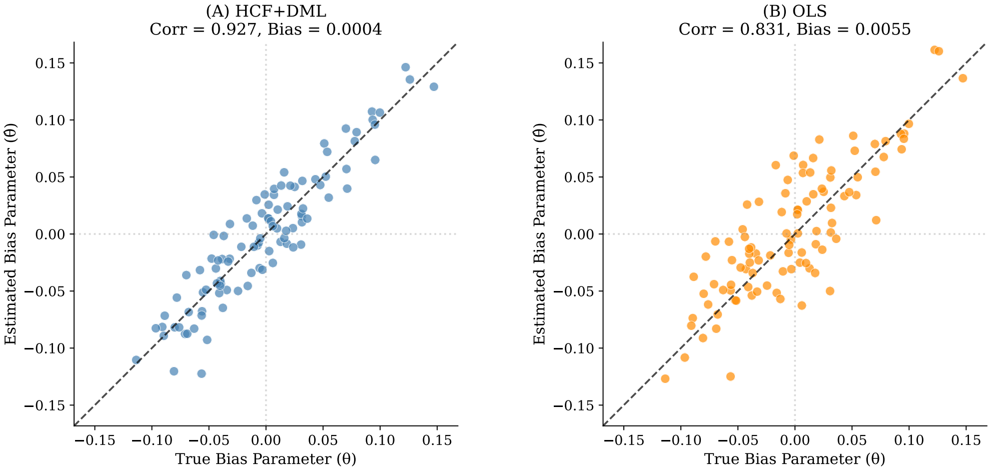


Figure 10: Estimated versus true bias parameters under complex confounding. Panel (A): HCF+DML estimates. Panel (B): OLS estimates. Dashed line indicates perfect recovery.

characteristic patterns drive rating variation, enabling valid causal inference about individual-level bias.

References

Arnold, D., W. Dobbie, and C. S. Yang (2018). Racial bias in bail decisions. *Quarterly Journal of Economics 133*(4), 1885–1932.

Arrow, K. J. (1973). The theory of discrimination. In O. Ashenfelter and A. Rees (Eds.), *Discrimination in Labor Markets*, pp. 3–33. Princeton, NJ: Princeton University Press.

Athey, S. and G. Imbens (2016a). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S. and G. Imbens (2016b). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Bach, P., V. Chernozhukov, S. Klaassen, M. S. Kurz, and M. Spindler (2024). DoubleML – double machine learning in python. BSD-3-Clause License. Documentation: https://docs.doubleml.org/stable/index.html.

Bach, P., V. Chernozhukov, M. S. Kurz, and M. Spindler (2022). DoubleML – An object-oriented implementation of double machine learning in Python. *Journal of Machine Learning Research 23*(53), 1–6.

Becker, G. S. (1957). *The Economics of Discrimination*. Chicago: University of Chicago Press.

Benjamini, Y. and Y. Hochberg (1995, January). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological) 57*(1), 289–300.

Bertrand, M. and E. Duflo (2017). Field experiments on discrimination. *Handbook of economic field experiments 1*, 309–393.

Bertrand, M. and S. Mullainathan (2004a). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review 94*(4), 991–1013.

Bertrand, M. and S. Mullainathan (2004b). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review 94*(4), 991–1013.

Candes, E. J. and B. Recht (2008). Exact low-rank matrix completion via convex optimization. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 806–812. IEEE.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018a, February). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018b). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal 21*(1), C1–C68.

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

Gao, C., Y. Zheng, W. Wang, F. Feng, X. He, and Y. Li (2024). Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems 42*(4), 1–32.

Gemser, G., M. van Oostrum, and M. A. A. M. Leenders (2007). The impact of film reviews on the box office performance of art house versus mainstream motion pictures. *Journal of Cultural Economics 31*(1), 43–63.

Griffith, A. and S. Peng (2023). Identification of network structure in the presence of latent, unobserved factors: A new result using Turán's theorem. Revise and Resubmit, Quantitative Economics.

Hernán, M. A. and J. M. Robins (2020). *Causal Inference: What If.* Chapman & Hall/CRC. Available at: https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Kline, P., E. K. Rose, and C. R. Walters (2022a). Systemic discrimination among large us employers. *The Quarterly Journal of Economics 137*(4), 1963–2036.

Kline, P., E. K. Rose, and C. R. Walters (2022b). Systemic discrimination among large U.S. employers. *Quarterly Journal of Economics 137*(4), 1963–2036.

Kline, P., E. K. Rose, and C. R. Walters (2023). A discrimination report card. *arXiv preprint arXiv:2306.13005*.

Kline, P. and C. Walters (2021). Reasonable doubt: Experimental detection of job-level employment discrimination. *Econometrica 89*(2), 765–792.

Koltchinskii, V., K. Lounici, and A. B. Tsybakov (2011). Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion.

Koren, Y., R. Bell, and C. Volinsky (2009a, August). Matrix Factorization Techniques for Recommender Systems. *Computer 42*(8), 30–37.

Koren, Y., R. Bell, and C. Volinsky (2009b). Matrix factorization techniques for recommender systems. *Computer 42*(8), 30–37.

Lauzen, M. M. (2025). The celluloid ceiling: Behind-the-scenes employment of women on the top grossing u.s. films. Technical report, Center for the Study of Women in Television and Film, San Diego State University.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *American Economic Review 62*(4), 659–661.

Smith, S. L., M. Choueiti, and K. Pieper (2015). Exploring the barriers and opportunities for independent women filmmakers: Phase iii. Technical report, Sundance Institute and Women In Film Los Angeles. Research conducted by USC Annenberg School for Communication and Journalism.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.