

An Agent-Based Simulation of Fairness Interventions in Hiring

Mattia Cerrato¹, Herbert Dawid², Stefan Kramer¹

¹Johannes Gutenberg University Mainz, Mainz, Germany

²Bielefeld University, Bielefeld, Germany

mcerrato@uni-mainz.de, hdawid@uni-bielefeld.de, kramer@informatik.uni-mainz.de

Abstract

Fairness interventions in hiring are often evaluated on static datasets, even though hiring decisions shape the future data on which firms train their models. We introduce a hiring-and-promotion simulation that makes this feedback explicit. In the model, firms observe biased proxy measurements, receive labels only for selected workers, and retrain on a cumulative datastore generated by their own decisions. This setup supports comparison among status-quo learning, observational debiasing, demographic-parity-style top-k selection, and repair of the measurement channel. Our results suggest that interventions with similar surface outcomes can differ substantially in how they alter the learned signal and the composition of the training data. The framework is intended as a testbed for studying fairness, dataset formation, and recoverability of bias mechanisms in dynamic organizational settings.

1 Introduction

A firm using a machine-learning model to screen job candidates faces a circular problem. The model is trained on the outcomes of workers the firm previously hired, but who the firm hired is itself the output of an earlier model, or human process, applied to noisy and possibly biased proxies of those workers' latent ability. A bias in any one of those proxies propagates into who gets hired, and from there into the labels the firm gets to learn from. The dataset on which the firm assesses its fairness in any given period is therefore not a fixed input but the cumulative record of its own past policies, with measurement biases compounded through retraining.

Three features of this setting separate it from a static framing. First, the signals firms use at hiring time are proxies, not direct observations of the latent traits they are trying to measure. A candidate's underlying ability is unobservable; the firm sees standardised test scores, CV features, and interview ratings. We will refer to the latent traits as belonging to the *construct space* and to the proxies as belonging to the *observation space*, following Friedler, Scheidegger, and Venkatasubramanian (2016, 2021). The literature on the cognitive-ability tests used as the first stage of personnel selection has documented systematic gender bias in this construct-to-observation map (Striebing et al. 2025),

large enough that the same latent skill produces a noticeably lower observed score for women. Second, the outcome labels firms learn from are *selectively observed*: a firm only sees how productive its hires turned out to be, not how productive its rejected candidates would have been (Lakkaraju et al. 2017). The training set is therefore a sample from the firm's own hiring policy, not from the candidate population. *Third*, the firm retrains its models on this cumulative training set as it grows, so that the policy applied at time t shapes the data on which the policy at time $t + 1$ is fit (Liu et al. 2018; Ensign et al. 2018).

Putting these three features together yields a distinction that is invisible when the dataset is taken as given but matters once the dynamics are analyzed: the distinction we analyze is between fairness interventions that act on the firm's *decisions* and fairness interventions that act on the firm's *measurements*. Decision-side interventions adjust who is hired, for instance by enforcing parity in hiring rates. Measurement-side interventions adjust the signal the firm sees, for instance by correcting a known bias in the pre-screening test before any scoring or training. In a one-shot setting the two can produce the same hire decisions. In a retraining setting they pull apart: a decision-side intervention can produce hire-rate parity round after round while the firm's prediction model continues to be fit on biased features, and therefore continues to under-predict the disadvantaged group even when its hiring outcomes look balanced.

We study this through an agent-based simulation of hiring, promotion, and learning. Candidates have latent worker and leadership skills, observed by the firm only through noisy and (in the case of the pre-screening test) biased proxies. Each round the firm scores candidates with a learned model, hires the top fraction, observes productivity outcomes only for those it hired, and retrains. Promotion introduces a second biased channel, through a group-conditional assessment of the leadership skills of workers. We compare four intervention families that act on different parts of the pipeline. A *status-quo* policy trains on the biased features as observed. An *oracle* policy knows and corrects the hiring-stage measurement bias before training and scoring; this is not a deployable policy, since the bias parameter is structurally unobserved in any real setting, but it serves as a useful upper bound on what measurement-channel repair could achieve. An *observational debiasing* policy constrains the

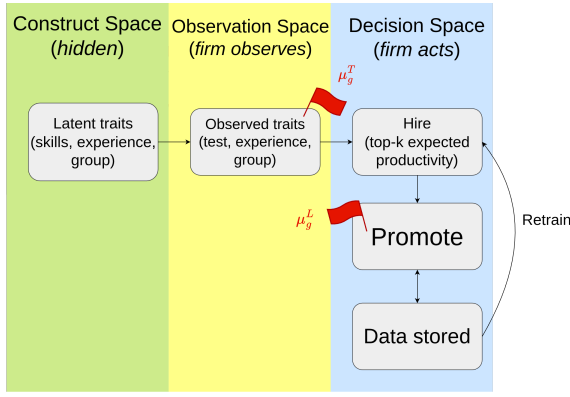


Figure 1: A schematic of our simulation. We model ML-based hiring as a biased process regulated by two factors: μ^T and μ^L , which are both group-dependent distortions of the maps between spaces construct, observation and decision space (Friedler, Scheidegger, and Venkatasubramanian 2021).

learned model to satisfy a group-fairness criterion without modifying the features it sees. A *demographic-parity top-k* allocation rule enforces equal group representation in the hired set without modifying the learned score.

The paper is organised around three empirical questions. First, when and how does the simulation produce structured disparities under the status-quo policy? Second, how do the four intervention families compare, both on surface fairness metrics such as hire-rate parity and group representation, and on the bias properties of the firm’s own prediction model? Third, can the firm recover the structure and magnitude of its own measurement bias from the data it accumulates? We address each question both under a broad parameter sweep that maps the simulation’s behaviour and under a calibrated scenario whose parameters are anchored to magnitudes from the cognitive-testing and within-firm promotion literatures (Striebing et al. 2025; Xu 2024).

We do not claim a complete account of how hiring discrimination arises in any particular labour market. The narrower claim, which the simulation lets us make precise, is that biased pre-screening, selective labels, and cumulative retraining together suffice to produce the kinds of disparities that fairness interventions are deployed against, and that the choice of intervention matters at the level of the prediction model the firm trains, not only at the level of its hire-rate outcomes. We take this as a reason to evaluate fairness in this setting by inspecting the prediction model and the dataset that produced it, and not only the visible hiring outcomes themselves.

To the best of our knowledge, agent-based models (ABMs) are not yet well-established in fairness research. We make the case for them in Section 2, after locating the present work in the literature on fair selection, construct-space and causal fairness, selective labels, and long-term fairness. Section 3 defines the simulation, the four interventions, and the calibration procedure. Section 4 reports the three empirical questions. Section 5 discusses what the cali-

bration achieves, what it does not, and where the work stands relative to the validation standards of the agent-based modelling literature.

2 Related Work and Motivation

2.1 Long-term fairness in machine learning

The recognition that static fairness criteria fail to capture how disparities evolve under repeated decisions and over time has produced a substantial literature on long-term fairness. Without pretense of completeness, we summarize here a few contributions which are relevant and closest to our present proposal. Liu et al. (2018) show that imposing demographic parity on a lending model can degrade the outcomes of the very group it was intended to protect, once feedback into the next round of decisions is taken into account. Compared to our proposal here, their model accounts for a singular timestep: one round of decisions and one round of population response, with no iteration of the dynamics. Ensign et al. (2018) document similar feedback loops in predictive policing, in which an initial allocation of patrols produces the observations that justify further allocation to the same area. Hu and Chen (2018) introduce a two-stage labor-market model and show that individual fairness criteria, applied at each stage, do not control group-level inequalities over time. D’Amour et al. (2020) take the broader methodological position that simulation is the natural evaluation tool in such settings, and accompany the argument with `ml-fairness-gym`, an open-source package built on the reinforcement learning (RL) conventions and on the OpenAI Gym APIs.

A more complete survey for the this line of work is given by Gohar et al. (2025). The authors identify four modeling frameworks that have come to structure the field: Pólya urn processes, one-step feedback models, reinforcement learning, and causal modeling. A brief subsequent discussion mentions game-theoretic and best-response models, social-network settings, online learning, and algorithmic recourse as additional but less established lines of inquiry. We note that agent-based simulation studies are not represented in either category. We take this absence as evidence that the methodology has not yet been recognized by the fairness community as a candidate framework for long-term fairness research in ML, a point we revisit below when presenting our argument for ABMs in this space (Section 2.3).

2.2 Agent-based modeling

Outside of fairness, ABM has been used productively for several decades to study the emergence of aggregate social patterns from individual behavior, including discrimination phenomena. The point of reference for any discussion of segregation is Schelling (1971), whose model showed that residential segregation can arise from individual relocation decisions without any aggregate discriminatory intent. Thus, ABMs have some explanatory potential in better understanding phenomena which are relevant to algorithmic fairness research. Epstein (2006) articulates the underlying view of ABMs as a potential explanation as a methodological commitment: a social phenomenon is explained when it

has been generated by the interaction of individually specified agents. In his formulation, “if you didn’t grow it, you didn’t explain its emergence.” (Epstein 2006, p.8). Bonabeau (2002), writing for a broader scientific audience, characterizes the conditions under which the methodology offers advantages over closed-form alternatives. He identifies three such conditions in particular: i) the presence of heterogeneous agents, whose attributes and decision-making strategies cannot be summarized by population-level averages; ii) the presence of nonlinear interactions among those agents, which preclude the kind of aggregation that equation-based approaches rely on; and iii) the presence of dynamic or complex interaction topologies, in which the structure of who interacts with whom changes over time and is itself part of the phenomenon under study.

As statistical methods become more prominent and employed in real-world settings, the economics literature has also developed models that study potential negative effects of ML over simulated societies. For instance, (Calvano et al. 2020) show the potential for “algorithmic collusion” in simulated product markets when algorithmic dynamic pricing is introduced. Within ML fairness, the foundational agent-based proposal is due to Cruz Cortés and Ghosh (2020), who develop an agent-based model coupled with reinforcement learning to study allocation in predictive policing, and argue explicitly for agent-based simulation as an evaluation framework for fairness interventions. Related work also exists outside the fairness community, most notably the marketing literature, where Ukanwa and Rust (2021) study two competing firms running biased and unbiased algorithms in a market with reputation effects. The contribution of Cruz Cortés and Ghosh has not, however, propagated into the long-term fairness literature in the way its framing might have suggested. Its absence from the recent taxonomy of Gohar et al. (2025) is consistent with this reading. We see this as an indication that the methodology has remained underappreciated, and develop in the following an argument for revising it.

2.3 An argument for ABMs in fairness research

We argue that agent-based modeling is well suited to several of the questions that motivate the long-term fairness literature. We start from noting the taxonomy given by Friedler, Scheidegger, and Venkatasubramanian (2021), where a distinction between the construct space, observation and decision space is drawn. All three spaces are taken to be metric spaces, that is, a distance between elements of the set is defined. The construct space is taken to be the abstract, unmeasurable distribution of characteristics relevant to the process at hand; the measurements of these is operationalized via the observation space; the decision space defines the process’ outcomes starting from observations. This distinction provides us with a conceptual vocabulary for locating where bias enters a decision pipeline. We raise the point that agent-based modeling is a natural way to render that vocabulary operational: each of the three spaces becomes an explicit object in the simulation, the maps between them become named mechanisms, and any bias along a given mechanism becomes a parameter that can be varied independently of the others. This allows researchers and practitioners significant

leeway to investigate which mechanisms may give rise not only to discrimination *per se*, but also to discrimination in machine learning and biases in training data. The remainder of this section develops the implications of this observation along two complementary lines.

The first concerns the relationship between models and the data they analyze. Conventional fairness analyses begin from a fixed dataset and examine the properties of decisions made on it. Agent-based models begin one step earlier, by specifying the mechanism through which the data come into being. In the setting of our interest, this includes who applies for a position, how each applicant is observed, who is hired, and who remains in the system long enough to be measured. With the mechanism in place, its components can be varied independently of one another, which supports two complementary uses. The first is to identify where bias arises and how it propagates, which the present work pursues by separating the channels of bias active at hiring and at promotion and varying them independently of one another. The second is to revisit existing fairness results (e.g., that fairness constraints carry an accuracy cost (Menon and Williamson 2018; Hardt, Price, and Srebro 2016)) which are, when read carefully, statements about the specific generative assumptions under which they hold, or equivalently statements about the maps between construct, observation and decision spaces. A parallel line of work pursues the same commitment to making the generative mechanism explicit, but through causal directed acyclic graphs rather than agent-level specifications. Creager et al. (2020) are the clearest exponent of this view, showing how a causal model of the dynamics supports the computation of interventional and counterfactual quantities through either simulation (when the dynamics are known) or off-policy estimation (when they are not). The agent-based approach we adopt is complementary to theirs: a fully specified agent-based model can be read as a structural causal model in which varying a parameter or mechanism is the operational form of an intervention (Marshall and Galea 2015), and the difference lies in the level of abstraction at which the subjects of the analysis enter the specification: either as nodes in a graph governed by structural equations, or as individually addressable entities with their own attributes. Different ABM simulation runs across different parameters may also be understood as a form of counterfactual example or “what-if” question: we refer the interested reader to (Marshall and Galea 2015) and (Valogianni and Padmanabhan 2022, Section 2) for a broader discussion of the relationship between ABMs and causal modeling.

The second concerns the role of subject-side agency in fairness simulations. Most simulation work in this space has been built around reinforcement learning conventions: an institution makes a sequence of decisions, an environment generates the consequences of those decisions, and the goal is to identify a time-dependent decision rule that satisfies a fairness criterion. As an exemplary case, consider D’Amour et al. (2020), whose `ml-fairness-gym` packages a set of long-term fairness scenarios as Gym-style environments. Nothing in principle prevents one from writing such an environment with multiple agents acting on their own objec-

tives. In practice, however, `ml-fairness-gym` casts the social situation into a single-agent environment, in which the institution is the only locus of explicit objective-setting and everything else, including the subjects of the decisions, is absorbed into the dynamics of a stochastic environment. Agent-based modeling makes the opposite commitment by default. In the standard definition (see Macal and North (2014)) an agent-based model is specified through its agents, their attributes and behaviours, their relationships and methods of interaction, and their environment, each of which the modeller must identify and program. Agents are taken to be autonomous, that is, to act on their own in response to the situations they encounter. To specify such a model is therefore to commit, as part of the specification itself, to identifying who the relevant subjects are, what attributes they carry across time, and what behaviour they exhibit independently of the institution’s decision rule. Axtell (2000) sets out the broader methodological case, distinguishing three uses for agent-based models: as classical Monte Carlo simulation of a system that can be solved analytically, as a complement to mathematical theorizing when the system can be written down but not fully solved, and as the only available tool when writing down equations is itself not useful. The present work falls into the second category. The bias mechanisms, the firm’s prediction models, and the promotion rule can each be written down on their own; we use the simulation to study what happens when they act together, over repeated rounds of hiring and retraining, under several different intervention policies.

3 Methodology

We study fairness in hiring through an agent-based simulation of a firm that hires from a renewing candidate pool, observes outcomes only for workers it has chosen to hire, promotes a subset of those workers, and retrains its predictive models on the data it has accumulated. The simulation is designed to make three properties of the setting explicit and adjustable: a bias in the hiring-stage measurement channel, a bias in the post-hire leadership evaluation, and the cumulative, selectively-labeled training dataset that results from the firm’s own decisions. We describe the components of the model in this section and defer the calibration of its parameters to Section 3.6.

3.1 Candidates and observed features

The candidate pool is populated by individuals indexed by i and characterized by three latent traits. Worker skill S_i captures the component of productivity relevant to the day-to-day execution of a role; leadership skill L_i captures the component relevant to managing others. We model these as separate latent dimensions because hiring and promotion are decisions about distinct kinds of competence, and conflating them would erase the structural difference between the two stages. Each candidate also has experience E_i and a binary group attribute $g_i \in \{0, 1\}$ that plays no causal role in productivity but mediates the observation channel below.

True worker and manager productivity are deterministic

Symbol	Description	Value
<i>Latent traits</i>		
μ_E, σ_E	Experience $\mathcal{N}(\mu_E, \sigma_E^2)$	2, 1
μ_S, σ_S	Skill $\mathcal{N}(\mu_S, \sigma_S^2)$	0, 1
$\ell_{\min}, \ell_{\max}$	Leadership $\mathcal{U}[\ell_{\min}, \ell_{\max}]$	0, 2
<i>Productivity</i>		
w_E, w_S	Worker-productivity weights in P^1	0.5, 0.5
η	Leadership elasticity in P^2	0.5
<i>Observation channels</i>		
ω_T	Test informativeness	0.7
$\sigma_{\hat{S}}$	Pre-hire skill noise std	0.3
$\sigma_{\hat{L}}$	Pre-hire leadership noise std	0.5
σ_{ρ}	Post-hire leadership noise std	0.2
σ_{ν^1}	Post-hire P^1 noise std	0.3
μ_0^T	Test bias, reference group	0
μ_1^T	Test bias, disadvantaged grp	−0.15
μ_0^L	Leadership bias, ref. grp	0
μ_1^L	Leadership bias, dis. grp	Sec. 3.6
<i>Hiring</i>		
α	Worker vs. manager weight in score	0.5
k_H	Hires per step	10
<i>Promotion and attrition</i>		
τ	Softmax temperature	Sec. 3.6
k_P	Promotions per step	Sec. 3.6
$T_{\min}$	Minimum tenure for eligibility	Sec. 3.6
δ	Per-step attrition probability	0
<i>Population and pool</i>		
N	Target firm employment	Sec. 3.6
$ C_0 $	Initial candidate pool size	200
$\Delta C_t $	New candidates per step	20
$\Pr(g_i = 1)$	Group share in candidate pool	0.5

Table 1: Simulation parameters.

functions of the latent traits:

$$P_i^1 = w_E E_i + w_S S_i, \quad P_i^2 = L_i^\eta P_i^1.$$

Worker productivity is a weighted sum of experience and skill; manager productivity scales worker productivity by leadership raised to an elasticity η that controls how much promotion potential depends on leadership relative to baseline worker performance. Latent traits are drawn i.i.d. across candidates and independently of g_i ,

$$E_i \sim \mathcal{N}(\mu_E, \sigma_E^2), \quad S_i \sim \mathcal{N}(\mu_S, \sigma_S^2),$$

$$L_i \sim \mathcal{U}[\ell_{\min}, \ell_{\max}],$$

with parameter values reported in Table 1.

At hiring time the firm does not observe any of S_i, L_i , or the productivities P_i^1, P_i^2 directly. It observes, instead, a vector of noisy or biased proxies:

$$\hat{E}_i = E_i, \quad \hat{S}_i = S_i + \epsilon_i^S, \quad \hat{L}_i = L_i + \epsilon_i^L,$$

with $\epsilon_i^S \sim \mathcal{N}(0, \sigma_{\hat{S}}^2)$ and $\epsilon_i^L \sim \mathcal{N}(0, \sigma_{\hat{L}}^2)$, and a pre-screening test score:

$$\hat{T}_i = \omega_T S_i + \mu_{g_i}^T.$$

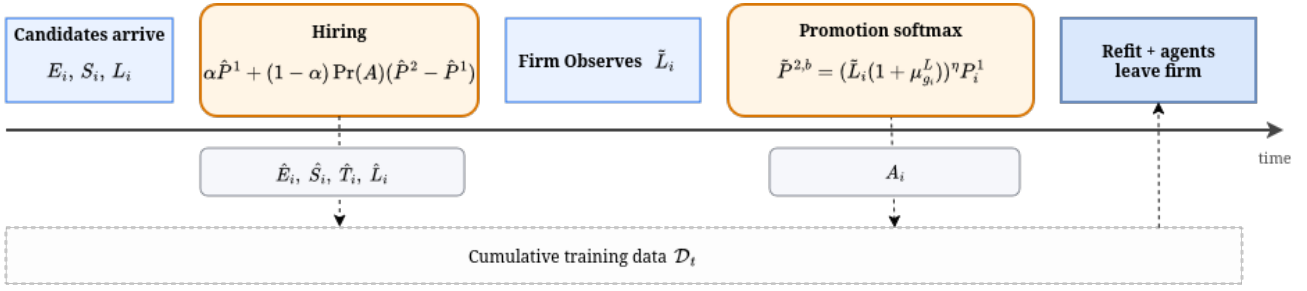


Figure 2: A sketch of one time step of the simulation. New candidates apply to the firm; the firm hires them according to two models, one estimating productivity and one estimating likelihood of future promotion; agents with a certain tenure are considered for promotion and selected via softmax over their estimated productivity; the models are then refit based on the firm’s updated data, and agents will leave the firm at random.

Noise standard deviations and the test informativeness ω_T are listed in Table 1. The test is a biased, weighted observation of worker skill: ω_T controls the test’s informativeness, and $\mu_{g_i}^T$ shifts the measured score by a group-dependent constant. We model $\hat{T}$ as a deterministic transformation of S given g , so its between-group dispersion is attributable to bias rather than to measurement noise. Negative values of μ_1^T correspond to a test that systematically underestimates otherwise comparable candidates from the disadvantaged group.

We operationalize this distortion as a shift in observation space rather than in construct space, thus taking the “we are all equal” position (Friedler, Scheidegger, and Venkatasubramanian 2021). We justify our position by pointing to the empirical literature on the kinds of tests firms use at the first stage of selection. Striebing et al. (2025) document that the cognitive ability tests routinely deployed as the first step of personnel selection in German mass-recruitment procedures (typically, condensed versions of standardized intelligence tests administered in twenty to thirty minutes) systematically disadvantage women, with women at a disadvantage in 35 of the 48 procedures they analyze across five German cities. Another line of work studies the case in which the score may be assumed to represent a self-assessment rather than an externally administered test, and finds bias on the observation side of the same kind: women report lower self-efficacy on cognitive ability tasks (Maertz Jr et al. 2005) and rate their own performance lower than equally productive men working in the same role (Abraham 2023). In both cases the latent skill distribution is taken to be group-independent, and the bias enters through the channel by which skill is measured, a construct-to-observation distortion of the kind discussed by Friedler, Scheidegger, and Venkatasubramanian (2021). The parameter μ_g^T in our model is the simplest operationalization of this channel. In what follows we report results both for arbitrary settings of μ_g^T , in order to map the simulation’s bias-generating behaviour, and for a value calibrated against the cognitive-test evidence. For the latter we follow Striebing et al. (2025), who report a WAIS-IV gender gap of Cohen’s $d \approx 0.22$; under our parameterisation this corresponds to $\mu_1^T \approx -d \cdot \omega_T \sigma_S$ at the default values of ω_T and σ_S (Table 1), treating the reference group as $\mu_0^T = 0$.

In several experiments the noisy skill observation $\hat{S}_i$ is suppressed and the firm uses only $\hat{E}_i$ and $\hat{T}_i$, reflecting settings in which raw skill is not directly assessable and the test is the only available proxy.

3.2 Hiring

Hiring is the firm’s principal decision. The firm aims to identify the candidates whose expected combined contribution, as workers in the near term, and as potential managers in the long term, is largest. It trains three predictive models from its accumulated training data: a regressor $\hat{P}^1$ for worker productivity, a classifier $\hat{\Pr}(A = 1)$ for the probability of eventual promotion, and a regressor $\hat{P}^2$ for manager productivity. The two productivity regressors are fitted by ridge regression and the promotion classifier is a logistic regression, in both cases using the default implementations of `scikit-learn` (Pedregosa et al. 2011). The first model uses the hiring-time features $(\hat{E}_i, \hat{S}_i, \hat{T}_i)$, or $(\hat{E}_i, \hat{T}_i)$ when skill is not observed; the latter two use the same features augmented with $\hat{L}_i$.

The hiring score combines worker-productivity and manager-potential contributions through a single weight $\alpha \in [0, 1]$:

$$\text{score}_i = \alpha \hat{P}_i^1 + (1 - \alpha) \hat{\Pr}(A_i = 1) (\hat{P}_i^2 - \hat{P}_i^1).$$

The second term is the expected incremental value of promotion: the gain a candidate would deliver as a manager, weighted by the firm’s estimate that the candidate will in fact be promoted. Subtracting $\hat{P}_i^1$ rather than adding $\hat{P}_i^2$ outright avoids double-counting the worker contribution. Each round the firm ranks the candidate pool by this score and extends offers to the top k_H .

Because the firm trains its three models only on workers it has actually hired, rejected candidates never produce outcome labels. The selective-labels problem is therefore a structural feature of the setting (Lakkaraju et al. 2017), and the firm’s policy in any given round shapes the data available for retraining in subsequent rounds, in the manner described by Liu et al. (2018).

3.3 Employment, promotion, and attrition

After a candidate is hired, the firm receives a noisy observation of worker productivity, $\tilde{P}_i^1 = P_i^1 + \nu_i^1$ with $\nu_i^1 \sim \mathcal{N}(0, \sigma_{\nu^1}^2)$, and a more accurate observation of leadership, $\tilde{L}_i = L_i + \rho_i$ with $\rho_i \sim \mathcal{N}(0, \sigma_{\rho}^2)$ and $\sigma_{\rho} < \sigma_{\tilde{L}}$ (Table 1). The reduced noise reflects that on-the-job interactions may yield more information about leadership than a pre-hire interview does.

Promotion is modeled as a stochastic selection from the set $\mathcal{W}_t$ of workers eligible at time t . Eligibility requires minimum tenure; among the eligible, the firm assesses leadership through a biased channel,

$$\tilde{L}_i^b = \tilde{L}_i (1 + \mu_{g_i}^L),$$

and uses it to compute a leadership-adjusted manager-productivity score $\tilde{P}_i^{2,b} = (\tilde{L}_i^b)^\eta \tilde{P}_i^1$. Promotion probabilities follow a temperature-controlled softmax,

$$\Pr(A_i = 1 \mid \mathcal{W}_t) = \frac{\exp(\tilde{P}_i^{2,b}/\tau)}{\sum_{j \in \mathcal{W}_t} \exp(\tilde{P}_j^{2,b}/\tau)},$$

which is a tractable smooth analogue of deterministic top- k_P ranking and lets the temperature τ interpolate between near-random and near-deterministic promotion. Workers also leave the firm at a fixed per-step attrition rate, which keeps the workforce composition from saturating and creates a steady demand for new hires.

As with the hiring-stage bias, we report results both for arbitrary settings of $\mu_{g_i}^L$ and for a calibrated value. We employ results reported by Xu (2024), who reports that in the US within-firm gender promotion gap such that women’s promotion probability is approximately 84% of men’s controlling for observable characteristics, against a baseline annual promotion rate of about 5.7%. We set the reference group to $\mu_0^L = 0$ and choose μ_1^L together with the softmax temperature τ to reproduce these two empirical moments under the simulated-moment matching procedure described next.

3.4 Retraining and dataset endogeneity

At each step, the firm appends the hire-time features and observed post-hire outcomes of every newly hired worker to its cumulative training dataset, together with the worker’s group label and an eventual promotion indicator. The three predictive models introduced in Section 3.2 are retrained on this growing dataset between rounds. Two consequences follow. First, the firm’s training data is not a sample from the candidate-pool distribution but from the distribution of candidates the firm chose to hire, which depends on the policy in use. Second, the firm’s policy in any given round shapes the data available for retraining in subsequent rounds, producing the kind of long-term feedback dynamics that have motivated the recent dynamic-fairness literature (Liu et al. 2018; Ensign et al. 2018; D’Amour et al. 2020; Gohar et al. 2025).

3.5 Interventions

We compare four hiring-side intervention families that act on different points of the pipeline. The first is the *status quo*,

in which the firm trains and scores using the biased features as observed. The second is an *oracle* repair of the measurement channel: the firm is given the true value of $\mu_{g_i}^T$ and subtracts the group-specific shift from $\tilde{T}_i$ before training and scoring. The oracle is not a deployable policy — the bias parameter is structurally unobserved in any real setting — but it serves as a benchmark for the most that any test-channel repair could achieve.

The third intervention is *observational debiasing*, in which the features are used as observed but the predictive models are fit under group-fairness constraints. Concretely, we use the Fairlearn toolkit (Weerts et al. 2023) for the binary promotion classifier and the reduction-based fair regression algorithm of Agarwal, Dudík, and Wu (2019) for the two productivity regressors. The fourth intervention is a *population-level demographic-parity top- k_H* rule: the learned score is left untouched but offers are allocated so that the share of hires from each group matches the share of candidates from that group in the applicant pool. Whenever necessary, we will thus assume that the firm also has access to a ground-truth measurement of the binary group an agent belongs to, which is an idealized assumption.

These four interventions separate three conceptually different modes of action on the pipeline: repairing the measurement channel (oracle), constraining the learner (observational debiasing), and constraining the final allocation (demographic-parity top- k_H).

3.6 Calibration

The simulation’s “bias parameters” $\mu_{g_i}^T$ and $\mu_{g_i}^L$ were introduced alongside each mechanism in Sections 3.1 and 3.3. We now discuss the problem of calibration. Overall, calibration may be defined at a high level as “[...] running the model with different parameters and testing, for each case, how well the model performs by comparing the output against empirical data” (McCulloch et al. 2022, Paragraph 1.3).

The promotion side of the model does not admit a transparent one-parameter mapping from an empirical promotion gap to a single structural coefficient, because the observed rates are jointly shaped by the promotion bias parameter μ_1^L , the softmax temperature τ , the promotion quota, and the size and composition of the eligible worker pool. For that reason we use a simulated minimum-distance approach (Grazzini and Richiardi 2015): we pick a small set of summary statistics (moments) from the simulation data, simulate the model under candidate parameters, and choose the configuration whose simulated moments are closest to the empirical ones. Moments are evaluated on a post-burn-in window, so that they are taken from the model’s statistical equilibrium, and averaged across seeds. We note that we use a discrete grid of pre-defined values rather than continuous optimization of

We treat one simulation step as one month. This fixes the temporal scale at which the simulation parameters apply. Then, as mentioned beforehand, we choose the hiring-stage bias with help from cognitive-test evidence: at the default values of ω_T and σ_S (Table 1) and Cohen’s $d \approx 0.22$ from Striebing et al. (2025), we have $\mu_1^T \approx -d \cdot \omega_T \sigma_S \approx -0.15$.

The promotion-side parameters are obtained by

simulated-moment matching against the data reported by Xu (2024). The two target moments are the aggregate monthly promotion rate $r^* = 0.057/12 \approx 0.00475$ and the group-1/group-0 promotion ratio $q^* = 0.84$. We minimise the equal-weight relative squared-error objective

$$L(\theta) = \left(\frac{r_{\text{sim}}(\theta) - r^*}{r^*} \right)^2 + \left(\frac{q_{\text{sim}}(\theta) - q^*}{q^*} \right)^2$$

over the parameter vector $\theta = (\mu_1^L, \tau, N, k_P, T_{\min})$, where N denotes the target employment level, k_P the number of promotions made each step, and $T_{\min}$ the minimum tenure required for promotion eligibility.

The simulated moments are computed from the late-run behavior of the model (the last half) rather than from a single promotion round. For each candidate parameter configuration and each random seed, we discard an initial burn-in window, retain the last 50% post-burn steps, and compute the promotion moments as averages over that window. The overall promotion-rate moment uses the worker-base denominator $n_{\text{promoted}}/n_{\text{workers}}$ pre-promotion at each step; the group-specific moments use the analogous group-level worker-base denominators, and the promotion-ratio moment is formed from those group-specific rates. We then average the resulting late-window moments across seeds. The objective $L(\theta)$ is therefore evaluated on a cross-seed average of approximately steady-state promotion behavior, rather than on transient values from the early part of the simulation.

Operationally, we solve this objective by a two-stage sweep. In the first stage we hold $\mu_1^L = 0$ and $\tau = 1$ and sweep the target employment level N , the promotion quota k_P , and the minimum-tenure threshold $T_{\min}$, retaining the configurations whose simulated monthly promotion rate is closest to r^* . This first stage isolates what is essentially the mechanical component of the calibration problem: the aggregate promotion rate is governed primarily by how many promotions can be made each step, how large the worker base is, and how many workers satisfy the tenure threshold, whereas μ_1^L and τ mainly affect which workers are promoted and therefore matter more for the relative group promotion gap than for the overall scale of promotion. In the second stage we condition on the best stage-one configuration and sweep over μ_1^L and τ , minimising $L(\theta)$ over both moments jointly. The resulting parameters are

$$\mu_1^L = -0.25, \quad \tau = 1.75, \quad N = 220, \\ k_P = 1, \quad T_{\min} = 10,$$

with achieved fit $r_{\text{sim}} = 0.00498$ against $r^* = 0.00475$ and $q_{\text{sim}} = 0.833$ against $q^* = 0.84$. Nearby alternatives in the same neighbourhood ($\mu_1^L \in [-0.25, -0.15]$, $\tau \in [1.25, 1.75]$, $T_{\min} \in \{10, 12, 14\}$) achieved comparable fit, so the process appears relatively stable. All remaining simulation parameters are held at the defaults introduced earlier in this section and in Table 1.

3.7 Outcome metrics

We report four families of outcomes throughout. To state the definitions compactly, let $\mathcal{C}_t$ denote the candidate pool at step t , $\mathcal{H}_t$ the workers currently employed by the firm,

$\mathcal{M}_t$ the managers currently employed by the firm, $\mathcal{N}_t \subseteq \mathcal{H}_t$ the workers hired at step t (the *hired cohort*), and $\mathcal{D}_t$ the cumulative training dataset accumulated up to step t .

Hiring and employment. The hiring-rate gap between groups at step t is

$$\Delta_t^{\text{hire}} = \Pr(i \in \mathcal{N}_t \mid g_i = 0) - \Pr(i \in \mathcal{N}_t \mid g_i = 1),$$

and the promotion-rate gap Δ_t^{promo} is defined analogously, with the conditioning set replaced by the worker base $\mathcal{H}_t$. We also report the share of group-1 candidates among current workers and managers,

$$\pi_{1,t}^W = \frac{|\{i \in \mathcal{H}_t : g_i = 1\}|}{|\mathcal{H}_t|}, \quad \pi_{1,t}^M = \frac{|\{i \in \mathcal{M}_t : g_i = 1\}|}{|\mathcal{M}_t|},$$

and per-capita firm productivity,

$$\bar{\Pi}_t = \frac{1}{|\mathcal{H}_t \cup \mathcal{M}_t|} \left(\sum_{i \in \mathcal{H}_t} P_i^1 + \sum_{i \in \mathcal{M}_t} P_i^2 \right),$$

formed by summing true worker productivity over current workers and true manager productivity over current managers, and dividing by the headcount. The corresponding total productivity is $\Pi_t = \bar{\Pi}_t \cdot |\mathcal{H}_t \cup \mathcal{M}_t|$ and is reported alongside the per-capita version when the size of the firm itself varies across conditions.

A separate diagnostic isolates the latent quality of newly hired workers by group. The *double-standard gap* is the difference between the two groups' mean true worker productivity within the hired cohort,

$$\Delta_t^{\text{ds}} = \mathbb{E}[P_i^1 \mid i \in \mathcal{N}_t, g_i = 0] - \mathbb{E}[P_i^1 \mid i \in \mathcal{N}_t, g_i = 1].$$

A negative value indicates that group-1 hires have, on average, higher true worker productivity than group-0 hires, that is, that they had to clear a higher effective bar to be hired.

Training data composition. The cumulative dataset is summarized by the share of group-1 examples,

$$\pi_{1,t}^{\mathcal{D}} = \frac{|\{i \in \mathcal{D}_t : g_i = 1\}|}{|\mathcal{D}_t|},$$

and by the group-conditional rate of positive promotion labels recorded in $\mathcal{D}_t$. The promotion gap is thus

$$\Delta_t^{\mathcal{D}, \text{promo}} = \Pr(A_i = 1 \mid i \in \mathcal{D}_t, g_i = 0) \\ - \Pr(A_i = 1 \mid i \in \mathcal{D}_t, g_i = 1).$$

These two quantities track, respectively, how the policy in use reshapes who appears in the firm's training data and how the distribution of labels within that data evolves over time.

Learned signal. The central diagnostic of the learned score is the group-differential prediction bias of the worker-productivity regressor,

$$b_t^{(g)} = \mathbb{E}[\hat{P}_i^1 - P_i^1 \mid i \in \mathcal{C}_t, g_i = g], \quad \Delta_t^{\text{pred}} = b_t^{(0)} - b_t^{(1)}.$$

A value of Δ_t^{pred} near zero indicates that the worker-productivity model does not differentially over- or under-value one group; values away from zero indicate that an intervention has changed surface outcomes without repairing the learned signal. This is the principal device through which we distinguish interventions that act on decisions from those that act on the model and the dataset that produced them.

Recoverability of the biased channel. The experiments include two diagnostics that ask whether the structural bias in the hiring-test channel is recoverable from the firm’s selectively-labelled data.

The first runs a causal-discovery procedure on a prefix of the cumulative dataset $\mathcal{D}_t$ of size n and asks whether the procedure recovers an edge between the group attribute g and the observed test $\hat{T}$ in the inferred graph. Causal-discovery algorithms reconstruct a graph over a set of variables from a sequence of conditional-independence tests on the data, adding an edge between two variables when they fail to become independent after conditioning on subsets of the remaining ones, and orienting edges using statistical and background-knowledge constraints (Spirtes and Glymour 1991). We run two such algorithms, the PC algorithm (Spirtes and Glymour 1991), which assumes no unobserved confounders, and the FCI algorithm (Spirtes 2001), which relaxes that assumption. We rely on the `causal-learn` library to do so. The variables exposed to the discovery step are $(g, \hat{T}, \hat{S}, \hat{L}, \hat{E}, \hat{P}^1, A)$, and the quantity of interest is whether the edge between g and $\hat{T}$ appears in the recovered graph $\hat{\mathcal{G}}_n$. We report the empirical rate across random seeds at which this happens,

$$r_n = \Pr(g \rightarrow \hat{T} \in \hat{\mathcal{G}}_n),$$

as a function of prefix size n . We compute probabilities as the rate of recovery over 10 random seeds. Values close to one indicate that the firm’s accumulated data is consistently informative about the existence of a group-conditional test channel; values close to zero indicate that the bias is not yet detectable from the available data.

The second diagnostic asks not only whether the bias channel is detectable but whether its magnitude can be recovered. We use a downstream estimator $\hat{\mu}_1^T(\mathcal{D}_t)$ for the hiring-test bias, formed from the firm’s accumulated data, and track its absolute calibration error $|\hat{\mu}_1^T(\mathcal{D}_t) - \mu_1^T|$ as a function of the simulation step t . Together the two diagnostics quantify whether the structural bias is both detectable from the firm’s endogenous training data and, if so, whether its magnitude is estimable to within a useful tolerance.

4 Experiments

We use the simulation introduced in Section 3 as a controlled environment for three empirical questions. The central parameters across all experiments are the hiring-stage measurement bias μ_1^T , the promotion-stage bias μ_1^L , the hiring-score weight α , and the candidate-pool group composition. We report two complementary kinds of result: broad parameter sweeps that map the simulation’s bias-generating behaviour, and a calibrated slice that fixes the bias parameters, softmax temperature, and steady-state workforce size to the values from Section 3.6. Our experiments are implemented with the `fairsim` library (<https://pibborn.github.io/fairsim>).

4.1 RQ1: when and how does bias emerge?

The first question is whether the simulation produces structured disparities under the status-quo policy, and how those

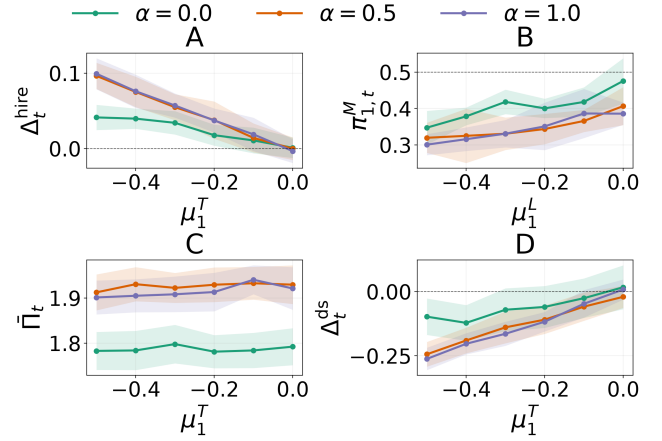


Figure 3: RQ1. Bias emergence under the status-quo policy. Panels: (A) Δ^{hire} versus μ_1^T , separate lines per $\alpha \in \{0, 0.5, 1\}$, averaged over μ_1^L ; (B) π_1^M versus μ_1^L at $\mu_1^T = -0.5$; (C) $\bar{\Pi}_t$ versus μ_1^T ; (D) Δ^{ds} versus μ_1^T . Shaded bands are 95% confidence intervals across seeds.

disparities depend on the bias parameters and on the scoring rule. The broad sweep runs the status-quo policy over $\mu_1^T, \mu_1^L \in \{0, -0.1, -0.2, -0.3, -0.4, -0.5\}$, the hiring-score weight $\alpha \in \{0, 0.5, 1\}$, and two candidate-pool compositions, $(0.5, 0.5)$ and $(0.7, 0.3)$. The calibrated slice fixes $\mu_1^T = -0.15, \mu_1^L = -0.25$ and varies the same α and composition values. Both use five seeds and one hundred simulation steps (~ 8 years, see Section 3.6), aggregated over the last twenty steps after a twenty-step burn-in.

Figure 3 reports four diagnostics from the sweep. The hire-rate gap Δ^{hire} grows monotonically with the magnitude of μ_1^T , the group-1 manager share π_1^M falls monotonically with the magnitude of μ_1^L at strong test bias, per-capita productivity $\bar{\Pi}$ declines mildly with μ_1^T , and the double-standard gap Δ^{ds} becomes increasingly negative as the test becomes more biased against group 1. The interpretation of the double-standard signal is that group-1 hires must clear a higher latent-productivity bar than group-0 hires when the test is biased against them.

In the calibrated slice, the hire-rate gap continues to scale with μ_1^T but the productivity decline visible at extreme bias does not survive. At empirically anchored magnitudes the hiring-stage disparity is the only signal that persists.

4.2 RQ2: intervention comparison

The second question is how the four intervention families introduced in Section 3.5, namely status quo, oracle repair of the measurement channel, observational debiasing, and population-level demographic-parity top- k_H , compare under biased measurement, selective labels, and cumulative re-training. We use a stationary regime with target headcount 40 (outside the calibrated slice $N = 220$, chosen so attrition and intake balance), attrition rate 0.05, a $(0.7, 0.3)$ candidate pool, $\alpha = 0.5$, five seeds, and one hundred steps, at $\mu_1^T = -0.5, \mu_1^L = 0$.

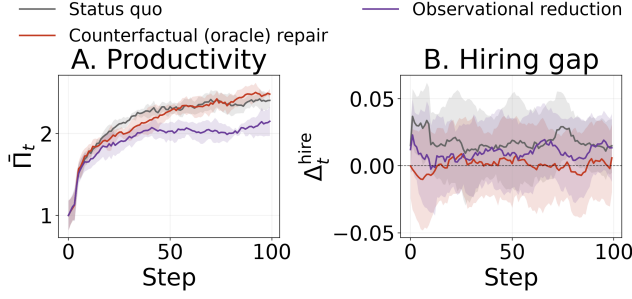


Figure 4: RQ2. Intervention comparison in a strong-bias experiment ($\mu_1^T = -0.5$, $\mu_L^T = 0$). Shaded bands are 95% confidence intervals across seeds.

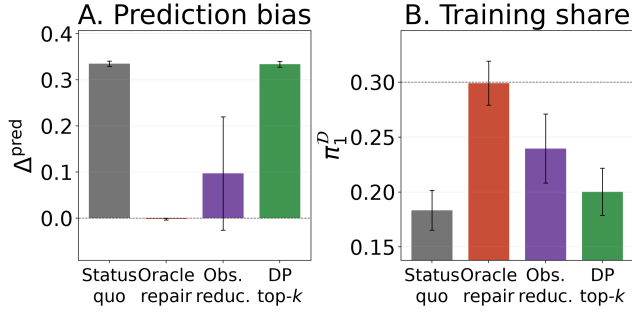


Figure 5: RQ2. Dataset-side comparison of the four interventions in the strong-bias experiment.

Figure 4 reports the interventions on Δ^{hire} , π_1^W , π_1^M , and $\bar{\Pi}$ in the strong-bias experiment. The oracle achieves the highest productivity and lowest bias, showing that here fair prediction of who to hire does not come at a fairness cost. Observational debiasing, on the other hand, is not aware of the underlying causal structure and thus struggles to maximize the productivity, even while maintaining a relatively low hiring gap.

The four policies do not, however, differ only on decision-space metrics, but on the resulting dataset collected following these policies. Figure 5 reports the group-1 dataset share π_1^D , and the prediction-bias gap Δ^{pred} across the four interventions. Both demographic-parity top- k_H and observational debiasing shift the hire-rate gap and the dataset composition relative to the status quo, but the prediction-bias gap remains essentially at its status-quo level for both. The oracle is the only intervention that also moves Δ^{pred} to near zero. Our reading is that interventions which constrain the learner or the allocation can adjust the firm’s surface outcomes while leaving its productivity model differentially biased across groups; only intervening on the measurement channel itself repairs the model.

4.3 RQ3: detectability of the biased test channel

The third question is whether the firm could infer the existence and magnitude of the test-channel bias from its own selectively labelled data. We use the two diagnostics described in Section 3.7: the causal-discovery edge recovery rate r_n on

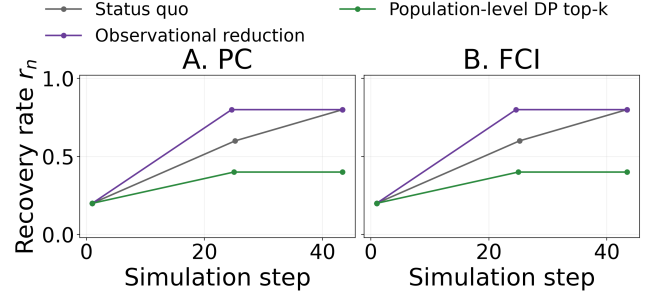


Figure 6: Edge-recovery rate r_n of the $g-\hat{T}$ channel under the three hiring-side policies at the calibrated bias $\mu_1^T = -0.15$. Panel A uses the PC algorithm; panel B uses FCI. Both algorithms run with a Fisher-Z conditional-independence test at $\alpha = 0.05$. Each point is the seed-wise mean over five independent simulation runs, evaluated at five cumulative-dataset prefix sizes; the x-axis shows the simulation step at which each prefix size is reached on average.

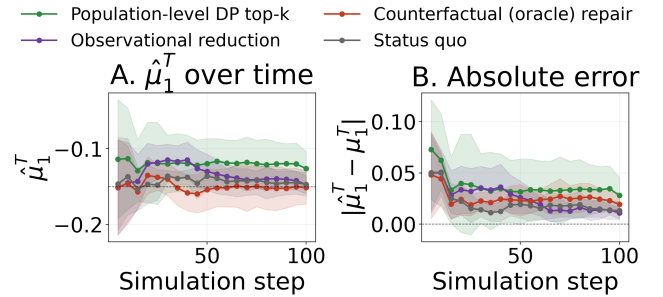


Figure 7: RQ3. Estimation of the hiring-test bias when $\mu_1^T = -0.15$. Left: per-policy estimator $\hat{\mu}_1^T(\mathcal{D}_t)$, with the true μ_1^T shown as a dashed horizontal line. Right: absolute error.

prefixes of the cumulative dataset, and the downstream estimator $\hat{\mu}_1^T(\mathcal{D}_t)$ together with its absolute calibration error.

At the calibrated bias $\mu_1^T = -0.15$, the three policies separate (Figure 6). Status quo and observational debiasing both reach $r_n = 0.8$ by the largest prefix, with observational debiasing arriving roughly eighteen steps earlier. Demographic-parity top- k_H plateaus at $r_n = 0.4$ and does not improve with additional data. PC and FCI agree throughout, so the effect is not an artifact of assuming causal sufficiency. Constraining the final allocation disturbs the joint distribution of g and $\hat{T}$ enough to remove roughly half the long-run statistical trace of the bias channel, while constraining the learner leaves the channel as detectable as the status quo, even if more data is necessary.

Figure 7 reports the magnitude diagnostic at the calibrated bias: the per-policy estimator $\hat{\mu}_1^T(\mathcal{D}_t)$ and its absolute error $|\hat{\mu}_1^T(\mathcal{D}_t) - \mu_1^T|$ over step t . The oracle and observational debiasing converge toward μ_1^T ; demographic-parity top- k_H drifts away from it, consistent with the edge-recovery plateau above.

4.4 Summary

The three experiments combine into a single substantive picture. On the first question (RQ1, when and how does bias emerge?), the simulation produces structured hiring-side disparities tracking μ_1^T , and the productivity decline visible at extreme bias does not survive at calibrated magnitudes (Figure 3 C). There is therefore no accuracy-fairness tradeoff to negotiate in the calibrated regime: the firm does not pay a productivity price for fair measurement. On the second question (RQ2, how do the four interventions compare?), separating them at the level of Δ^{pred} , the group-differential prediction bias of the firm’s productivity model, reveals a hierarchy that surface metrics conceal. Demographic-parity top- k_H rebalances the allocation without touching the learned signal; observational debiasing slightly improves both surface fairness and the underlying dataset composition, but does not close the prediction-bias gap to the extent the oracle does. Only intervening on the measurement channel itself approaches causal repair. On the third question (RQ3, is the bias channel recoverable from the firm’s own data?), the answer depends on which policy generated that data. Under status quo and observational debiasing, the $g \rightarrow \hat{T}$ channel becomes recoverable and $\hat{\mu}_1^T(\mathcal{D}_t)$ converges toward the truth on a moderate sample size. Under demographic-parity top- k_H , recovery plateaus at roughly half the long-run rate of the other two policies and the magnitude estimator drifts away from μ_1^T . The firm that intervenes on allocation also loses the statistical trace it would need to detect and repair the underlying problem.

5 Conclusion

We have presented an agent-based simulation of a hiring-and-promotion pipeline that makes biased measurement, selective labels, and cumulative retraining separately tunable mechanisms. Across three research questions we showed that hiring-side bias produces structured disparities under the status quo (RQ1); that interventions which constrain the learner or the allocation can shift the hire-rate gap and group representation while leaving the firm’s productivity-prediction model differentially biased across groups (RQ2); and that the biased measurement channel is detectable and estimable from the firm’s own selectively labelled data, depending on the policy (RQ3). The values we use for μ_1^T , μ_1^L , the softmax temperature τ , and the steady-state workforce size are calibrated against magnitudes drawn from the cognitive-testing literature and the within-firm promotion literature.

The simulation has been calibrated against empirical magnitudes but not validated in the sense the term is used in the agent-based modelling literature. Following Fagiolo et al. (2019), calibration matches model parameters to a small set of empirical moments, whereas output validation asks whether the model reproduces empirical regularities that were *not* used to fix the parameters. Our simulation matches the two moments it was tuned to by construction, but we have not shown that it reproduces further labour-market regularities, such as tenure-conditional promotion rates, productivity distributions, or cohort-level dynamics, that we

did not target. Fagiolo et al. (2019) reproduce a four-level empirical-validity taxonomy originally due to Axtell and Epstein. A Level-0 model is a caricature of reality, useful as a graphical or conceptual device but making no empirical claims. A Level-1 model is in qualitative agreement with empirical macro-structures, established by comparing simulated and empirical distributions (firm-size distributions, wage curves, and so on). Level 2 requires quantitative agreement at the macro scale, via on-board statistical estimation routines. Level 3 requires quantitative agreement at the level of individual agents, established from cross-sectional and longitudinal analysis of the agent population. Against this scheme our simulation is a Level-0 model with calibrated parameters: we have matched two specific moments but not demonstrated broader distributional agreement. This taxonomy was developed for ABMs whose ambition is predictive macro-modelling, in the tradition of competing with or replacing DSGE models, and its higher levels reflect that ambition. Our purpose is different. We use the simulation to display mechanisms by which biased training data can arise under standard fairness interventions, not to predict labour-market outcomes. Grüne-Yanoff (2009) draws a related distinction within the philosophy of simulation between *actual* and *potential* explanation: a potential explanation specifies behaviours and interactions that suffice to generate the phenomenon of interest, without claiming that they are the actual causes of any empirically observed instance of it. Hoffman, Sharma, and Watts (2017) make a parallel distinction between prediction and explanation in social systems more generally. The contribution we offer is closer to a potential explanation in this sense: a demonstration that the combination of biased measurement, selective labels, and cumulative retraining can produce the kinds of disparities that fairness interventions are deployed against, and that the choice of intervention is consequential not only for the hire-rate gap and group representation but also for whether the firm’s productivity model continues to under-predict the disadvantaged group.

The natural next steps are systematic output validation against labour-market regularities beyond the calibration moments, sensitivity analysis under non-stationary candidate distributions, and extension to multi-firm market settings in which firms compete for the same applicants. We also note that the mechanisms we operationalise, particularly the cumulative-dataset feedback, are not specific to hiring and could in principle be transferred to other selective-label domains in which long-term fairness is studied. The substantive claim of the paper, that an evaluation of fairness in such domains should include inspection of the firm’s prediction model and the dataset that produced it rather than only its hire-rate outcomes, is what we seek to make possible with the present contribution.

Acknowledgements

MC and SK gratefully acknowledge the support of the “TOPML: Trading Off Non-Functional Properties of Machine Learning” project funded by the Carl Zeiss Foundation, grant number P2021-02-014.

References

- Abraham, L. 2023. The gender gap in performance reviews. *Journal of Economic Behavior & Organization*, 214: 459–492.
- Agarwal, A.; Dudík, M.; and Wu, Z. S. 2019. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA*.
- Axtell, R. 2000. Why Agents? On the Varied Motivations for Agent Computing in the Social Sciences. Technical Report Working Paper No. 17, Center on Social and Economic Dynamics, The Brookings Institution.
- Bonabeau, E. 2002. Agent-Based Modeling: Methods and Techniques for Simulating Human Systems. *Proceedings of the National Academy of Sciences*, 99(suppl. 3): 7280–7287.
- Calvano, E.; Calzolari, G.; Denicolo, V.; and Pastorello, S. 2020. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10): 3267–3297.
- Creager, E.; Madras, D.; Pitassi, T.; and Zemel, R. 2020. Causal modeling for fairness in dynamical systems. In *International conference on machine learning*, 2185–2195. PMLR.
- Cruz Cortés, E.; and Ghosh, D. 2020. An invitation to system-wide algorithmic fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 235–241.
- D’Amour, A.; Srinivasan, H.; Atwood, J.; Baljekar, P.; Sculley, D.; and Halpern, Y. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.
- Ensign, D.; Friedler, S. A.; Neville, S.; Scheidegger, C.; and Venkatasubramanian, S. 2018. Runaway Feedback Loops in Predictive Policing. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, volume 81 of *Proceedings of Machine Learning Research*, 160–171.
- Epstein, J. M. 2006. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton, NJ: Princeton University Press.
- Fagiolo, G.; Guerini, M.; Lamperti, F.; Moneta, A.; and Roventini, A. 2019. Validation of agent-based models in economics and finance. In *Computer simulation validation: fundamental concepts, methodological frameworks, and philosophical perspectives*, 763–787. Springer.
- Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2016. On the (Im)possibility of Fairness. *arXiv preprint*. Revised version published as Friedler, Scheidegger, Venkatasubramanian, “The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making”, *Communications of the ACM* 64(4):136–143, 2021.
- Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. 2021. The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Communications of the ACM*, 64(4): 136–143.
- Gohar, U.; Tang, Z.; Wang, J.; Zhang, K.; Spirtes, P.; Liu, Y.; and Cheng, L. 2025. Long-Term Fairness Inquiries and Pursuits in Machine Learning: A Survey of Notions, Methods, and Challenges. *Transactions on Machine Learning Research*.
- Grazzini, J.; and Richiardi, M. 2015. Estimation of ergodic agent-based models by simulated minimum distance. *Journal of Economic Dynamics and Control*, 51: 148–165.
- Grüne-Yanoff, T. 2009. The explanatory potential of artificial societies. *Synthese*, 169(3): 539–555.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29 (NeurIPS 2016)*.
- Hofman, J. M.; Sharma, A.; and Watts, D. J. 2017. Prediction and explanation in social systems. *Science*, 355(6324): 486–488.
- Hu, L.; and Chen, Y. 2018. A Short-term Intervention for Long-term Fairness in the Labor Market. In *Proceedings of the 2018 World Wide Web Conference (WWW)*, 1389–1398.
- 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 *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 275–284.
- Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3150–3158. PMLR.
- Macal, C.; and North, M. 2014. Introductory tutorial: Agent-based modeling and simulation. In *Proceedings of the winter simulation conference 2014*, 6–20. IEEE.
- Maertz Jr, C. P.; Bauer, T. N.; Mosley Jr, D. C.; Posthuma, R. A.; and Campion, M. A. 2005. Predictors of self-efficacy for cognitive ability employment testing. *Journal of Business Research*, 58(2): 160–167.
- Marshall, B. D.; and Galea, S. 2015. Formalizing the role of agent-based modeling in causal inference and epidemiology. *American journal of epidemiology*, 181(2): 92–99.
- McCulloch, J.; Ge, J.; Ward, J. A.; Heppenstall, A.; Polhill, J. G.; and Malleson, N. 2022. Calibrating agent-based models using uncertainty quantification methods. *Journal of Artificial Societies and Social Simulation*, 25(2).
- Menon, A. K.; and Williamson, R. C. 2018. The cost of fairness in binary classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 107–118. PMLR.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, 12: 2825–2830.
- Schelling, T. C. 1971. Dynamic Models of Segregation. *Journal of Mathematical Sociology*, 1(2): 143–186.

- Spirtes, P. 2001. An anytime algorithm for causal inference. In *International Workshop on Artificial Intelligence and Statistics*, 278–285. PMLR.
- Spirtes, P.; and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1): 62–72.
- Striebing, C.; Schneider, J.; Uhler, L.; and Hartmann, F. 2025. The Fairness Imperative: Challenging Gender Bias in Cognitive Ability Tests. *Zeitschrift für Arbeits- und Organisationspsychologie*, 69(1): 28–32.
- Ukanwa, K.; and Rust, R. T. 2021. Algorithmic bias in service. *USC Marshall School of Business Research Paper*.
- Valogianni, K.; and Padmanabhan, B. 2022. Causal ABMs: Learning plausible causal models using agent-based modeling. In *The KDD’22 Workshop on Causal Discovery*, 3–29. PMLR.
- Weerts, H.; Dudík, M.; Edgar, R.; Jalali, A.; Lutz, R.; and Madaio, M. 2023. Fairlearn: Assessing and Improving Fairness of AI Systems. *Journal of Machine Learning Research*, 24(257): 1–8.
- Xu, J. 2024. The gender gap in executive promotions. *Journal of Corporate Finance*, 89: 102680.