Long-term Effects of Fairness Metrics on Population Dynamics

Mirthe Dankloff[1][0009−0009−4439−5749], Yining Yuan[2][0009−0000−3093−5526], Nirav Ajmeri[2][0000−0003−3627−097X], and Vahid Yazdanpanah[3][0000−0002−4468−6193]

Vrije Universiteit Amsterdam, Amsterdam, The Netherlands m.e.dankloff@vu.nl
University of Bristol, Bristol, United Kingdom
{yining.yuan,nirav.ajmeri}@bristol.ac.uk
University of Southampton, Southampton, United Kingdom
V.Yazdanpanah@soton.ac.uk

Abstract. Algorithmic fairness is often treated as a static property, overlooking that individuals may disengage from systems they perceive as unfair. We introduce a dynamic notion of perceived fairness in a lending scenario that captures how repeated unjust denials and observation of peer outcomes can drive applicants to opt out. Through a multi-agent simulation framework on synthetic data, we measure how different fairness metrics affect long-term population retention and feature dynamics. Our results show that without fairness constraints, apparent fairness improvements arise from the selective opt-out of disadvantaged applicants (survivorship bias). Demographic parity and equal opportunity reduce immediate retention disparities but do not guarantee long-term fairness; demographic parity, in particular, overcorrects participation dynamics, accumulating long-term unfairness. We compare this against a causal fairness model that achieves a balanced retention rate and lower long-term unfairness. Our findings highlight the need to assess long-term fairness in settings with endogenous participation, where individual decisions are shaped by perceived fairness and peer effects, beyond static fairness constraints.

Keywords: Long-Term Fairness · Social Influence · Agent-Based Simulation.

1 Introduction

Algorithmic systems are increasingly used for resource allocation decisions such as lending and other public services, making fair treatment between demographic groups a critical concern [2,6,9]. Fairness metrics (e.g., demographic parity) have been proposed to equalize the outcomes between groups through statistical constraints [1,6]. However, prior research has shown that applying such metrics over time can produce unintended effects, such as overlending to disadvantaged groups or amplifying disparities, highlighting the need to study long-term fairness dynamics [3,8,12]. Existing research on long-term fairness primarily model feature dynamics, examining how decisions affect future applicant qualifications

(e.g., credit scores) [3,8,12]. These studies adopt a decision-maker's perspective, assuming static participation, and overlooking that applicants may disengage from systems they perceive as unfair. Spillover effects, where decisions about some applicants influence the behavior of others, are also often ignored [10,11]. Satisfying group-level fairness is thus no guarantee that the outcomes are perceived as fair by individual applicants [4,15].

Against this background, we ask: *to what extent do lending decisions lead to fairer outcomes over time when considering perceived fairness, feature dynamics, and population dynamics?* We address this question by evaluating long-term fairness under endogenous participation driven by perceived fairness and peer effects.

Building on work on *perceived fairness* in psychology and algorithmic decision-making [13,14], we introduce a variation of this notion that captures both individual experience and social observation within a multi-agent simulation. *We define perceived (un)fairness as the extent to which applicants perceive outcomes as favorable to themselves and their peer group.* We introduce a fixed social network through which applicants can observe peer outcomes. Applicants may opt out following repeated unjust rejections or when they observe that their demographic group is denied loans more often than other group members. In the lending context, the applicant's opt-out behavior may refer to switching to competing lenders, which could reduce the amount of customers and representative data for the bank's future decision models. This creates population dynamics with opt-out behavior resulting from perceived unfairness. To evaluate these dynamics, we propose retention rate and retention disparity among demographic groups as metrics that capture differential opt-out.

We evaluate our framework in a synthetic lending setting [7], comparing demographic parity, equal opportunity, and causal fairness. Our results show that perceived unfairness can induce selective opt-out among disadvantaged groups, reducing measured unfairness through population change rather than improved treatment. This "survivorship bias"[5] also occurs in baseline conditions without fairness metrics. While demographic parity and equal opportunity reduce immediate retention disparities, they do not guarantee long-term fairness. In contrast, the causal fairness model [7] achieves balanced retention and lower long-term unfairness. These findings highlight trade-offs between accuracy, retention, and long-term fairness that are not observable in static settings. Our work emphasizes the need to jointly evaluate fairness constraints with endogenous participation dynamics to ensure that fairness interventions serve the citizens they are designed to protect.

2 Method

In this section, we describe the lending environment and social network (2.1), the perceived fairness state and opt-out behavior (2.2), the bank's decision model (2.3), and the evaluation metrics (fairness and retention) (2.4).

2.1 Bank Loan Simulation and Social Network

Each applicant i has a binary protected attribute $S \in \{0, 1\}$ (e.g., age), a vector of dynamic features $\mathbf{X}_t \in \mathbb{R}^d$ (e.g., credit risk score at time t), and a ground-truth label $Y_{i,t} \in \{-1, 1\}$ indicating whether the applicant can pay back ($Y_{i,t} = 1$) or default ($Y_{i,t} = -1$). At each time step, the lender makes a binary loan decision $D_{i,t} \in \{-1, 1\}$, where $D_{i,t} = 1$ denotes approval and $D_{i,t} = -1$ denial. Each applicant then chooses the action $A_{i,t} \in \{0, 1\}$ to apply ($A_{i,t} = 1$) or opt out ($A_{i,t} = 0$) in the lending process. Our synthetic datapoint has $|\mathbf{X}_1| = 2$ non-protected features sampled from S-specific Gaussian distributions and Y sampled from a ground-truth model. The detailed generation process for further steps is given in Appendix A.1.

Instead of simulating the behavior of individuals in isolation, we model how fairness is mediated through sequential social interactions. Applicants are embedded in a fixed, undirected social network. This can take spillover effects into account, where one applicant's rejection can influence another applicant's decision to participate [11]. Each applicant is connected to a predefined set of neighbors $\mathcal{N}_i$ with degree $n = 10$, where 80% of neighbors share the same protected attribute ($S_i = S_j = s, j \in \mathcal{N}_i$) and 20% of neighbors belong to the other demographic group ($S_i = s, S_j = s', j \in \mathcal{N}_i$). We choose this ratio to reflect the tendency for individuals to connect more often with similar individuals.

2.2 Perceived Fairness and Opt-out Behavior

Research on perceived fairness highlights two dimensions: whether individuals view their own outcomes as just and whether they perceive the process as legitimate relative to others' outcomes [13,14]. Inspired by this, we model perceived fairness as a deterministic belief state formed by applicants, unobservable by the lender, and separate from the fairness metrics used for evaluation. Furthermore, we assume that applicants have full knowledge of their own application details and history but observe only the decision outcomes of their peers.

At each time step t, an applicant i observes the outcomes of its connected neighbors $\mathcal{N}_i$ to form a peer-outcome signal $O_{i,t} \in \{0, 1\}$. $\mathcal{N}_{i,t}^{\text{act}} \subseteq \mathcal{N}_i$ denotes the subset of neighbors who are active at time t. The applicant computes the observed rejection rate for each group as $r_s = \frac{\sum_{j \in \mathcal{N}_{i,t}^{\text{act}},\, S_j = s} \mathbb{1}[D_{j,t} = -1]}{|\{j \in \mathcal{N}_{i,t}^{\text{act}} : S_j = s\}| + \epsilon}$. In this equation, ϵ in the denominator is a small constant introduced to prevent division by zero when agent i has no active neighbors of group s. The peer-outcome signal $O_{i,t} = 1$ if $r_{S_i} > r_{1-S_i}$, otherwise $O_{i,t} = 0$. This signal reflects whether the applicant's in-group experiences a higher rejection rate than the out-group, generating a social spillover effect that influences perceived fairness. Let $P_{i,t} = D_{i,t-1}$ store the last decision received from the bank.

We define perceived unfairness as $U_{i,t} = \mathbf{1}[D_{i,t} = -1 \wedge ((P_{i,t} = -1 \wedge Y_{i,t} = 1) \vee O_{i,t} = 1)]$. The first condition represents an applicant who can pay back but is unjustly denied a loan in two consecutive time steps, while the second captures if the in-group rejection rate exceeds the out-group rejection rate for a

rejected agent. The agent opts out $A_{i,t+1} = 0$ if it perceives unfairness ($U_{i,t} = 1$), or because it is correctly identified as fraudulent ($D_{i,t} = Y_{i,t} = -1$). Figure 1 demonstrates an example of three consecutive simulation steps. A_5 opts out at t_2 because it was correctly identified as fraud, which does not trigger perceived unfairness. A_3 opts out at t_3 because it was denied a loan in two consecutive time steps, and A_1 will opt out at the next step (t_4) due to spillover effects.

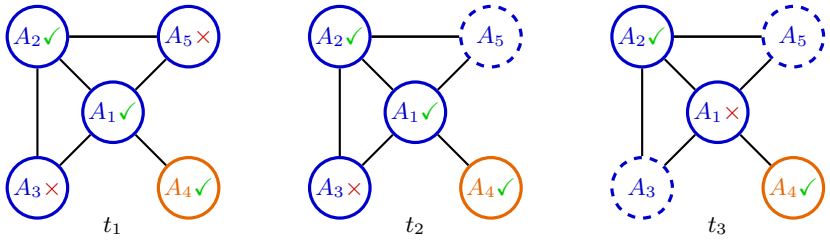

Fig. 1: Blue and orange agents denote different protected groups. Dashed circles represent $A_{i,t} = 0$. ✓ means application approved, × means application rejected.

2.3 Bank's Decision Model

We evaluate three fairness constraints. *Demographic Parity (DP)* requires equal approval rates between groups defined by the protected attribute, *Equal Opportunity (EO)* requires equal true positive rates between groups, and *Long-term Counterfactual Fairness (LCF)* requires that an agent's approval probability at a future horizon t^* be invariant to the protected attribute. The bank's decision model is optimized using Algorithm 1. The bank optimizes prediction accuracy, while setting either DP, EO or LCF as upper-bounded constraints. These are compared against a baseline model without fairness constraints.

2.4 Evaluation metrics

We evaluate fairness from two perspectives. From the lender's perspective, we measure *accuracy* (true positive rate at each step) and *t-step (long-term)*. We adopt the t-step definition from [7], which captures whether cumulative feature trajectories create divergent outcomes between demographic groups over time, where 0 is perfectly fair. To capture the population dynamics, we introduce three retention metrics: the *retention rate* $Rt = \frac{1}{N}\sum_{i=1}^{N} A_{i,t}$, the *retention disparity* $\Delta t = |R_t^{S=0} - R_t^{S=1}|$ which captures the differential opt-out between the groups, and the *relative retention ratio* $\rho = \frac{R_t^{S=0}}{R_t^{S=1}}$ where values below 1 indicate disproportionate opt-out for the protected group.

Algorithm 1 Bank's Decision Model

Input: $X = [S, \mathbf{X}] \in \mathbb{R}^{n \times (d+1)}, Y \in \{-1, 1\}^n$
Parameters: Weight vector $\mathbf{w} = (w_s, w_1, \ldots, w_d)$, intercept c, logistic function σ
for each applicant i at time t **do**
$p_{i,t} \leftarrow \sigma(w_s s_i + \sum_{j=1}^{d} w_j X_{i,t}^j + c)$,
$D_{i,t} \leftarrow 2\,\mathbf{1}[p_{i,t} > 0.5] - 1$
end for
Solve: $\min_{\mathbf{w},\, c}\ \mathcal{L}(\mathbf{w}, X)$ subject to $\mathcal{C}(\mathbf{w}, X) \leq \tau$
Optional Fairness constraints:
DP: $| P(D_t{=}1 \mid S{=}0) - P(D_t{=}1 \mid S{=}1) | \leq \tau$
EO: $| P(D_t{=}1 \mid Y_t{=}1, S{=}0) - P(D_t{=}1 \mid Y_t{=}1, S{=}1) | \leq \tau$
LCF: $| P(D_{t^*} = 1 \mid S = s, X = x) - P(D_{t^*} = 1 | S = s', X = x) | \leq \tau$
return Optimized $\mathbf{w}^*, c^*$

3 Preliminary Results

Accuracy The baseline achieves the highest accuracy ($t_5 = 0.734$) across all time steps compared to the fairness-aware approaches. EO has a slightly lower accuracy ($t_5 = 0.690$) followed by LCF ($t_5 = 0.683$) and DP ($t_5 = 0.685$), suggesting a modest accuracy trade-off when optimizing with fairness constraints. Detailed results for comparisons between different methods that evaluate accuracy, long-term unfairness, retention rate, and retention disparity are provided in the Appendix A.2.

Long-term fairness t-step Figure 2 shows that LCF achieves the lowest cumulative unfairness across all time steps, demonstrating that incorporating multi-step counterfactual reasoning into the fairness constraint [7] remains effective even under population dynamics. In contrast, DP and EO constraints enforce group fairness at each step but do not account for the cumulative feature trajectories induced by their decisions. The resulting mismatch accumulates over time, resulting in both DP and EO exceeding the long-term unfairness of the baseline model without population dynamics, consistent with findings in [8]. When comparing the simulation with and without population dynamics in Figure 2, the baseline and EO exhibit higher unfairness in a static setting when the full population is retained. DP shows a higher unfairness with population dynamics from step 4. LCF shows almost no difference between the settings, demonstrating that long-term fairness does not depend on selective opt-out of disadvantaged groups.

Retention rates Figure 3 shows that the baseline condition exhibits a severe retention disparity between groups, with the protected group ($S = 0$) leaving at a higher rate. This indicates a survivorship bias where long-term fairness seems to improve because disadvantaged applicants left the system [5]. The relative retention ratio (right axis) confirms this: the baseline ratio decreases and remains below 1.0. While DP and EO reduce this gap, they do so inconsistently: EO narrows but does not eliminate the imbalance, while DP reverses the disparity so that the unprotected group leaves faster (relative retention ratio above 1.0).

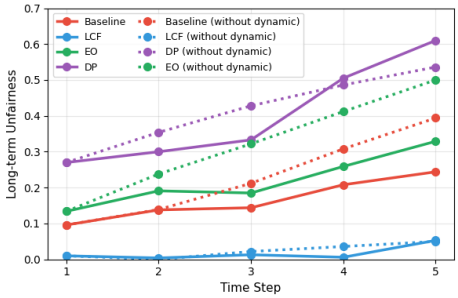

Fig. 2: Long-term unfairness with (solid) and without (dotted) population dynamics for each fairness metric. LCF maintains near-zero unfairness in both settings; DP has the highest long-term unfairness.

The LCF model achieves near equal retention with a ratio close to 1.0 throughout while maintaining counterfactual fairness.

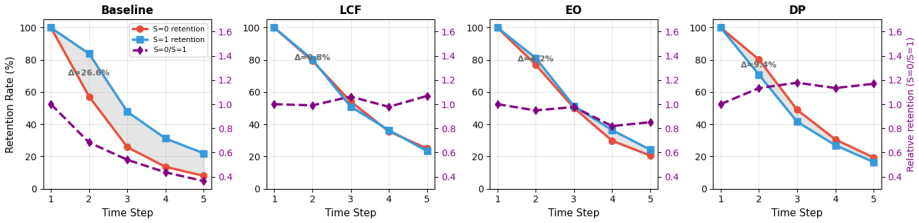

Fig. 3: Retention rates per group (left-axis) for $S = 0$ (red) and $S = 1$ (blue); relative retention ratio (right axis) for $S = 0/S = 1$ (purple dashed), values below 1.0 indicate disproportionate opt-out for $S = 0$. Baseline shows severe retention disparity ($\Delta = 26.6\%$), LCF achieves near parity ($\Delta = 0.8\%$) and DP reverses disparity.

4 Conclusions and Future Work

In this work, we evaluate long-term fairness under endogenous participation driven by perceived fairness and peer effects. We introduce perceived fairness as an internal belief state that captures past experience and social observation. Unlike prior studies that model feature evolution but assume static populations [3,8,12], our framework integrates population dynamics driven by perceived fairness, exposing trade-offs between accuracy, short-term fairness, and long-term fairness. While the unconstrained baseline achieves the highest accuracy, it shows high retention disparities, driven by selective opt-out of the protected group.

Demographic parity and equal opportunity reduce the immediate retention disparity, but do not guarantee long-term fairness. Demographic parity overcorrects retention in earlier time steps, resulting in reverse long-term disparities. In contrast, the causal fairness model consistently achieves low long-term unfairness with balanced retention rates between groups. Our findings emphasize the importance of a citizen-centric perspective in evaluating algorithmic decision-making: when fairness interventions fail to account for how citizens perceive and react to decisions, this may inadvertently drive the most vulnerable individuals out of essential public services, undermining the fairness goals they aim to achieve.

To build on this workshop paper, we plan to validate our framework on real-world datasets and study additional fairness metrics (e.g., predictive equality). The current results are derived under the assumption of a fixed network degree and group-mixing ratios. Future work will address these constraints by performing a sensitivity analysis on neighborhood size and the proportion of connections between demographic groups. Moreover, we can extend this work by generalizing the two-step denial memory to variable lengths, modeling dynamic social networks where peer groups evolve over time, and incorporating positive spillover effects arising from favorable outcomes. Additional directions include introducing competing lenders to capture push-pull dynamics and modeling the lender as a learning agent that optimizes long-term objectives, such as retention and perceived fairness.

References

1. Barocas, S., Hardt, M., Narayanan, A.: Fairness and machine learning. Recommender systems handbook **1**, 453–459 (2020)
2. Barocas, S., Selbst, A.D.: Big data's disparate impact. Calif. L. Rev. **104**, 671 (2016)
3. D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., Halpern, Y.: Fairness is not static: deeper understanding of long term fairness via simulation studies. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 525–534 (2020)
4. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R.: Fairness through awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. pp. 214–226 (2012)
5. Gupta, P., MacAvaney, S.: On survivorship bias in ms marco. In: proceedings of the 45th International ACM SIGIR conference on research and development in information retrieval. pp. 2214–2219 (2022)
6. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
7. Hu, Y., Zhang, L.: Achieving long-term fairness in sequential decision making. Proceedings of the AAAI Conference on Artificial Intelligence **36**(9), 95499557 (Jun 2022). https://doi.org/10.1609/aaai.v36i9.21188
8. Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., Hardt, M.: Delayed impact of fair machine learning. In: International Conference on Machine Learning. pp. 3150–3158. PMLR (2018)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
10. Mehrotra, A., Sachs, J., Celis, L.E.: Revisiting group fairness metrics: The effect of networks. Proceedings of the ACM on Human-Computer Interaction **6**(CSCW2), 1–29 (2022)
11. Narayanan, A.: What if algorithmic fairness is a category error? Contemporary Debates in the Ethics of Artificial Intelligence pp. 77–95 (2025)
12. Rateike, M., Valera, I., Forré, P.: Designing long-term group fair policies in dynamical systems. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 20–50 (2024)
13. Tyler, T.R.: Social justice: Outcome and procedure, 35 intl j. Psychol **117**, 119–20 (2000)
14. Wang, R., Harper, F.M., Zhu, H.: Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. p. 114. CHI '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376813, https://doi.org/10.1145/3313831.3376813
15. Zhou, W., et al.: Group vs. individual algorithmic fairness (2022)

A Appendix

A.1 Synthetic Dataset

We generate a 5-step synthetic dataset with 4000 agents for training and 1000 for testing. At $t = 1$, S is sampled with equal probability, and $\mathbf{X}_1$ is sampled from S-specific Gaussian distributions. The true repayment is sampled from a ground-truth decision model $Y_t = \sigma(h_{\theta_*}(\cdot))$, where $\sigma(\cdot)$ is the sigmoid function, $h_{\theta_*}(\cdot)$ is a fixed mapping of the probability of whether an applicant would default $P(Y_t) = \sigma(h_{\theta^*}(\mathbf{X}_t, S))$, $Y_t \sim 2 \cdot \text{Bernoulli}(P(Y_t)) - 1$. D_t is sampled from a separate $\sigma(h_{\theta_*}(\cdot))$ as $D_t \sim 2 \cdot \text{Bernoulli}(P(D_t)) - 1$. $\mathbf{X}_{i,t+1}$ is generated according to the update rule below:

$$\mathbf{X}_{i,t+1} = \begin{cases} \mathbf{X}_{i,t} - \lambda \cdot \theta_t + b & D_{i,t} = 1, Y_{i,t} = -1 \\ \mathbf{X}_{i,t} + \lambda \cdot \theta_t + b & D_{i,t} = 1, Y_{i,t} = 1 \\ \mathbf{X}_{i,t} + b & D_{i,t} = -1 \end{cases} \tag{1}$$

where λ controls the sensitivity of the update to the predicted decisions, and $b = S \cdot b_1 + (1 - S) \cdot b_0$ is a small increment at each time step. The parameters are set as $\lambda = 0.5$, $b_0 = 0.2$, $b_1 = 1.0$.

A.2 Detailed Results Table

Table 1 lists the results of the evaluation metrics in our simulation.

Table 1: Accuracy, long-term unfairness (t-step), retention rate, and retention disparity across fairness metrics over 5 time steps.

Metric	Fairness Metric	t_1	t_2	t_3	t_4	t_5
Accuracy (%)	Baseline	70.4	73.5	72.5	71.8	73.4
	LCF	68.5	70.1	69.9	68.7	68.3
	EO	68.6	69.2	69.9	68.6	69.0
	DP	67.7	67.4	68.1	68.4	68.5
t-step	Baseline	0.096	0.138	0.144	0.208	0.244
	LCF	0.010	0.004	0.013	0.006	0.053
	EO	0.134	0.191	0.185	0.259	0.329
	DP	0.270	0.300	0.333	0.505	0.610
Retention (%)	Baseline	100	70.5	36.8	22.4	15.0
	LCF	100	79.8	52.6	36.0	24.2
	EO	100	79.1	50.8	33.1	22.4
	DP	100	75.5	45.3	28.8	18.0
Disparity (%)	Baseline	0.0	26.6	22.0	17.6	14.0
	LCF	0.0	0.8	3.2	0.8	1.6
	EO	0.0	4.2	1.2	6.6	3.6
	DP	0.0	9.4	7.4	3.6	2.8