Temporal Dynamics of Qualification Rates: Expanding on POMDP-based Fairness Frameworks

Adi Levi
The Faculty of Data and Decisions Science | Technion
Strategic and Societal Aspects of Machine Learning
30/08/23

Abstract

This study elaborates on and extends the work of [1], which employed the POMDP framework to investigate how qualification rates evolve under the influence of institutional decisions and individual actions. The original research observe the long-term implications of imposing various fairness constraints on diverse subgroups within a population. It concluded that short-term fairness measures may not necessarily translate into long-term equity and that the outcomes can differ depending on the underlying feature distributions and transition dynamics. In an effort to deepen the understanding of these results, this paper present the sketch proofs of [1]'s core findings. Additionally, we introduce an expanded model that incorporates the effects of external interventions, such as community programs and mentoring initiatives, on qualification states. This enhancement provides a more nuanced and comprehensive perspective on the temporal dynamics of qualification rates.

1 Summary

1.1 The Problem

Building on prior research in the field [1] this paper seeks to understand the long-term impact of algorithmic (fair) decisions on the welfare of different subgroups in a population. Specifically, it examines whether or when short-term fairness constraints are promoting long-term equality. Distinguishing itself from previous works, this paper extends the focus to sequential frameworks, while addressing their limitations by considering, how individuals may strategically adapt to policies as well as, how the sensitive attributes can influence these dynamics. The study also developed ways to handle unobservable qualification states and differing dynamics across groups, making it a more realistic representation of the world.

1.2 The Model

The study employs a Partially Observed Markov Decision Process (POMDP) which is a sequential decision-making model with discrete time intervals. In this approach, when basing decisions on an individual's current attributes and features, previous decisions and features are not considered. The decision-maker evaluates the current features of individuals, represented by X_t, against variables of interest, such as the (unobserved) qualification state, Y_t. The goal is to make a decision, D_t, that maximizes immediate utility while adhering to specific constraints. After a decision is made, individuals then decide how much effort, T^s_{yd}, to invest to either maintain or enhance their qualification status for the upcoming stage, Y_{t+1}. This effort not only influences their next-stage features, X_{t+1} but also modifies the overall qualification rates of the population, α^s_t. In other words, the system outlines how qualification rates evolve over time and are influenced by the decisions taken and the individuals' subsequent responses.

1.3 Methods

The following equation captures the transition dynamics of qualification rates over time of both groups (a and b):

$$\alpha_{t+1}^{s} = g^{0s}\ (\alpha_t^a, \alpha_t^b\ ) \cdot (1-\alpha_t^s) + g^{1s}(\alpha_t^a, \alpha_t^b) \cdot \alpha_t^s, s \in \{a, b\} \tag{1}$$

The first and second terms capture the expected qualification rate of unqualified and qualified individuals at the next time step respectively. The dynamics system can reach equilibrium if $\alpha_{t+1}^s = \alpha_t^s$ holds. Therefore, the system has equilibrium if there exists a solution to the **balanced equations** defined as:

$$\forall s \in \{a, b\} : \frac{1}{\alpha^s} - 1 = \frac{1 - g^{1s}\left(\theta^s\left(\alpha^a, \alpha^b\right)\right)}{g^{0s}\left(\theta^s\left(\alpha^a, \alpha^b\right)\right)} \tag{2}$$

The equilibrium of this dynamic system can provide insight into the long-term population properties and is highly dependent on the different transition probabilities that specify different user dynamics. This paper focuses on the 2 following user dynamics:

Lack of motivation effect - $T_{01}^s \leq T_{00}^s\ and\ T_{11}^s \leq T_{10}^s$: Individuals that were accepted by the institute are **less** likely to remain\become qualified than individuals that were rejected by the institute. Meaning acceptance **decreases** the motivation to remain\become qualified.

Leg up effect- $T_{01}^s \geq T_{00}^s\ and\ T_{11}^s \geq T_{10}^s$: Individuals that were accepted by the institute are **more** likely to remain\become qualified than individuals that were rejected by the institute. Meaning acceptance **boosts** the motivation to remain\become qualified.

Lemma 1: Let $(\gamma^a(x), \gamma^b(x))$ be a pair of qualification profiles for groups $\mathcal{G}_a$ and $\mathcal{G}_b$ at t. Let threshold pairs $(\theta_{UN}^{a^*}, \theta_{UN}^{b^*})$ and $(\theta_C^{a^*}, \theta_C^{b^*})$ be the unconstrained and fair optimal thresholds under constraint C, respectively. Then we have $\gamma^a(\theta_{UN}^{a^*}) = \gamma^b(\theta_{UN}^{b^*}) = \frac{u_-}{u_+ + u_-}$ and

$$p_a\gamma^a\left(\theta_{DP}^{a^*}\right) +\ p_b\gamma^b\left(\theta_{DP}^{b^*}\right) = \frac{u_-}{u_+ +\ u_-}; \frac{p_a\alpha^a}{\gamma^a(\theta_{EqOpt}^{a^*})} + \frac{p_b\alpha^b}{\gamma^b(\theta_{EqOpt}^{b^*})} = \frac{p_a\alpha^a}{\frac{u_-}{u_+ +\ u_-}} + \frac{p_b\alpha^b}{\frac{u_-}{u_+ +\ u_-}} \tag{3}$$

Lemma 1 describes how we can find the optimal fair threshold.

Theorem 1: Consider a dynamic (1) with a threshold policy $\theta^s\left(\alpha^a,\ \alpha^b\right)$ that is continuous in α^a and α^b. $\forall T_{yd}^s\ \in (0,1)$, there exists at least one equilibrium $\left(\hat{\alpha}^a, \hat{\alpha}^b\right)$.
Theorem 1 asserts the existence of at least one equilibrium for the given dynamical system under a threshold policy that is continuous with respect to the qualification rates of both groups.

Given that the optimal policy is a threshold policy, and that the system has equilibrium, then:

$$g^{ys}\ (\alpha^a, \alpha^b\ ) = T_{y0}^s \cdot \mathbb{G}_y^s\ (\theta^s\ (\alpha^a, \alpha^b\ )) + T_{y1}^s \cdot (1 - \mathbb{G}_y^s\ (\theta^s\ (\alpha^a, \alpha^b\ ))) \tag{4}$$

Therefore, we get:

$$h^s(\theta^s\ (\alpha^a, \alpha^b\ )) := \frac{1 - g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))} = \frac{1 - (T_{10}^s \cdot \mathbb{G}_1^s(\theta^s\ (\alpha^a, \alpha^b\ )) + T_{11}^s \cdot (1 - \mathbb{G}_1^s(\theta^s\ (\alpha^a, \alpha^b\ ))))}{T_{00}^s \cdot \mathbb{G}_0^s(\theta^s\ (\alpha^a, \alpha^b\ )) + T_{01}^s \cdot (1 - \mathbb{G}_0^s(\theta^s\ (\alpha^a, \alpha^b\ )))} \tag{5}$$

In order to be able to compare the long-term impacts of different fairness constraints on the qualification rates, we limit the number of possible equilibria by assuming that there is a unique equilibrium.

Theorem 2: Consider a decision-making system with dynamics (1) and either unconstrained or fair optimal threshold policy. Let $h^s\left(\theta^s\left(\alpha^a,\ \alpha^b\right)\right) := \frac{1\ - g^{1s}\left(\theta^s\left(\alpha^a,\ \alpha^b\right)\right)}{g^{0s}\left(\theta^s\left(\alpha^a,\ \alpha^b\right)\right)}$, $s \in a, b$. Under Assumptions 1 and 2, a sufficient condition for (1) to have a unique equilibrium is as follows, $\forall s \in a, b$:

- **Under lack of motivation**, $\left|\frac{\partial h^s\left(\theta^s\left(\alpha^a,\ \alpha^b\right)\right)}{\partial\alpha^{-s}}\right| < 1, \forall\alpha^s \in [0,1]$ where $-s := \{a, b\} \setminus s$
- **Under Leg up**, $\left|\frac{\partial h^s\left(\theta^s\left(\alpha^a,\ \alpha^b\right)\right)}{\partial\alpha^{-s}}\right| < 1$ and $\left|\frac{\partial h^s\left(\theta^s\left(\alpha^a,\ \alpha^b\right)\right)}{\partial\alpha^{s}}\right| < 1, \forall\alpha^a,\ \alpha^b \in [0,1]$

1.4 Key Findings

The paper discovers that short-term fairness interventions may not lead to long-term equity, and the effects can vary based on feature distributions and transitions. Moreover, a small change in either can lead to contrarian results.

Natural Equality: Both groups attain similar qualification rates without external fairness interventions, i.e., $\hat{\alpha}^a_{UN} = \hat{\alpha}^b_{UN}$.

Theorem 3: For any feature distribution $G^s_y(x)$ and $\forall \alpha_{UN} \in (0,1)$, there exist transitions $\left\{T^s_{yd}\right\}_{y,d,s}$ satisfying either lack of motivation effect or Leg up effect such that $\hat{\alpha}^a_{UN} = \hat{\alpha}^b_{UN} = \alpha_{UN}$. In this case, if $G^a_y(x) \neq G^b_y(x)$ (resp. $G^a_y(x) = G^b_y(x)$), then imposing either C =DP or EqOpt fair optimal policies will violate (resp. maintain) equality, i.e.,$\hat{\alpha}^a_C \neq \hat{\alpha}^b_C$ (resp. $\hat{\alpha}^a_C = \hat{\alpha}^b_C$).

Theorem 3 shows that for any equilibrium α_{UN}, there exist model parameters that ensure both groups converge to this equilibrium. If both groups have different feature distributions, applying fairness constraints like DP or EqOpt could exacerbate inequality. Given that feature distributions between groups often differ, applying fairness interventions might do more harm than help.

Natural inequality is more prevalent than natural equality, primarily because $G^s_y(x)$ and T^s_{yd} typically differ across groups. Thus, addressing fairness requires understanding the roots of inequality.

Theorem 4: Under the assumption that $G^a_y(x) = G^b_y(x)$, $T^a_{yd} \neq T^b_{yd}$ we have that:
Under lack of motivation effect, DP and EqOpt fairness exacerbate inequality, i.e., $\left|\hat{\alpha}^a_C - \hat{\alpha}^b_C\right| \geq \left|\alpha^a_{UN} - \hat{\alpha}^b_{UN}\right|$; Under Leg up effect, DP and EqOpt fairness mitigate inequality, i.e., $\left|\hat{\alpha}^a_C - \hat{\alpha}^b_C\right| \leq \left|\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN}\right|$. Moreover, the disadvantaged group remains disadvantaged in both cases, i.e., $\left(\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN}\right)\left(\hat{\alpha}^a_C - \hat{\alpha}^b_C\right) \geq 0$.

This theorem shows that under identical feature distributions, but differing transition dynamics, fairness constraints can either mitigate or worsen inequality, depending on whether the “leg-up” effect or “lack of motivation” effect is predominant. Only when the “leg-up” effect is more prominent than the “lack of motivation” effect then imposing fairness can reduce inequality.

Condition 2: Under the assumption that $\frac{G^a_0(x)}{G^b_0(x)}$ is strict increasing in x, $G^a_1(x) = G^b_1(x), \forall x$ and $T^a_{yd} = T^b_{yd}$, we will define $\hat{x}$ as the x for which $G^a_0(\hat{x}) = G^b_0(\hat{x})$ holds.

Theorem 5: Under condition 2 and the leg up effect, if $\frac{u_+}{u_-} \geq \frac{G^s_0(\hat{x})}{G^s_1(\hat{x})}\frac{(1-T_{10})}{T_{00}}$, we have:

- $\hat{\alpha}^a_{UN} > \hat{\alpha}^b_{UN}$ and $\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN} > \hat{\alpha}^a_{EqOpt} - \hat{\alpha}^b_{EqOpt} \geq 0$ hold, i.e., EqOpt fairness always mitigates inequality and the disadvantaged group $\mathcal{G}_b$ remains disadvantaged.
- DP fairness may either (1) mitigate inequality, i.e., $\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN} > \hat{\alpha}^a_{DP} - \hat{\alpha}^b_{DP} \geq 0$; or (2) flip the disadvantaged group from $\mathcal{G}_b$ to $\mathcal{G}_a$, i.e., $\hat{\alpha}^b_{DP} \geq \hat{\alpha}^a_{DP}$.

There’s a balance point that defines the impact of fairness interventions, $\frac{G^s_0(\hat{x})}{G^s_1(\hat{x})}\frac{(1-T_{10})}{T_{00}}$, derived from both current qualifications and future transitions. It compares the benefit of accepting a qualified individual against the cost of accepting an unqualified one. Theorem 5 highlights that if the benefits of accepting a qualified individual far outweigh the costs of accepting an unqualified one, EqOpt will always reduce inequality, while, DP might either reduce inequality or flip the disadvantaged group.

1.5 Limitations

To capture individuals’ abilities to improve/maintain future qualifications they used the transition functions, T^s_{yd}. The properties of the equilibrium are highly dependent on these transitions.
Therefore, since their analysis and conclusions rely on this set of values, then the conclusions are highly sensitive to minor changes in these transitions. In practice, these transitions can be extremely hard to measure due to the complexity of human behaviors and environmental factors.

2 Implementation

In this section, we will provide a proof sketch of the main theorems - 3,4,5.
From the proof of Theorem 1, we get that the equilibrium $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$ is the intersection of $\mathcal{C}_1$ and $\mathcal{C}_2$.

From the proof of Theorem 2, we have the following results;

1. Under the Lack of motivation effect $\psi_C^a\left(\alpha^b\right)$ and $\psi_C^b\left(\alpha^a\right)$ are non-increasing in α^b and α^a respectively. Therefore, to have an exact one intersection between $\mathcal{C}_1$ and $\mathcal{C}_2$, $-1 < \frac{\partial \psi_C^b(\alpha^a)}{\partial \alpha^a} \leq 0$ and $-1 < \frac{\partial \psi_C^a\left(\alpha^b\right)}{\partial \alpha^b} \leq 0$ must hold.
2. Under the Leg-up effect $\psi_C^a\left(\alpha^b\right)$ and $\psi_C^b\left(\alpha^a\right)$ are non-decreasing in α^b and α^a respectively. Therefore, to have an exact one intersection between $\mathcal{C}_1$ and $\mathcal{C}_2$, $0 \leq \frac{\partial \psi_C^b(\alpha^a)}{\partial \alpha^a} < 1$ and $0 \leq \frac{\partial \psi_C^a\left(\alpha^b\right)}{\partial \alpha^b} < 1$ must hold.
3. h^s is non-increasing in α^s under the Leg-up effect and non-decreasing in α^s under the Lack of motivation effect.
4. h^s is non-decreasing in θ^s under the Leg-up effect and non-increasing in θ^s under the Lack of motivation effect.
5. $g^{1s}\left(\theta_{UN}^s\left(\alpha\right)\right)$ and $g^{0s}\left(\theta_{UN}^s\left(\alpha\right)\right)$ are the convex combination of T_{11}^s, T_{10}^s and T_{01}^s, T_{00}^s respectively. Therefore, we have that $\forall \alpha^b, \alpha^a \in [0,1]$,
   under the Leg up effect: $0 < \frac{(1-T_{11})}{T_{01}} \leq h^s\left(\theta^s\left(\alpha^a, \alpha^b\right)\right) \leq \frac{(1-T_{10})}{T_{00}} < \infty$.
   under the Lack of motivation effect: $0 < \frac{(1-T_{10})}{T_{00}} \leq h^s\left(\theta^s\left(\alpha^a, \alpha^b\right)\right) \leq \frac{(1-T_{11})}{T_{01}} < \infty$.

2.1 Theorem 3

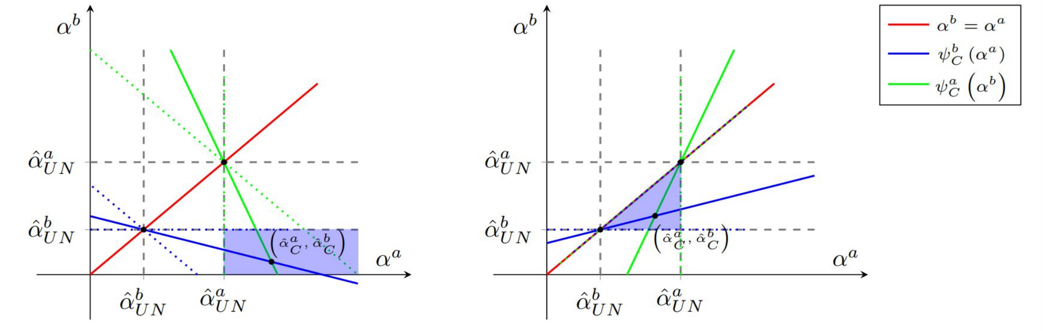
1. Show that $\forall \mathbb{G}_y^s\left(x\right), \alpha_{UN}, s \in \{a, b\}$ an equitable equilibrium is attained, i.e., $\hat{\alpha}_{UN}^s = \alpha_{UN}$:
   - The equilibrium is a convex combination of both T_{00}^s, T_{01}^s and T_{10}^s, T_{11}^s from the balanced equation (2), i.e.,
   $$\begin{aligned}\hat{\alpha}_{UN}^s = T_{11}^s \cdot \left(1 - \mathbb{G}_1^s\left(\theta_{UN}^s\left(\hat{\alpha}_{UN}^s\right)\right)\right) + T_{10}^s \cdot \mathbb{G}_1^s\left(\theta_{UN}^s\left(\hat{\alpha}_{UN}^s\right)\right) = \\ T_{00}^s \cdot \mathbb{G}_0^s\left(\theta_{UN}^s\left(\hat{\alpha}_{UN}^s\right)\right) + T_{01}^s \cdot \left(1 - \mathbb{G}_0^s\left(\theta_{UN}^s\left(\hat{\alpha}_{UN}^s\right)\right)\right)\end{aligned} \quad (6)$$
   - From the properties of convex combinations we get , $\forall \alpha_{UN}$, $\exists T_{00}^s, T_{01}^s$ and $\exists T_{10}^s, T_{11}^s$ that bounds α_{UN} between them such that (6) holds with $\hat{\alpha}_{UN}^s = \alpha_{UN}, \forall s \in \{a, b\}$.
2. Show that if $G_y^a\left(x\right) \neq G_y^b\left(x\right), \forall y \in \{0,1\}$, then $\hat{\alpha}_C^a \neq \hat{\alpha}_C^b$:
   - If $\tilde{\alpha}_C^b \neq \tilde{\alpha}_C^a$, then the intersection of $\mathcal{C}_1$ and $\mathcal{C}_2$ (i.e., $(\hat{\alpha}_C^a, \hat{\alpha}_C^b)$) is not on the line $\alpha^b = \alpha^a$, because both $\mathcal{C}_1$ and $\mathcal{C}_2$ are non-increasing\non-decreasing together.
   - If $\alpha^a = \alpha^b = \alpha_{UN}$, then $\frac{G_0^a(\theta_{UN}^a(\alpha_{UN}))}{G_1^a\left(\theta_{UN}^a(\alpha_{UN})\right)} = \frac{G_0^b\left(\theta_{UN}^b(\alpha_{UN})\right)}{G_1^b\left(\theta_{UN}^b(\alpha_{UN})\right)}$ because $\gamma^a\left(\theta_{UN}^a\left(\alpha_{UN}\right)\right) = \gamma^b\left(\theta_{UN}^b\left(\alpha_{UN}\right)\right) = \frac{u_-}{u_+ + u_-}$ according to lemma 1. Therefore, if $G_y^a\left(x\right) \neq G_y^b\left(x\right)$ then $\theta_{UN}^s\left(\alpha_{UN}\right) \neq \theta_C^s\left(\alpha_{UN}, \alpha_{UN}\right)$ must hold. Specifically, there are only 2 possible options that satisfy equation 3;
     (a) $\theta_{UN}^a\left(\alpha_{UN}\right) > \theta_C^a\left(\alpha_{UN}, \alpha_{UN}\right)$ and $\theta_{UN}^b\left(\alpha_{UN}\right) < \theta_C^b\left(\alpha_{UN}, \alpha_{UN}\right)$
     (b) $\theta_{UN}^a\left(\alpha_{UN}\right) < \theta_C^a\left(\alpha_{UN}, \alpha_{UN}\right)$ and $\theta_{UN}^b\left(\alpha_{UN}\right) > \theta_C^b\left(\alpha_{UN}, \alpha_{UN}\right)$
   - WOLG, If (a) hold, then from result 4,under the Leg-Up effect, $h^b\left(\theta_{UN}^b\left(\alpha_{UN}\right)\right) < h^b\left(\theta_C^b\left(\alpha_{UN}, \alpha_{UN}\right)\right)$ and $h^a\left(\theta_{UN}^a\left(\alpha_{UN}\right)\right) > h^a\left(\theta_C^a\left(\alpha_{UN}, \alpha_{UN}\right)\right)$. From result 3 we get $\tilde{\alpha}_C^b < \hat{\alpha}_{UN}^b = \hat{\alpha}_{UN}^a < \tilde{\alpha}_C^a$. Similarly, we can show that under Lack of motivation effect $\tilde{\alpha}_C^b > \hat{\alpha}_{UN}^b = \hat{\alpha}_{UN}^a > \tilde{\alpha}_C^a$. Therefore, $\tilde{\alpha}_C^b \neq \tilde{\alpha}_C^a$.
3. Show that if $G_y^a\left(x\right) = G_y^b\left(x\right), \forall y \in \{0,1\}$, then $\hat{\alpha}_C^a = \hat{\alpha}_C^b$:
   - If $\alpha^a = \alpha^b = \alpha$, then $\gamma^a\left(x\right) = \gamma^b\left(x\right)$ and $\mathcal{P}_C^a\left(x\right) = \mathcal{P}_C^b\left(x\right)$ $\forall C \in \{DP, EqOpt\}$, which implies $\theta_C^a\left(\alpha, \alpha\right) = \theta_C^b\left(\alpha, \alpha\right)$.
   - $\gamma^a\left(\theta_C^a\left(\alpha, \alpha\right)\right) = \gamma^b\left(\theta_C^b\left(\alpha, \alpha\right)\right) = \frac{u_-}{u_+ + u_-}$, according to the optimal fair policy equation (3). Therefore, $\gamma^s\left(\theta_{UN}^s\left(\alpha\right)\right) = \gamma^s\left(\theta_C^s\left(\alpha, \alpha\right)\right) = \frac{u_-}{u_+ + u_-}, \forall s \in \{a, b\} \Rightarrow \theta_{UN}^s\left(\alpha\right) = \theta_C^s\left(\alpha, \alpha\right)$ hold under any α.
   - Therefore, $h^s\left(\theta_{UN}^s\left(\alpha\right)\right) = h^s\left(\theta_C^s\left(\alpha, \alpha\right)\right)$, meaning $\hat{\alpha}_C^a = \hat{\alpha}_C^b$, because there is only one intersection.

2.2 Theorem 4

1. Show that $\hat{\alpha}^b_{UN} = \psi^b_C\left(\hat{\alpha}^b_{UN}\right)$ and $\hat{\alpha}^a_{UN} = \psi^a_C\left(\hat{\alpha}^a_{UN}\right)$ hold:

   - As we saw in the proof sketch of Theorem 3, when $G^a_y(x) = G^b_y(x)$ and $\alpha^a = \alpha^b = \alpha$ then, $\theta^s_{UN}(\alpha) = \theta^s_C(\alpha, \alpha)$ holds under any α.
   - Since $\hat{\alpha}^s_{UN}$ is the solution to balanced equation 2, then $l^s\left(\hat{\alpha}^s_{UN}\right) = h^s\left(\theta^s_{UN}\left(\hat{\alpha}^s_{UN}\right)\right) = h^s\left(\theta^s_C\left(\hat{\alpha}^s_{UN}, \hat{\alpha}^s_{UN}\right)\right)$, which implies $\hat{\alpha}^s_{UN} = \psi^s_C\left(\hat{\alpha}^s_{UN}\right), \forall s \in \{a, b\}$.

2. From the results of Theorem 2, I will present a possible representation under both effects:



(a) Under Lack Of Motivation Effect

(b) Under Leg-Up Effect

   The dotted green and blue lines represent the range of $\psi^a_C\left(\alpha^b\right)$ and $\psi^b_C\left(\alpha^a\right)$ respectively under each of the effects according to Theorem 2. The blue area represents all possible values for the intersection of $\mathcal{C}_1$ and $\mathcal{C}_2$.

3. Show that under the Leg-Up effect $\left|\hat{\alpha}^a_C - \hat{\alpha}^b_C\right| \geq \left|\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN}\right|$:
   From (b) we can see that $\hat{\alpha}^a_C > \hat{\alpha}^b_C$, $\hat{\alpha}^b_{UN} < \hat{\alpha}^a_C < \hat{\alpha}^a_{UN}$ and $\hat{\alpha}^b_{UN} < \hat{\alpha}^b_C < \hat{\alpha}^a_{UN}$ as demonstrated from the blue area. Therefore we have,$\left|\hat{\alpha}^a_C - \hat{\alpha}^b_C\right| \geq \left|\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN}\right|$.

4. Show that under the Lack of motivation effect $\left|\hat{\alpha}^a_C - \hat{\alpha}^b_C\right| \leq \left|\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN}\right|$:
   From (a) we can see that $\hat{\alpha}^a_C > \hat{\alpha}^b_C$, $\hat{\alpha}^b_C < \hat{\alpha}^b_{UN}$ and $\hat{\alpha}^a_C > \hat{\alpha}^a_{UN}$ as demonstrated from the blue area. Therefore we have,$\left|\hat{\alpha}^a_C - \hat{\alpha}^b_C\right| \leq \left|\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN}\right|$.

2.3 Theorem 5

1. Show that if $\frac{u_+}{u_-} \geq \frac{(1-T_{10})}{T_{00}}\beta(\hat{x})$ under the Leg-Up effect, then $\hat{\alpha}^b_{UN} < \hat{\alpha}^a_{UN}$:

   - From condition 2, $\frac{G^a_0(x)}{G^b_0(x)}$ is strict increasing in x, $G^a_0(\hat{x}) = G^b_0(\hat{x})$ and $\forall x, G^a_1(x) = G^b_1(x)$. Therefore we have, $\forall x > \hat{x}, G^a_0(x) > G^b_0(x)$ and $\forall x < \hat{x}, G^a_0(x) < G^b_0(x)$. Thus if $\alpha^a = \alpha^b = \alpha$, then $\forall x < \hat{x}, \gamma^b(x) < \gamma^a(x)$ and $\forall x > \hat{x}, \gamma^b(x) > \gamma^a(x)$.
   - Let $\bar{\alpha}$ such that $\gamma^b(\hat{x}) = \gamma^a(\hat{x}) = \frac{1}{\beta(\hat{x})\left(\frac{1}{\alpha}-1\right)+1} = \frac{u_-}{u_+ + u_-}$ where $\beta(\hat{x}) := \frac{G^a_0(\hat{x})}{G^a_1(\hat{x})} = \frac{G^b_0(\hat{x})}{G^b_1(\hat{x})}$. From lemma 1 we have, $\gamma^a\left(\theta^a_{UN}(\alpha)\right) = \gamma^b\left(\theta^b_{UN}(\alpha)\right) = \frac{u_-}{u_+ + u_-}$ and since as α increase so as $\frac{1}{\beta(\hat{x})\left(\frac{1}{\alpha}-1\right)+1}$, we have, $\forall \alpha > \bar{\alpha}$, $\gamma^a\left(\theta^a_{UN}(\alpha)\right) = \gamma^b\left(\theta^b_{UN}(\alpha)\right) < \gamma^b(\hat{x}) = \gamma^a(\hat{x})$.
   - Since both $\theta^a_{UN}(\alpha)$ and $\theta^b_{UN}(\alpha)$ are smaller than $\hat{x}$ and $\gamma^a\left(\theta^a_{UN}(\alpha)\right) = \gamma^b\left(\theta^b_{UN}(\alpha)\right)$ then $\forall \alpha > \bar{\alpha}, \theta^a_{UN}(\alpha) < \theta^b_{UN}(\alpha) < \hat{x}$. Therefore, since CDF is always monotonic increasing we get $\forall y \in \{0,1\}, \mathbb{G}^a_y\left(\theta^a_{UN}(\alpha)\right) < \mathbb{G}^b_y\left(\theta^b_{UN}(\alpha)\right) \Rightarrow \forall \alpha > \bar{\alpha}$,
     $h^a\left(\theta^a_{UN}(\alpha)\right) < h^b\left(\theta^b_{UN}(\alpha)\right)$. Therefore, according to result 5 we have, $\forall \alpha > \bar{\alpha}$,
     $\frac{(1-T_{11})}{T_{01}} < h^a\left(\theta^a_{UN}(\alpha)\right) < h^b\left(\theta^b_{UN}(\alpha)\right) < \frac{(1-T_{10})}{T_{00}}$.
   - Because both $\hat{\alpha}^b_{UN}, \hat{\alpha}^a_{UN}$ are the solutions to the balanced equation, then $\frac{1}{\hat{\alpha}^b_{UN}} - 1$ must be bounded by $\frac{(1-T_{11})}{T_{01}}$ and $\frac{(1-T_{10})}{T_{00}}$ too under the Leg-Up effect.
     Hence, if $\beta(\hat{x})\left(\frac{1}{\bar{\alpha}} - 1\right) = \frac{u_+}{u_-} \geq \frac{(1-T_{10})}{T_{00}}\beta(\hat{x}) \Rightarrow \frac{1}{\bar{\alpha}} - 1 > \frac{(1-T_{10})}{T_{00}}$, then $\bar{\alpha} \leq \hat{\alpha}^b_{UN}$.
     Therefore, $\hat{\alpha}^b_{UN} < \hat{\alpha}^a_{UN}$ must hold under Leg up effect, according to result 3.

2. Show that if $\frac{u_+}{u_-} \geq \frac{(1-T_{10})}{T_{00}}\beta\left(\hat{x}\right)$ under the Leg-Up effect for EqOpt fair policy, then $\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN} > \hat{\alpha}^a_{EqOpt} - \hat{\alpha}^b_{EqOpt} \geq 0$:
   - Under EqOpt, when $\alpha^a = \alpha^b = \alpha$, $\mathcal{P}^a_{EqOpt}(\theta^a_{EqOpt}(\alpha,\alpha)) = G^a_1(\theta^a_{EqOpt}(\alpha,\alpha)) = G^b_1(\theta^b_{EqOpt}(\alpha,\alpha)) = \mathcal{P}^b_{EqOpt}(\theta^b_{EqOpt}(\alpha,\alpha))$ must hold. Therefore, $\forall\alpha \geq \bar{\alpha}$, $\theta^a_{EqOpt}\left(\alpha,\alpha\right) = \theta^b_{EqOpt}\left(\alpha,\alpha\right)$.
   - $\theta^a_{UN}\left(\alpha\right) < \theta^a_{EqOpt}\left(\alpha,\alpha\right) = \theta^b_{EqOpt}\left(\alpha,\alpha\right) < \theta^b_{UN}\left(\alpha\right) < \hat{x}$ must hold, otherwise Equation 3 will be violated. Hence, $\forall\alpha > \bar{\alpha}, h^b\left(\theta^b_{EqOpt}\left(\alpha,\alpha\right)\right) < h^b\left(\theta^b_{UN}\left(\alpha\right)\right)$ and $h^a\left(\theta^a_{EqOpt}\left(\alpha,\alpha\right)\right) > h^a\left(\theta^a_{UN}\left(\alpha\right)\right)$ according to result 4.
   - Because $\tilde{\alpha}^s_{EqOpt}$ is the solution to the balanced equation $h^s\left(\theta^s_{EqOpt}\left(\alpha,\alpha\right)\right) = \frac{1}{\alpha^s} - 1$ then result 3, $\tilde{\alpha}^a_{EqOpt} < \hat{\alpha}^a_{UN}, \tilde{\alpha}^b_{EqOpt} > \hat{\alpha}^b_{UN}$.
   - If $\forall\alpha > \bar{\alpha}, \theta^a_{EqOpt}\left(\alpha,\alpha\right) = \theta^b_{EqOpt}\left(\alpha,\alpha\right)$ then, $\mathbb{G}^a_1\left(\theta^a_{EqOpt}\left(\alpha,\alpha\right)\right) = \mathbb{G}^b_1\left(\theta^b_{EqOpt}\left(\alpha,\alpha\right)\right)$ and $\mathbb{G}^a_0\left(\theta^a_{EqOpt}\left(\alpha,\alpha\right)\right) \leq \mathbb{G}^b_0\left(\theta^b_{EqOpt}\left(\alpha,\alpha\right)\right)$ when $\theta^s_{EqOpt} < \hat{x}$. Thus, $\forall\alpha > \bar{\alpha}, h^a\left(\theta^a_{EqOpt}\left(\alpha,\alpha\right)\right) < h^b\left(\theta^b_{EqOpt}\left(\alpha,\alpha\right)\right) \Rightarrow \tilde{\alpha}^a_{EqOpt} \geq \tilde{\alpha}^b_{EqOpt}$
   - We saw in proof of Theorem 4 that $\hat{\alpha}^a_C$ and $\hat{\alpha}^b_C$ are always between $\tilde{\alpha}^b_C$ and $\tilde{\alpha}^a_C$, therefore, $\hat{\alpha}^b_{UN} < \tilde{\alpha}^b_{EqOpt} \leq \hat{\alpha}^b_{EqOpt} \leq \hat{\alpha}^a_{EqOpt} \leq \tilde{\alpha}^a_{EqOpt} < \hat{\alpha}^a_{UN}$, which mean $\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN} > \hat{\alpha}^a_{EqOpt} - \hat{\alpha}^b_{EqOpt} \geq 0$.
3. Show that under the Leg-Up effect for DP fair policy, $\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN} > \hat{\alpha}^a_{DP} - \hat{\alpha}^b_{DP} \geq 0$ or $\hat{\alpha}^a_{DP} \leq \hat{\alpha}^b_{DP}$:
   - Similarly to the way we showed under the EqOpt constraint, we can show that under the Leg-Up effect $\tilde{\alpha}^a_{DP} < \hat{\alpha}^a_{UN}, \tilde{\alpha}^b_{DP} > \hat{\alpha}^b_{UN}$.
   - $\theta^a_{DP}\left(\alpha,\alpha\right) > \theta^b_{DP}\left(\alpha,\alpha\right)$ can result in either $\mathbb{G}^a_0\left(\theta^a_{DP}\left(\alpha,\ \alpha\right)\right) \leq \mathbb{G}^b_0\left(\theta^b_{DP}\left(\alpha,\ \alpha\right)\right)$ or $\mathbb{G}^a_0\left(\theta^a_{DP}\left(\alpha,\ \alpha\right)\right) \geq \mathbb{G}^b_0\left(\theta^b_{DP}\left(\alpha,\ \alpha\right)\right)$, depending on whether $\theta^a_{DP}\left(\alpha,\alpha\right)$ and $\theta^b_{DP}\left(\alpha,\alpha\right)$ are smaller or bigger than $\hat{x}$ respectively. Therefore, both $\tilde{\alpha}^a_{DP} \leq \tilde{\alpha}^b_{DP}$ and $\tilde{\alpha}^a_{DP} \geq \tilde{\alpha}^b_{DP}$ are likely to occur.
   - If $\tilde{\alpha}^a_{DP} \geq \tilde{\alpha}^b_{DP}$ then, similarly to the way we showed under the EqOpt constraint, we can show that under the Leg-Up effect $\hat{\alpha}^a_{UN} - \hat{\alpha}^b_{UN} > \hat{\alpha}^a_{DP} - \hat{\alpha}^b_{DP} \geq 0$.
   - We saw in proof of Theorem 4 that $\hat{\alpha}^a_C$ and $\hat{\alpha}^b_C$ are always between $\tilde{\alpha}^b_C$ and $\tilde{\alpha}^a_C$, therefore if $\tilde{\alpha}^a_{DP} \leq \tilde{\alpha}^b_{DP}$ then, $\hat{\alpha}^a_{DP} \leq \hat{\alpha}^b_{DP}$ because now the intersection is above the $\alpha^b = \alpha^a$ line, where the y coordinate is bigger than the x coordinate.

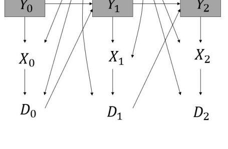
3 Extension

3.1 Motivation

External factors such as community programs, mentoring initiatives, or educational campaigns can significantly influence the qualification rates. In real-world scenarios, qualification rates aren't solely determined by an individual's features or the institute's decisions. By integrating this into the model, we can have a more realistic view of how qualification rates evolve over time and develop fairness policies that take into account these external influences.

3.2 Model Settings

Consider a new variable $E_t = e \in [0,1]$ that captures the intensity or the presence of such interventions at any given time step t ($e = 0$ implies no external intervention, and $e = 1$ implies full intensity of the intervention). I assume that the intervention's effect isn't immediate but instead has a delayed effect, therefore the next qualification state, Y_{t+1} depends on E_t. Moreover, I assume that the intervention at E_t is in reaction to the effects (or lack thereof) of the intervention at E_{t-1} and is not adaptive to the qualification state at the previous step,Y_{t-1}, therefore E_{t+1} depends only on E_t.



The institute needs to be aware of the nature and intensity of interventions for this model to work effectively. Ergo, we assume that this information is observable to the institute.

External interventions will affect how the qualification state transitions, since Y_{t+1} depends on E_t. We can incorporate this by modifying the transition probabilities:

$$T^s_{yde} = P\left(Y_{t+1} = 1 |\, Y_t = y, D_t = d, S = s, E_t = e\right)$$

Since E_{t+1} depends on E_t and s, its dynamics can be captured by:

$$P\left(E_{t+1} = e' \mid E_t = e, S = s\right) = f(e, s)$$

Where f is some function that models the relationship between the current state of the system and the next intervention's intensity.

We assume that the institute either doesn't have direct immediate costs/benefits from the intervention or that these costs/benefits are not significant compared to the expected future utility changes. Therefore, any cost or gain from the intervention is indirectly reflected in the future qualification states, thus, the utility function and the reward function equations would remain as they are.

Since external interventions affect the qualification state, we define the qualification rate and qualification profile as:

$$\alpha^s_{te} = P(Y_t = 1|S = s, E_{t-1} = e)$$
$$\gamma^s_{te}(x) = P(Y_t = 1|S = s, X_t = x, E_{t-1} = e)$$

We define the new dynamics for the qualification rate as follows:

$$\alpha^s_{t+1,e} = P\left(Y_{t+1} = 1|S = s, E_t = e\ \right) = \left(1 - \alpha^s_{t,e}\right) \cdot g^{0se}\left(\alpha^a_t, \alpha^b_t\right) + \alpha^s_{t,e} \cdot g^{1se}\left(\alpha^a_t, \alpha^b_t\right)$$

Where $g^{yse}\left(\alpha^a_t, \alpha^b_t\right) = E_{G^s_y(x)}\left[T^s_{y0e} \cdot (1 - \pi^s_t(x)) + T^s_{y1e} \cdot \pi^s_t(x)\right]$[see 5]

We can see that $\alpha^s_t = \int_e \alpha^s_{t,e} \cdot P(E_t = e|E_{t-1} = e', S = s)$

If we can prove that when $t \to \infty$ then $E_{t+1} = E_t$, i.e., the dynamics of the external intervention reach equilibrium, then we can prove the rest of the main theorems in this paper with some modifications.

3.3 Limitations

While the proposed extension addresses certain limitations of the original model, it remains an incomplete representation of real-world complexities. As the number of variables or effects incorporated into the model grows, so does its complexity, often exponentially. Within the scope of external intervention effects, I haven't accounted for numerous potential scenarios. Examples include the immediate influence of interventions on qualification rates and the utility function, and the impact of each qualification rate on the intensity of the following intervention (suggests the possibility of an intervention policy that adapts to the change in the qualification states).

Another assumption made is the observability of intervention intensities. However, given that these interventions might be initiated by third parties, the institution would need a mechanism to access such data. While interventions spearheaded by governmental bodies may be transparent due to the expected collaboration and transparency with the institute, interventions from private entities or competitors may not be as forthcoming with information about their methodologies in determining intensity.

4 Conclusion & Future Work

In this comprehensive exploration of long-term qualification rates, influenced by algorithmic fairness decisions, I learned several key insights. First, short-term fairness interventions, while well intent, do not always guarantee long-term equity. Depending on underlying feature distributions and transition dynamics, the imposition of fairness constraints could alleviate or intensify inequality, illustrating the importance of strategic planning in policy and algorithm development. Often observed disparities of feature distributions and motivational factors to an institute's decision can lead to 'Natural Inequality'. Such disparities spotlight the necessity to delve deeper into the root causes of inequality.

In addition, the current extension offers a framework for further investigation in various avenues:

- **Adaptive Policy Formulation.** A potential avenue is the development of adaptive policies that factor in how qualification rates influence the intensity of subsequent interventions.
- **Equilibrium Dynamics of Interventions.** A pivotal area of study for this extension lies in contrasting the equilibrium dynamics associated with external interventions. Establishing this equilibrium, along with corresponding modifications, could lead to more robust and generalizable theorems.
- **Navigating Information Asymmetry.** The model's assumption of observable intervention intensities, particularly from third-party sources, opens up a discussion on information asymmetry. Future research could focus on developing mechanisms to handle such scenarios, especially when dealing with non-transparent private entities or competitors.

References

[1] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, pages 3150–3158. PMLR, 2018.

5 Supplementary Material

$\mathcal{G}_s$	demographic group, $s \in \{a, b\}$
S	sensitive attribute that the groups are distinguished by.
p_s	group proportion of s. i.e., $p_s := P(S = s)$
X_t	feature at time t. $X_t = x \in R$
Y_t	hidden qualification state at time t. $Y_t = y \in \{0, 1\}$
D_t	institute's decision at t, (reject or accept). $D_t = d \in \{0, 1\}$
$\pi_t^s(x)$	policy for $\mathcal{G}_s$ at t,$\pi_t^s\left(x\right) := P\left(D_t = 1 \mid X_t = x, S = s\right)$
$G_y^s\left(x\right)$	feature distribution of unqualified (y = 0) or qualified (y = 1) people from G_s , i.e., $P\left(X_t = x \mid Y_t = y,\ S = s\right)$
$\mathbb{G}_y^s\left(x\right)$	CDF of $G_y^s\left(x\right)$, i.e.,$\mathbb{G}_y^s\left(x\right) = \int_{-\infty}^{x} G_y^s\left(z\right) dx$
α_t^s	qualification rate of $\mathcal{G}_s$ at t, $\alpha_t^s := P(Y_t = 1 \mid S = s)$
$\gamma_t^s\left(x\right)$	qualification profile of $\mathcal{G}_s$ at t (probability that an individual with features x from group $\mathcal{G}_s$ is qualified at t), i.e., $P(Y_t = 1 \mid X_t = x, S = s) = \frac{1}{\left(\frac{1}{\alpha_t^s} - 1\right)\left(\frac{G_0^s(x)}{G_1^s(x)}\right) + 1}$, $x \in R$
T_{yd}^s	transition probability of $\mathcal{G}_s$ given that the qualification state is y and the institute's decision is d, i.e., $P\left(Y_{t+1} = 1 \mid Y_t = y, D_t = d, S = s\right)$
$\hat{\alpha}^s$	qualification rate of $\mathcal{G}_s$ at the equilibrium under policy with constraint $C \in \{UN,\ DP,\ EqOpt\}$
g^{ys}	expected qualification rate at the next step of $\mathcal{G}_s$ for individuals with qualification state y at the next time step. i.e., $g^{ys}\left(\alpha_t^a, \alpha_t^b\right) := E_{(X_t \mid Y_t = y, S = s)}\left[\left(1 - \pi_t^s\left(X_t\right)\right) T_{y0}^s + \pi_t^s\left(X_t\right) T_{y1}^s\right]$
θ_C^s	threshold in a threshold policy for $\mathcal{G}_s$ under constraint C, i.e., $\pi_t^s\left(x\right) = I\left[x \geq \theta_C^s\right]$
u_+	benefit that the institute gains by accepting a qualified individual
u_-	cost incurred to the institute by accepting an unqualified individual
h^s	measure of the relative change in the qualification rates of $\mathcal{G}_s$ (fully agreed to qualified), under the threshold policy $\theta^s\left(\alpha^a, \alpha^b\right)$. (How difficult or easy for unqualified people to become qualified, compared to qualified ones to maintain their status, within the given policy). $h^s := \frac{1 - g^{1s}(\theta^s(\alpha^a, \alpha^b))}{g^{0s}(\theta^s(\alpha^a, \alpha^b))}$
$\mathcal{P}_C^s$	a probability distribution over X_t that specifies the fairness metric C. $\mathcal{P}_{EqOpts}^s(x)\ = G_1^s(x)$ $\mathcal{P}_{DP}^s(x)\ =\ \left(1 - \alpha_t^s\right) G_0^s(x)\ + \alpha_t^s G_1^s(x)$
$-s$	The opposite group, $-s := \{a, b\} \setminus s$.
$\Psi^s\left(\alpha^{-s}\right)$	Balanced Set - the set of all qualification rates that satisfy the balanced equations, i.e., $\forall \alpha^{-s} \in [0, 1]$ the balanced set w.r.t. dynamics is: $\Psi^s\left(\alpha^{-s}\right) := \left\{\bar{\alpha}^s : \frac{1}{\bar{\alpha}^s} - 1 = \frac{1 - g^{1s}\left(\theta^s\left(\bar{\alpha}^s, \alpha^{-s}\right)\right)}{g^{0s}\left(\theta^s\left(\bar{\alpha}^s, \alpha^{-s}\right)\right)}\right\}$
$\psi^s\left(\alpha^{-s}\right)$	balanced function - maps α^{-s} to its' unique $\bar{\alpha}^s$ value, when the size of the balanced set is one $\forall \alpha^{-s} \in [0, 1]$. i.e., $\exists \alpha^s \in [0, 1]$ s.t. $\alpha^s = \psi^s\left(\alpha^{-s}\right)$.
$\mathcal{C}_1$	Represents the line where for each α^b, the corresponding α^a is given by $\psi_C^a\left(\alpha^b\right)$ on the 2D plane, where the x-axis is α^a and the y-axis is α^b. $\mathcal{C}_1 = \left\{\left(\alpha^a, \alpha^b\right) :\ \alpha^a = \psi_C^a\left(\alpha^b\right), \alpha^b \in [0, 1]\right\}$
$\mathcal{C}_2$	Represents the line where for each α^a, the corresponding α^b is given by $\psi_C^b\left(\alpha^a\right)$ on the 2D plane, where the x-axis is α^a and the y-axis is α^b. $\mathcal{C}_2 = \left\{\left(\alpha^a, \alpha^b\right) :\ \alpha^b = \psi_C^b\left(\alpha^a\right), \alpha^a \in [0, 1]\right\}$
$\tilde{\alpha}_C^s$	the solutions to $\begin{cases} \alpha^s = \psi_C^s\left(\alpha^{-s}\right) \\ \alpha^{-s} = \alpha^s \end{cases}$, i.e., $\tilde{\alpha}_C^s = \psi_C^s\left(\tilde{\alpha}^s\right)$.

$$
\begin{gathered}
\alpha^s_{t+1,e} = P\left(Y_{t+1}=1|S=s,E_t=e\ \right) = \int_x \sum_{y,d} P\left(Y_{t+1}=1,\ Y_t=y,\ D_t=d, X_t=x|S=s,E_t=e\right)dx = \\
\int_x \sum_{y,d} P\left(Y_{t+1}=1|\ Y_t=y,\ D_t=d, X_t=x,\ S=s, E_t=e\right) \cdot P\left(D_t=d|\ X_t=x,\ S=s\right) \\
\cdot P\left(X_t=x|\ Y_t=y,\ S=s\right) \cdot P\left(Y_t=y|\ S=s,E_t=e\right)dx = \\
\int_x \sum_{d} P\left(Y_{t+1}=1|\ Y_t=0,\ D_t=d, X_t=x,\ S=s, E_t=e\right) \cdot P\left(D_t=d|\ X_t=x,\ S=s\right) \cdot P\left(X_t=x|\ Y_t=0,\ S=s\right) \cdot \\
P\left(Y_t=0|\ S=s,E_t=e\right)dx + \int_x \sum_{d} P\left(Y_{t+1}=1|\ Y_t=1,\ D_t=d, X_t=x,\ S=s, E_t=e\right) \cdot \\
P\left(D_t=d|\ X_t=x,\ S=s\right) \cdot P\left(X_t=x|\ Y_t=1,\ S=s\right) \cdot P\left(Y_t=1|\ S=s,E_t=e\right)dx \\
= \int_x \left(\sum_d T^s_{0de} \cdot P\left(D_t=d|\ X_t=x,\ S=s\right) \cdot G^s_0\left(x\right) \cdot \left(1-\alpha^s_{t,e}\right) \right) dx + \\
\int_x \left(\sum_d T^s_{1de} \cdot P\left(D_t=d|\ X_t=x,\ S=s\right) \cdot G^s_1\left(x\right) \cdot \alpha^s_{t,e} \right) dx = \\
\left(1-\alpha^s_{t,e}\right) \cdot \int_x G^s_0\left(x\right) \cdot \left(T^s_{00e} \cdot \left(1-\pi^s_t\left(x\right)\right) + T^s_{01e} \cdot \pi^s_t\left(x\right)\right)dx + \alpha^s_{t,e} \cdot \int_x G^s_1\left(x\right) \cdot \left(T^s_{10e} \cdot \left(1-\pi^s_t\left(x\right)\right) + T^s_{11e} \cdot \pi^s_t\left(x\right)\right)dx = \\
\left(1-\alpha^s_{t,e}\right) \cdot E_{G^s_0(x)}\left[T^s_{00e} \cdot \left(1-\pi^s_t\left(x\right)\right) + T^s_{01e} \cdot \pi^s_t\left(x\right)\right] + \alpha^s_{t,e} \cdot E_{G^s_1(x)}\left[T^s_{10e} \cdot \left(1-\pi^s_t\left(x\right)\right) + T^s_{11e} \cdot \pi^s_t\left(x\right)\right] = \\
= \left(1-\alpha^s_{t,e}\right) \cdot g^{0se}\left(\alpha^a_t, \alpha^b_t\right) + \alpha^s_{t,e} \cdot g^{1se}\left(\alpha^a_t, \alpha^b_t\right)
\end{gathered}
$$