Online Appendix to

Discrimination as Retaliation

by Till Wicker

B1 Extensions to Conceptual Framework in Section 2

B1.1 Paternalistic Discrimination (Buchmann et al., 2024)

To allow for paternalistic discrimination, the utility function of the employer also contains a fraction α_{eg} of the expected on-the-job welfare of the worker, where employers differ in whether they internalize their perception of the worker's perception of welfare, or internalize their own perception of workers' welfare. In particular, the utility function of the manager becomes:

$$\max_{L_{A,t},L_{B,t}} \underbrace{Y(L_{A,t},\theta_A,L_{B,t},\theta_B) - \sum_{g\in\{A,B\}} L_{g,t}w_g}_{\text{Firm Profit}} - \underbrace{\sum_{g\in\{A,B\}} L_{g,t}f\Big(d_g,F(\chi_{g,t})\Big)}_{\text{Non-Pecuniary Costs}} - \underbrace{\sum_{g\in\{A,B\}} L_{g,t}\alpha_{g,t}\mathcal{W}_{g,t}}_{\text{Other-regarding utility}}$$

where $\mathcal{W}_{g,t}$ is the manager's perception of the workers' perception of welfare, which is defined with respect to the outside option. The worker's welfare consists of their wage, and disutility of working ($\mathcal{W} = \mathbb{E}_i[w_g - u_g(c)]$). Paternalistic employers internalize their own perceptions of the worker's welfare.

B1.2 Experience-based Discrimination (Lepage, 2024)

The employer's utility function remains as specified:

$$\max_{L_{A,t},L_{B,t}} Y(L_{A,t},\theta_A,L_{B,t},\theta_B) - \sum_{g\in\{A,B\}} L_{g,t}w_g - \sum_{g\in\{A,B\}} L_{g,t}f(d_g,F(\chi_{g,t}))$$

However, we now incorporate dynamic belief updating based on accumulated experiences with each group. In particular, at $t = 0$, employers have prior beliefs about group g's

productivity: $\hat{\theta}_{g,0} \sim N(\hat{\mu}_{g,0}, 1/\hat{\tau}_{g,0})$.

After each hiring decision, employers observe realized productivity $\theta_{g,i}$ for each hired worker i from group g. Following Bayesian updating combined with experience-based learning:

$$\hat{\mu}_{g,t+1} = \alpha_\mu \hat{\mu}_{g,t} + (1-\alpha_\mu)\left[\beta_g(\chi_{g,t}) \cdot \overline{\theta}_{g,obs,t} + (1-\beta_g(\chi_{g,t})) \cdot \hat{\mu}_{g,t}\right]$$
$$\hat{\tau}_{g,t+1} = \alpha_\tau \hat{\tau}_{g,t} + (1-\alpha_\tau)\left[\frac{H_{g,t}}{\text{var}(\theta_{g,obs,t})}\right]$$

where $\alpha_\mu, \alpha_\tau \in [0,1]$ are experience weights (higher values place more weight on past beliefs), $\beta_g(\chi_{g,t}) \in [0,1]$ is the experience-dependent learning rate from new observations, $H_{g,t}$ is the cumulative number of workers hired from group g up to time t, and $\overline{\theta}_{g,obs,t} = \frac{1}{H_{g,t}} \sum_{i=1}^{H_{g,t}} \theta_{g,i}$ is the sample mean of observed productivity.

Lepage (2024) illustrates that the learning rate itself depends on past experiences:

$$\beta_g(\chi_{g,t}) = \beta_0 \cdot \exp(-\gamma \cdot F(\chi_{g,t})) \quad \text{where } \gamma > 0, \beta_0 \in (0,1]$$

This specification captures the psychological mechanism whereby negative experiences with respect to the productivity of hired workers make employers less receptive to contradictory information. As such, past experiences can have an effect on current discrimination through two channels:

1. Learning about group-level productivity, as a result of past experiences (Experience-based discrimination, Lepage (2024))

2. Endogenously updating non-pecuniary costs of hiring workers from a specific group (Retaliatory discrimination)

B1.3 Inaccurate Statistical Discrimination

The employer's utility function remains as specified in Equation (1):

$$\max_{L_{A,t}, L_{B,t}} Y(L_{A,t}, \theta_A, L_{B,t}, \theta_B) - \sum_{g \in \{A,B\}} L_{g,t} w_g - \sum_{g \in \{A,B\}} L_{g,t} f(d_g, F(\chi_{g,t}))$$

Following Bohren et al. (2025a), we now explicitly distinguish between true and subjective productivity distributions:

True Productivity Distribution: Worker productivity for group g is drawn from $\theta_g \sim N(\mu_g, 1/\tau_g)$ with true signal precision η_g.

Subjective Beliefs: Employers hold potentially inaccurate subjective beliefs $\hat{\psi} \equiv (\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g)$ about group g's productivity distribution and signal precision, where:

$$\hat{\theta}_g \sim N(\hat{\mu}_g, 1/\hat{\tau}_g) \tag{1}$$

$$\text{Subjective signal precision:} \quad \hat{\eta}_g \geq 0, \quad \hat{\eta}_g \neq \eta_g \tag{2}$$

Inaccurate Statistical Discrimination occurs when $(\hat{\mu}_g, \hat{\tau}_g, \hat{\eta}_g) \neq (\mu_g, \tau_g, \eta_g)$ for some group g.

Following Bohren et al. (2025a), employers make hiring decisions based on their subjective posterior beliefs. After observing worker's group identity g and signal s, the employer forms a posterior belief using Bayes' rule with subjective distributions:

$$\hat{\mu}_{g,t}(s) = \frac{\hat{\tau}_g \hat{\mu}_g + \hat{\eta}_g s}{\hat{\tau}_g + \hat{\eta}_g} \tag{3}$$

B1.4 Systemic Discrimination

Systemic discrimination captures how discrimination in other decisions indirectly contributes to disparities by affecting relevant attributes for a given decision, which in turn generates disparities in outcomes. This extension demonstrates how retaliatory discrimination at the focal node can coexist with systemic discrimination arising from other nodes in the decision system.

Following Bohren et al. (2025b), we embed the retaliatory discrimination model within a broader system of interconnected decision nodes. The system consists of a set of nodes $N \equiv \{1, ..., N\} \cup \{n^*\}$, where n^* represents the focal hiring node from our baseline model. Each non-focal node $n = 1, ..., N$ represents a decision task where:

- Worker i has productivity $\theta\_i^n \in \Theta^n$ for task n
- An employer observes the worker's group G_i and signal $S_i^n \in \mathcal{S}^n$

- The employer selects action $A_i^n \in \mathcal{A}^n$ according to action rule $A^n(G_i, S_i^n)$. The action is to either hire the worker, or not.

At the focal node n^* (our baseline hiring decision):

- Worker productivity is $\theta_i^* \in \Theta^*$
- Signal is $S_i^* \in \mathcal{S}^*$
- Action is $A_i^* \in \mathcal{A}^*$

As discussed in Bohren et al. (2025b), actions at other nodes can affect productivity and signals at the focal node. For example, S_i^* may include performance evaluations from other nodes, or focal-node productivity θ_i^* may depend on training received at node n.

Incorporating this within the retaliatory discrimination model, the employer's utility function at the focal node becomes:

$$\max_{L_{A,t}, L_{B,t}} Y(L_{A,t}, \theta_A, L_{B,t}, \theta_B) - \sum_{g \in \{A,B\}} L_{g,t} w_g - \sum_{g \in \{A,B\}} L_{g,t} f(d_g, F(\chi_{g,t}))$$

However, the signal S_i^* and/or productivity Y_i^* now depend on actions at other nodes:

$$S_i^* = S^*(G_i, A_i^1, A_i^2, ..., A_i^N, \xi_i)\theta_i^* \qquad = \theta^*(G_i, A_i^1, A_i^2, ..., A_i^N, \zeta_i) \tag{4}$$

where ξ_i and ζ_i represent other factors affecting signals and productivity.

As such, retaliatory discrimination at one node can have consequences for future nodes, and hence lead to more, systematized discrimination.

B2 Pilot Data Insights

B2.1 Recognition of Nationality by Name

Table B1: Correctly Identified Nationality by Name During Pilot.

	Name and Nationality: Pilot	
	Ugandan Name	Eritrean Name
	(1)	(2)
Correctly Identified by Eritrean	97.33%	96.00%
Incorrectly Identified by Eritrean	2.67%	4.00%
Correctly Identified by Ugandan	100.00%	94.67%
Incorrectly Identified by Ugandan	0.00%	5.33%
N	50	50

Notes: Refers to data collected in a pilot study in December 2024 with 25 Eritrean refugees, and 25 Ugandans. None of these individuals participated in the final study. Pilot participants were shown the name of their fellow co-worker and manager (in stage 1 of the experiment), and of two workers (in stage 2 of the experiment).

B2.2 Average Quality of Envelopes Made

Table B2: Quality of Envelopes and Time Taken During Pilot.

	Quality of Envelope	Time Taken to Make Envelopes
	(1)	(2)
Eritrean Workers	0.53	383.91
Ugandan Workers	0.51	380.61
N	50	50

Notes: Refers to data collected in a pilot study in December 2024 with 25 Eritrean refugees, and 25 Ugandans. None of these individuals participated in the final study. Quality of Envelope is based on five defined categories, each evaluated on a $\{0, 1\}$ scale, and averaged. Time taken to make envelopes is measured in seconds. On average, individuals made 4 envelopes. This also refers to the data shared with participants in the study.

B2.3 Word Clouds of Reasons Why Given Set Number of Tasks

During the pilot study, detailed beliefs were elicited about the expectations of the number of tasks the participant would receive, and why they believed they ultimately received the number of tasks that they did. These were converted into word clouds (using a free word cloud generator). Below, I illustrate the word clouds for when the participants received two, four, and six out of eight possible tasks:

Figure B1. Word Clouds

(a) Assigned **Two** Tasks

(b) Assigned **Four** Tasks

(c) Assigned **Six** Tasks

B3 Balance Table

B3.1 Uganda: Lab-in-the-Field

Table B3: Balance Table: Uganda.

	(T1)		(T2)		(T3)		(T4)		F-test	
	Computer Manager				***Ugandan Manager***					
	(2,6)		***(4,4)***		***(2,6)***		***(4,4)***			
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	F-stat/P-value
Age	56	30.714	55	30.418	56	31.071	57	30.772	224	0.094
		(7.586)		(5.570)		(6.760)		(5.846)		0.963
Ugandan friends	56	2.589	55	2.509	56	2.661	57	2.333	224	0.486
		(1.797)		(1.373)		(1.552)		(1.314)		0.692
Arrival Year	56	2017.232	55	2016.418	56	2017.214	57	2016.544	224	0.513
		(4.884)		(4.003)		(5.263)		(3.689)		0.674
Attitudes Towards Ugandans	56	-0.120	55	0.122	56	-0.022	57	0.017	224	1.885
		(0.642)		(0.557)		(0.457)		(0.499)		0.133
Empathy Index	56	0.135	55	-0.137	56	-0.010	57	-0.105	224	3.707**
		(0.552)		(0.438)		(0.442)		(0.454)		0.012
Retaliation Index	56	-0.103	55	0.105	56	0.227	57	0.200	224	2.215*
		(0.826)		(0.606)		(0.728)		(0.825)		0.087

Notes: Columns (T1), (T2), (T3), and (T4) show the average value (and standard deviation) for respondents in each of the four treatment arms: Ugandan manager (2,6), Ugandan manager (4,4), Computer manager (2,6), and Computer manager (4,4), where values in parentheses indicate the allocation of the manager in the first stage of the game. The F-test reports the joint test for orthogonality, including both the F-statistic and associated p-value. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

B3.2 America: Online Experiment

Table B4: Balance Table: USA.

	(T1)		(T2)		(T3)		(T4)		(T5)		(T6)		F-test	
		(2,6)		***Coethnic Manager*** ***(4,4)***		***(6,2)***		***(2,6)***		***Non-Coethnic Manager*** ***(4,4)***		***(6,2)***		
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	F-stat/P-value
Age	108	40.824	105	39.648	109	40.872	99	38.960	109	38.697	109	38.550	639	1.145
		(10.340)		(9.779)		(9.613)		(9.788)		(10.863)		(10.321)		0.335
Total Approvals	108	2650.694	105	2201.610	109	2412.633	99	2638.556	109	2035.055	109	2517.532	639	1.158
		(2665.565)		(2215.184)		(2502.470)		(2419.970)		(1798.169)		(2529.070)		0.329
Ethnicity: African American	108	0.481	105	0.495	109	0.477	99	0.485	109	0.523	109	0.450	639	0.249
		(0.502)		(0.502)		(0.502)		(0.502)		(0.502)		(0.500)		0.940
USA National	108	1.000	105	0.981	109	0.982	99	0.990	109	0.963	109	1.000	639	1.515
		(0.000)		(0.137)		(0.135)		(0.101)		(0.189)		(0.000)		0.183
Student	108	0.093	105	0.133	109	0.055	99	0.071	109	0.119	109	0.128	639	1.235
		(0.291)		(0.342)		(0.229)		(0.258)		(0.326)		(0.336)		0.291
Employed	108	0.093	105	0.086	109	0.073	99	0.061	109	0.101	109	0.064	639	0.372
		(0.291)		(0.281)		(0.262)		(0.240)		(0.303)		(0.246)		0.868
Detailed Elicitation	108	0.509	105	0.457	109	0.495	99	0.455	109	0.523	109	0.523	639	0.401
		(0.502)		(0.501)		(0.502)		(0.500)		(0.502)		(0.502)		0.848

Notes: Columns (T1), (T2), (T3), (T4), (T5), and (T6) show the average value (and standard deviation) for respondents in each of the six treatment arms: Same Ethnicity Manager (2,6), Same Ethnicity Manager (4,4), Same Ethnicity Manager (5,2) Other Ethnicity Manager (2,6), Other Ethnicity Manager (4,4), and Other Ethnicity Manager (6,2), where values in parentheses indicate the allocation of the manager in the first stage of the game. The F-test reports the joint test for orthogonality, including both the F-statistic and associated p-value. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table B5: Balance Table: USA, Detailed Belief Elicitation.

	No		Detailed Beliefs		t-test	
Variable	N	Mean/(SD)	N	Mean/(SD)	N	p-value
Age	323	40.050	316	39.139	639	0.257
		(10.357)		(9.907)		
Total Approvals	323	2456.734	316	2355.522	639	0.590
		(2416.707)		(2330.703)		
Ethnicity: African American	323	0.489	316	0.481	639	0.837
		(0.501)		(0.500)		
USA National	323	0.978	316	0.994	639	0.100
		(0.146)		(0.079)		
Student	323	0.093	316	0.108	639	0.536
		(0.291)		(0.310)		
Employed	323	0.068	316	0.092	639	0.271
		(0.252)		(0.289)		

Notes: Columns show the average value (and standard deviation) for respondents who either provided detailed beliefs, or did not. The t-test reports the associated p-value. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

B3.3 Minimal Group Paradigm

The Minimal Group Paradigm study was conducted on Besample, an online survey platform similar to Prolific that surveys participants across many countries. For this study, 320 men were recruited from Kenya, Ethiopia, Ghana and Nigeria. Appendix Table B6 presents the balance table for this sample:

Table B6: Balance Table: Minimal Group Paradigm.

	(T1)		(T2)		(T3)		(T4)		(T5)		(T6)		F-test	
						Treatment Groups								
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	F-stat/P-value
Urban	53	0.849	51	0.725	58	0.845	55	0.891	53	0.887	56	0.786	326	1.539
		(0.361)		(0.451)		(0.365)		(0.315)		(0.320)		(0.414)		0.177
Employed	53	0.585	51	0.431	58	0.500	55	0.382	53	0.547	56	0.482	326	1.182
		(0.497)		(0.500)		(0.504)		(0.490)		(0.503)		(0.504)		0.318
Age	53	27.887	51	26.882	58	28.914	55	27.673	53	29.698	56	28.929	326	1.190
		(6.883)		(5.945)		(7.373)		(6.449)		(7.360)		(6.954)		0.314
Nationality: Ghana	53	0.340	51	0.451	58	0.345	55	0.400	53	0.321	56	0.411	326	0.568
		(0.478)		(0.503)		(0.479)		(0.494)		(0.471)		(0.496)		0.725
Nationality: Kenya	53	0.208	51	0.118	58	0.190	55	0.182	53	0.151	56	0.125	326	0.522
		(0.409)		(0.325)		(0.395)		(0.389)		(0.361)		(0.334)		0.759
Nationality: Nigeria	53	0.245	51	0.255	58	0.293	55	0.291	53	0.264	56	0.304	326	0.150
		(0.434)		(0.440)		(0.459)		(0.458)		(0.445)		(0.464)		0.980
Highest Schooling: Primary	53	0.000	51	0.000	58	0.000	55	0.018	53	0.000	56	0.000	326	0.985
		(0.000)		(0.000)		(0.000)		(0.135)		(0.000)		(0.000)		0.427
Highest Schooling: Secondary	53	0.264	51	0.275	58	0.259	55	0.255	53	0.189	56	0.268	326	0.276
		(0.445)		(0.451)		(0.442)		(0.440)		(0.395)		(0.447)		0.926
Highest Schooling: Bachelors	53	0.547	51	0.627	58	0.603	55	0.600	53	0.679	56	0.643	326	0.446
		(0.503)		(0.488)		(0.493)		(0.494)		(0.471)		(0.483)		0.816
Highest Schooling: Masters	53	0.132	51	0.059	58	0.086	55	0.073	53	0.094	56	0.071	326	0.442
		(0.342)		(0.238)		(0.283)		(0.262)		(0.295)		(0.260)		0.819
Highest Schooling: Ph.D.	53	0.019	51	0.000	58	0.017	55	0.000	53	0.000	56	0.018	326	0.573
		(0.137)		(0.000)		(0.131)		(0.000)		(0.000)		(0.134)		0.721
Highest Schooling: Vocational	53	0.038	51	0.020	58	0.034	55	0.055	53	0.038	56	0.000	326	0.643
		(0.192)		(0.140)		(0.184)		(0.229)		(0.192)		(0.000)		0.667

Notes: Columns (T1)–(T6) show the average value (and standard deviation) for respondents in each of the six treatment arms. The F-test reports the joint test for orthogonality, including both the F-statistic and associated p-value. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

B3.4 Dictator Game

Table B7: Balance Table: Dictator Game.

	(T1)		(T2)		(T3)		(T4)		(T5)		(T6)		F-test	
			Coethnic Manager				***Non-Coethnic Manager***							
	(2,6)		***(4,4)***		***(6,2)***		***(2,6)***		***(4,4)***		***(6,2)***			
Variable	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	Mean/(SD)	N	F-stat/P-value
Age	62	39.081	61	37.656	62	36.758	62	38.919	61	38.475	61	39.230	369	0.558
		(9.479)		(9.832)		(9.634)		(9.958)		(9.944)		(11.920)		0.732
Total Approvals	62	3597.194	61	2571.344	62	2553.984	62	2426.806	61	2648.803	61	3200.656	369	2.229*
		(2893.761)		(2150.441)		(2255.177)		(2179.148)		(2632.518)		(2373.807)		0.051
Ethnicity: African American	62	0.226	61	0.230	62	0.226	62	0.242	61	0.246	61	0.246	369	0.032
		(0.422)		(0.424)		(0.422)		(0.432)		(0.434)		(0.434)		0.999
USA National	62	1.000	61	1.000	62	1.000	62	1.000	61	1.000	61	1.000	369	
		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)		(0.000)		
Student	62	0.129	61	0.000	62	0.097	62	0.065	61	0.115	61	0.098	369	1.728
		(0.338)		(0.000)		(0.298)		(0.248)		(0.321)		(0.300)		0.127
Employed	62	0.129	61	0.148	62	0.048	62	0.097	61	0.066	61	0.082	369	1.018
		(0.338)		(0.358)		(0.216)		(0.298)		(0.250)		(0.277)		0.406

Notes: Columns (T1), (T2), (T3), (T4), (T5), and (T6) show the average value (and standard deviation) for respondents in each of the six treatment arms: Same Ethnicity Manager (2,6), Same Ethnicity Manager (4,4), Same Ethnicity Manager (5,2) Other Ethnicity Manager (2,6), Other Ethnicity Manager (4,4), and Other Ethnicity Manager (6,2), where values in parentheses indicate the allocation of the manager in the first stage of the game. The F-test reports the joint test for orthogonality, including both the F-statistic and associated p-value. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

B4 Regression Tables - Uganda Experiment

B4.1 Heterogeneity

Table B8: HTE: Ugandan Friends
Allocation of Tasks to Ugandan Worker in Stage 2.

	Allocation of Tasks to U_2 in Stage 2	
	Ugandan Manager in Stage 1	Computer Manager in Stage 1
	(1)	(2)
Stage 1: Negative	-0.50***	-0.44*
	(0.18)	(0.23)
Ugandan Friends	-0.38**	0.15
	(0.18)	(0.19)
Interaction Term	-0.18	0.29
	(0.25)	(0.27)
Control Group Mean	3.63	3.55
Control Group S.D.	0.70	0.66
N	113	111

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Ugandan worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Stage 1: Negative* is a dummy variable equal to 1 if the manager in the first round was Ugandan, and hence refers to treatments T3 and T4. *Ugandan Friends* refers to the number of Ugandan friends the participant reported to have. The *Interaction Term* refers to *Stage 1: Negative* interacted with *Ugandan Friends*. Column (1) reports results for the sub-sample who had a Ugandan manager in stage 1 (T3 and T4), while column (2) reports results for the sub-sample who had a Ugandan manager in stage 1 (T1 and T2). Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group (T2 and T4, respectively). Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table B9: HTE: Empathy
Allocation of Tasks to Ugandan Worker in Stage 2.

	Allocation of Tasks to U_2 in Stage 2	
	Ugandan Manager in Stage 1	Computer Manager in Stage 1
	(1)	(2)
Stage 1: Negative	-0.70***	-0.25
	(0.20)	(0.25)
Empathy	0.11	0.12
	(0.24)	(0.20)
Interaction Term	0.11	-0.14
	(0.28)	(0.28)
Control Group Mean	3.63	3.55
Control Group S.D.	0.70	0.66
N	113	111

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Ugandan worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Stage 1: Negative* is a dummy variable equal to 1 if the manager in the first round was Ugandan, and hence refers to treatments T3 and T4. *Empathy* refers to an inverse-covariance weighted index of five 5-item Likert scale questions: "Other people's misfortunes do not disturb me a great deal"; "It upsets me to see someone being treated disrespectfully"; "I am not really interested in how other people feel"; "When I see someone being treated unfairly, I do not feel very much pity for them"; "When I see someone being taken advantage of, I feel protective towards him/her." The *Interaction Term* refers to *Stage 1: Negative* interacted with *Empathy*. Column (1) reports results for the sub-sample who had a Ugandan manager in stage 1 (T3 and T4), while column (2) reports results for the sub-sample who had a Ugandan manager in stage 1 (T1 and T2). Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group (T2 and T4, respectively). Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table B10: HTE: Retaliation
Allocation of Tasks to Ugandan Worker in Stage 2.

	Allocation of Tasks to U_2 in Stage 2	
	Ugandan Manager in Stage 1	Computer Manager in Stage 1
	(1)	(2)
Stage 1: Negative	-0.67***	-0.15
	(0.18)	(0.18)
Retaliate	-0.16	0.13
	(0.20)	(0.19)
Interaction Term	0.03	-0.42
	(0.28)	(0.27)
Control Group Mean	3.63	3.55
Control Group S.D.	0.70	0.66
N	113	111

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Ugandan worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Stage 1: Negative* is a dummy variable equal to 1 if the manager in the first round was Ugandan, and hence refers to treatments T3 and T4. *Retaliate* refers to an inverse-covariance weighted index of two 7-item Likert scale questions: "If someone does me a favor, I am ready to return it to them"; "If someone treats me unfairly, I'll take the opportunity to get back at them." The *Interaction Term* refers to *Stage 1: Negative* interacted with *Retaliate*. Column (1) reports results for the sub-sample who had a Ugandan manager in stage 1 (T3 and T4), while column (2) reports results for the sub-sample who had a Ugandan manager in stage 1 (T1 and T2). Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group (T2 and T4, respectively). Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table B11: HTE: Attitudes Towards Ugandans
Allocation of Tasks to Ugandan Worker in Stage 2.

	Allocation of Tasks to U_2 in Stage 2	
	Ugandan Manager in Stage 1	Computer Manager in Stage 1
	(1)	(2)
Stage 1: Negative	-0.46**	-0.15
	(0.18)	(0.24)
Attitudes Towards Ugandans	0.28	0.34*
	(0.17)	(0.21)
Interaction Term	-0.41	-0.15
	(0.26)	(0.27)
Control Group Mean	3.63	3.55
Control Group S.D.	0.70	0.66
N	113	111

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Ugandan worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Stage 1: Negative* is a dummy variable equal to 1 if the manager in the first round was Ugandan, and hence refers to treatments T3 and T4. *Attitudes Towards Ugandans* refers to an inverse-covariance weighted index of four 5-item Likert scale questions: "Ugandans are friendly and good people"; "Eritreans are well integrated with Ugandans"; "Ugandan employers discriminate against me because I am an Eritrean"; "I have just as many opportunities to find formal work as my Ugandan neighbors." The *Interaction Term* refers to *Stage 1: Negative* interacted with *Attitudes Towards Ugandans*. Column (1) reports results for the sub-sample who had a Ugandan manager in stage 1 (T3 and T4), while column (2) reports results for the sub-sample who had a Ugandan manager in stage 1 (T1 and T2). Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group (T2 and T4, respectively). Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table B12: HTE: Years in Uganda
Allocation of Tasks to Ugandan Worker in Stage 2.

	Allocation of Tasks to U_2 in Stage 2	
	Ugandan Manager in Stage 1	Computer Manager in Stage 1
	(1)	(2)
Stage 1: Negative	-0.50***	-0.31
	(0.16)	(0.21)
Years in Uganda	0.08	-0.03
	(0.20)	(0.21)
Interaction Term	-0.31	-0.03
	(0.27)	(0.28)
Control Group Mean	3.63	3.55
Control Group S.D.	0.70	0.66
N	113	111

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Ugandan worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Stage 1: Negative* is a dummy variable equal to 1 if the manager in the first round was Ugandan, and hence refers to treatments T3 and T4. *Years in Uganda* refers to the number of years the participant has lived in Uganda. The *Interaction Term* refers to *Stage 1: Negative* interacted with *Years in Uganda*. Column (1) reports results for the sub-sample who had a Ugandan manager in stage 1 (T3 and T4), while column (2) reports results for the sub-sample who had a Ugandan manager in stage 1 (T1 and T2). Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group (T2 and T4, respectively). Robust standard errors are in parentheses.***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

B5 Regression Tables - USA Experiment

B5.1 Heterogeneity

Table B13: HTE: Discriminatory Attitudes
Allocation of Tasks to Non-Coethnic Worker in Stage 2.

	(1)	(2)	(3)
		Number of Tasks Allocated to Non-Coethnic Worker	
	Whole Sample	Below Median Discrim. Index	Above Median Discrim. Index
Stage 1: Non-Coethnic Manager	-0.03	-0.00	-0.12
	(0.04)	(0.04)	(0.11)
Stage 1: Positive	-0.09	-0.04	-0.30*
	(0.08)	(0.10)	(0.15)
Stage 1: Negative	-0.02	0.07	-0.09
	(0.06)	(0.06)	(0.12)
Stage 1: Non-Coethnic & Positive	0.16	0.10	0.34**
	(0.10)	(0.17)	(0.17)
Stage 1: Non-Coethnic & Negative	-0.18*	-0.40**	0.14
	(0.10)	(0.18)	(0.14)
T1 Mean	4.02	4.02	4.02
T1 S.D.	0.38	0.38	0.38
N	639	228	234

Notes: The outcome variable is the number of tasks allocated to the Non-Coethnic worker by the participant in the second stage of the experiment, and ranges from 0 to 8. Regression results are reported separately for (1) the whole sample; (2) participants with a below-median discrimination score; (3) participants with an above-median discrimination score. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Stage 1: Non-Coethnic Manager* is a dummy variable equal to 1 if the manager in the first round was non-coethnic, and hence refers to treatments T4-T6. *Stage 1: Negative* is a dummy variable equal to 1 if the allocation of the manager in the first round was (6 ; 2), and hence refers to treatments T1 and T4. *Stage 1: Positive* is a dummy variable equal to 1 if the allocation of the manager in the first round was (2 ; 6), and hence refers to treatments T3 and T6. The *Interaction Terms* refers to *Stage 1: Non-Coethnic Manager* interacted with *Stage 1: Negative*, and *Stage 1: Positive*, respectively. Control mean and standard deviation refer to the mean value and standard deviation of the outcome in the control group (Coethnic manager with (4 ; 4) allocation in the first stage). Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

Table B14: HTE: Future Rounds
Black and White Men.

	(1)	(2)	(3)
	Number of Tasks Allocated to Non-Coethnic Worker		
	Whole Sample	Black Men	White Men
Treatment: *Future Rounds*	0.17*	0.10	0.19
	(0.11)	(0.09)	(0.20)
Status Quo Mean	3.90	3.91	3.88
Status Quo S.D.	0.72	0.51	0.94
N	148	85	63

Notes: Intention to Treat estimates. The outcome variable is the number of tasks allocated to the Non-Coethnic worker by the participant in the second stage of the experiment, and ranges from 0 to 8. These are reported separately for (1) the whole sample; (2) Black men; (3) White men. Control variables are selected using the post double LASSO machine learning algorithm outlined in Belloni et al. (2014). *Future Rounds* refers to the treatment arm where the experimental instructions heightened the salience of future rounds. *Status Quo* mean and standard deviation refer to the mean value and standard deviation of the outcome in the treatment arm where the salience of future rounds was not made salient (and hence equivalent to T4 of Figure 3). Robust standard errors are in parentheses. ***, ** and * represent significant differences at the 1, 5 and 10% level, respectively.

B6 Anecdotal Evidence of Retaliatory Discrimination

After the online experiment, participants were asked whether the notion of retaliatory discrimination resonated with their own past experiences. Below are some of the responses of participants:

> "I have experienced situations where I felt unfairly treated or discriminated against by someone from a different ethnic group. Sometimes, this led to feelings of frustration or resentment. In some cases, I noticed that I unconsciously responded by being stricter or less cooperative toward others from the same group as the person who treated me unfairly, even though they had nothing to do with the original incident."

> "I have felt treated unfairly at work or school because of my ethnicity. Even if I didn't confront the person, it sometimes affected how I acted toward others in similar situations."

> "I once experienced being treated unfairly by a supervisor, which later made me feel less inclined to cooperate with another colleague from the same background, even though they weren't personally responsible."

> "I once faced bias from a supervisor, and later felt tempted to be less cooperative with another coworker from their group."

> "I once felt unfairly treated by a manager during a group project, he consistently gave me the least desirable tasks. Later when I had to assign roles in a different setting to someone from his same background, I found myself feeling tempted to be less fair."

> "If a Black boss acts aggressively toward me, I may replicate the same behavior toward my Black subordinates."

> "In a previous job, after being treated unfairly by a manager from a certain background, I caught myself being less cooperative with another colleague from that same background."

> "In a previous job,a manager from from one ethnic group consistently gave me fewer opportunities. Later, when I had authority over someone from the same background,

I had to consciously check my own bias to avoid unfair treatment. It was a wake up call about how resentment can linger if we're not self aware."

"In my final year at university, I once felt a lecturer graded my work unfairly compared to my classmates from his own ethnic group. A week later, I was in a group assignment where one teammate was from the same group as that lecturer. I noticed I was initially less willing to cooperate, but I made a conscious effort not to let the previous incident influence me."

B7 Retaliatory Discrimination Micro-Foundations

Prior to the launch of the online experiment on Prolific, I pre-registered four theoretical micro-foundations of retaliatory discrimination on the AEA RCT Registry (AEARCTR-0016047). The function $f\Big(d_g, F(\chi_{g,t})\Big)$ specified in Section 2 is increasing in both arguments, however, the functional form of f is important for understanding the dynamic evolution of discriminatory tastes. Below I present four different specifications of $f(d_g, F(\chi_{g,t}))$ — Retaliatory Tit-for-Tat (social preferences), Bayesian Updating, Motivated Beliefs, and Memory Recall.

B7.1 Reciprocity-Based Tit-for-Tat (Social Preferences)

$f(d_g, F(\chi_{g,t}))$ is simplified to only consider the experience in the last period, with an individual of group g: $f(d_g, \chi_{g,t-1})$. In particular, it takes an additively separable form consisting of the individual's discriminatory prior, d_g, and a tit-for-tat reciprocity-based update, following Rabin (1993):

$$f(d_g, \chi_{g,t-1}) = d_g + \underbrace{\phi[d_{g,t}(\chi_{g,t-1}) - d_g]}_{\text{tit-for-tat}}$$

where $d_{g,t}(\chi_{g,t-1})$ captures the (perceived) discrimination as a result of interactions with individuals of group g in the previous period, $t-1$. ϕ captures the re-activeness of the tit-for-tat response, where a value of $\phi = 1$ corresponds to fully retaliatory behavior, and $\phi = 0$ no retaliatory behavior (and hence a return to the static, taste-based model of Becker (1957)). Values in between ($0 < \phi < 1$) capture partial tit-for-tat, and $\phi > 1$ captures over-retaliation.

B7.2 Bayesian Updating

Employers have a prior distribution of their discriminatory distaste for workers from group g distributed according to the density function $h(d_g)$ of population-level discriminatory preferences with a prior mean $\bar{d}_g$. Employers interact with workers of group g, and these interactions inform the information set of the employer, $I_{g,t}$. Through each interaction, an i.i.d. signal is drawn from the group-specific distaste distribution $D_g|d_g \sim G(x_g)$, characterized by the mean d_g, finite variance, and density function $g(x_g)$. Conditional on d_g, experiences with

workers of group g are independent $(\chi_{g,t}|d_g \overset{\text{i.i.d.}}{\sim} p(x_g|d_g))$, and hence the joint likelihood of the signals is $p(I_{g,t}|d_g) = \prod_{k \in I_{g,t}} g(\chi_{g,t})$. As a result, the posterior distribution of d_g given the observed signals $I_{g,t}$ is:

$$d_g|I_{g,t} = \frac{\prod_{k \in I_{g,t}} g_{d_g}(\chi_{g,t})h(d_g)}{\int \prod_{k \in I_{g,t}} g_{d_g}(\chi_{g,t})h(d_g)\, dd_g}$$

and hence

$$f(d_g, F(\chi_{g,t})) = E\Big[\frac{\prod_{k \in I_{g,t}} g_{d_g}(\chi_{g,t})h(d_g)}{\int \prod_{k \in I_{g,t}} g_{d_g}(\chi_{g,t})h(d_g)\, dd_g}\Big]$$

To derive a closed-form solution for the posterior mean of the employer's discriminatory distastes, $f(d_g, F(\chi_{g,t}))$, I will assume that both the prior distribution $h(d_g)$ and likelihood function $p(x_g|d_g)$ follow a Gaussian normal distribution. In particular, $h(d_g) \sim \mathcal{N}(d_g, \frac{1}{\tau_{d_g}})$, and $p(x_g|d_g) \sim \mathcal{N}(\bar{x}_{g,t}, \frac{1}{\tau_{x_{g,t}}})$.[1] Subsequently,

$$f(d_g, F(\chi_{g,t})) = \frac{d_g \cdot \tau_{d_g} + \bar{x}_{g,t} \cdot \tau_{x_{g,t}} \cdot n_t}{\tau_{d_g} + \tau_{x_{g,t}} \cdot n_t}$$

where $n_t = |I_{g,t}|$, the number of signals the employer received up until time t of individuals in group g - each with the same signal precision τ_{d_g}.[2]

Given the standard setting with Gaussian prior and signal distributions, the posterior distribution can also be written as a linear combination of the two:

$$f(d_g, F(\chi_{g,t})) = \omega_g d_g + (1 - \omega) p(x_g|d_g)$$

where ω_g is the weight on the prior mean for group g, and defined as:

$$\omega_g = \frac{\tau_{d_g}}{n_t \cdot \tau_{x_{g,t}} + \tau_{d_g}}$$

and hence increasing in the precision of the prior distribution versus the signal.

[1] By setting $\tau_{d_g} \to \infty$, the prior distribution of tastes becomes a single value, akin to Becker (1957).

[2] The assumption of equal signal precision can easily be dropped, as is the case with motivated beliefs, see below.

B7.3 Motivated Beliefs

Following the large theoretical and empirical literature documenting that individuals do not always update as a Bayesian does, I incorporate motivated beliefs within the Bayesian updating model.[3] In particular, a motive function is introduced, in line with Thaler (2024). The motive function is applied to individual experience $\chi_{g,t}$ the employer had with individuals from group g at time t. In particular, the motive function affects the interpretation of the experience:

$$\tilde{\tau}_{g,t} = \tau_{\chi_{g,t}} \cdot M(d_g, \chi_{g,t})$$

where $M(d_g, \chi_{g,t})$ is defined as:

$$M(d_g, \chi_{g,t}) = \begin{cases} 1+\alpha & \text{if } \text{sign}(\chi_{g,t}) \cdot \text{sign}(d_g) \geq 0 \quad \text{[confirming signal]} \\ 1-\beta & \text{if } \text{sign}(\chi_{g,t}) \cdot \text{sign}(d_g) < 0 \quad \text{[contradicting signal]} \end{cases}$$

$\alpha > 0$ and $\beta \in (0, 1)$ capture the additional weight or discount applied to experiences confirming and contradicting the employer's prior d_g, compared with the Bayesian standard. $\alpha > \beta$ captures motivated reasoning, as an employer places more weight on past experiences that align with their discriminatory priors d_g than past experiences that go against their priors. As such, the likelihood function is now based on the motive-adjusted precision: $p(\chi_{g,t}|d_g, \tilde{\tau}_{g,t})$. As a consequence, the posterior distribution of d_g given the observed past experiences $I_{g,t}$ is:

$$d_g|I_{g,t} = \frac{\prod_{k \in I_{g,t}} p(\chi_{g,t}|d_g, \tilde{\tau}_{g,t}) h(d_g)}{\int \prod_{k \in I_{g,t}} p(\chi_{g,t}|d_g, \tilde{\tau}_{g,t}) h(d_g)\, dd_g}$$

and hence

$$f(d_g, F(\chi_{g,t})) = E\Big[\frac{\prod_{k \in I_{g,t}} p(\chi_{g,t}|d_g, \tilde{\tau}_{g,t}) h(d_g)}{\int \prod_{k \in I_{g,t}} p(\chi_{g,t}|d_g, \tilde{\tau}_{g,t}) h(d_g)\, dd_g}\Big]$$

Using the same Gaussian assumption as under the Bayesian Updating model to obtain

[3] Motivated beliefs can also be referred to as the confirmation bias (Rabin and Schrag, 1999) and reference-dependent preferences (Kőszegi and Rabin, 2006).

a closed form solution, the final posterior mean becomes:

$$f(d_g, F(\chi_{g,t})) = \frac{d_g \cdot \tau_{d_g} + \sum_{k \in I_{g,t}} \tau_{x_{g,t}} \cdot M(\chi_{g,t}, d_g) \cdot \bar{x}_{g,t}}{\tau_{d_g} + \sum_{k \in I_{g,t}} \tau_{x_{g,t}} \cdot M(\chi_{g,t}, d_g)} = \frac{d_g \cdot \tau_{d_g} + \sum_{k \in I_{g,t}} \tilde{\tau}_{g,t} \cdot \bar{x}_{g,t}}{\tau_{d_g} + \sum_{k \in I_{g,t}} \tilde{\tau}_{g,t}}$$

B7.4 Recall of Memories

In line with Bordalo et al. (2024), the average *distaste* parameter is a weighted average of the employer's "true", static taste for discrimination, and their experience-based discriminatory tastes. The relative weighting of the static $(1-\rho)$ and dynamic component (ρ) is assumed to be exogenous, however this assumption can be relaxed.[4] $d_g(\kappa_{g,t})$ captures the situation-specific discriminatory factor that is influenced by the recall of past experiences.

$$f(d_g, F(\chi_{g,t})) = \underbrace{\sum_{g \in \{A,B\}} L_{g,t}\Big((1-\rho) \cdot d_g + \rho \cdot d_g(\kappa_{g,t})\Big)}_{\text{Distaste}}$$

The current decision of the employer comes with a cue $\kappa_{g,t}$, and the employer has a database $F(\tilde{e})$ of relevant past experiences, consisting of experiences with individuals of group g $(F(\tilde{e}_g))$, and other groups g' $(F(\tilde{e}_{g'})$. The cue $\kappa_{g,t}$ is characterized by several defining attributes, including the group affiliation of the individual engaging with (g), and context (c).[5] Following Kahana (2012); Bordalo et al. (2020, 2024); Miserocchi (2023), recall of experiences is characterized by their similarity to the current setting, and interference. Interference refers to the case when the recall of a given memory is weakened by other memories that are more similar to the cue of the current situation, $\kappa_{g,t}$. The similarity of an experience with an individual of group g in time period $t-k$, $\chi_{g,t-k} \equiv (d_g, c)$ to the cue κ_t of the current decision is given by the multiplicatively separable distance:

$$S(\chi_{g,t-k}, \kappa_{g,t}) \equiv S_1(d_g - d_{g,t-k}|) S_2(|c_t - c_{t-k}|)$$

$S_j : R_+ \Rightarrow R_+$ is decreasing for $j = 1, 2$. A more tractable expression is the exponential

[4] For example, an employer with no experience may only rely on their distaste, and hence $\rho = 0$. As employers get more experience, they may place more emphasis on past experiences, $\rho > 0$.

[5] For simplicity, I drop the price (q) and quantity (q) cues, which were included in Bordalo et al. (2020), from the equations. I therefore focus only on the group affiliation and context. In line with Bordalo et al. (2020), context also captures non-hedonic attributes, such as the timing of the experience.

specification, with the form:

$$S(\chi_{g,t-k}, \kappa_{g,t}) = exp\{-\delta[(d_g - d_{g,t-k})^2 + (c_t - c_{t-k})^2]\}$$

where $\delta \geq 0$ captures the importance of similarity in recall. The likelihood of past experiences being recalled is a function of how similar/relevant they are. Similarity is a function of how close the past experience was to the discriminatory taste of the employer (d_g), and the contextual relatedness (c_t). It follows that the weight assigned to memory $\chi_{g,t-k}$ after the cue $\kappa_{g,t}$ is given is as follows:

$$w(\chi_{g,t-k}, \kappa_{g,t}) = \frac{S(\chi_{g,t-k}, \kappa_{g,t})}{\int S(\tilde{e}, \kappa_{g,t}) dF(\tilde{e})}$$

where $F(\tilde{e})$ captures the entire distribution of past experiences. These experiences are not group-specific and do not have to be exclusively domain-specific: for example, discrimination perceived in the housing market can be a relevant past experience even if the cue $\kappa_{g,t}$ refers to a labor market situation. Aggregating over past memories results in the memory-based discrimination taste of cue κ_t:[6]

$$d_g(\kappa_{g,t}) \equiv \int d_{g,t-k} w(\chi_{g,t-k}, \kappa_{g,t}) dF(\tilde{e})$$

Bias in the recall of memories across groups g can occur for two distinct reasons:

1. Fewer memories for a particular group, $F(\tilde{e}_g) \neq F(\tilde{e}'_g)$;
2. Differential recall of positive and negative experiences across groups due to differing discriminatory tastes $d_g \neq d_{g'}$.

The first bias can arise for a variety of reasons: self-fulfilling prophecies (Coate and Loury, 1993; Glover et al., 2017; Gagnon et al., 2025), limited experimentation (Lepage, 2024), systemic discrimination (Bohren et al., 2025b), or the reliance on stereotypes (Miserocchi, 2023),

[6]This specification does not explicitly differentiate between the timing of when occurred, as timing is implicitly captured in the context parameter, c_t. This can be addressed in two ways: by including the duration distance in the similarity function, of by incorporating a time-specific weighting function, as in Malmendier and Wachter (2024).

among others. The differential availability of memories is likely particularly pronounced in cases where one of the two groups is a minority with whom interaction is limited.

The second bias highlights how, even in the case of an identical underlying distribution of experiences across both groups ($F(\tilde{e}_g) = F(\tilde{e}'_g)$), the weight assigned to memories differs across groups g and g'. This is because the similarity of, and hence weight assigned to, memories is a function of the employer's group-specific distaste parameter d_g. Without loss of generality, we assume that for a given employer e, the fixed distaste d_g is greater for group B than group A: $d_A < d_B$. When recalling memories, even if the memory set for both groups is identical ($F(\tilde{e}_A) = F(\tilde{e}_B)$), the employer is more likely to recall memories in line with their discrimination taste, d_g. Given $d_A < d_B$, the employer will recall more negative past experiences of Group B individuals than Group A individuals, skewing the memory recall to be more negative against Group B individuals. This results in a larger memory-based discrimination parameter ($d_b(\kappa_{b,t}) > d_a(\kappa_{a,t})$, and hence stronger discriminatory behavior in the current decision with cue $\kappa_{g,t}$.

B8 Theory Model: Anticipated Discrimination

Following Buchmann et al. (2024), worker i supplies labor if the expected utility from working is weakly greater than their outside option, which is normalized to zero. While the focus is on the extensive margin of labor supply (working vs. not), insights directly translate to the intensive margin. Workers of group g receive a wage from their employer equivalent to w_g. Expected costs of working for employer k are defined as c_{igk}, and the disutility of working $u_i(.)$ is continuously differentiable and $\frac{\partial u_i}{\partial c_{igk}} > 0$, with $u_i(0) = 0$.

The cost of working for worker i of group g for employer k is c_{igk}, which is a linear combination of the group-specific costs c_g, individual-specific costs c_i, and employer-group specific costs, c_{gk}: $c_{igk} = c_i + c_g + c_{gk}$. Employer-group specific costs are unknown to workers, however workers form beliefs based on the group g of the employer, as well as their past experiences. In particular, employer-group specific costs are larger when the employer is of a different group than the worker ($k_g \neq i_g$). Empirical support for worker's preference to work for employers of a similar background to theirs comes from Hellerstein and Neumark (2008) and Giuliano et al. (2009).

Assuming that the outside option is zero, and job applications are costless, worker i of group g only supplies labor if:

$$\mathbb{E}[w_g - u_i(c_{igk})] \geq 0$$

The worker's expectation of c_{gk} also depends on their past experiences with employers of the same group as employer k. In particular, negative past experiences with employers of the same group g as employer k increase the worker's expectations of the employer-group specific costs: $\frac{\partial c_{gk}}{\partial \chi_{g,k,t}} > 0$. As such, negative past experiences with an employer of the same group g as employer k can increase the employer-group specific costs for worker i (c_{gk}), and hence the cost of working c_{igk}. Given $u_i(.)$ is monotonically increasing in c_{igk}, this increases the expected disutility from work, and hence reduces the labor supplied by worker i. This illustrates how retaliatory discrimination (and negative past experiences) can provide a micro-foundation for anticipated discrimination.

B9 Theory Model: Future Rounds

In each round, two players of different groups ($g \in \{A, B\}$) are randomly paired. Each player only plays the game once. Players can take one of two actions: they can either discriminate (D), or not discriminate (N). Akin to a prisoners dilemma, both players discriminating generates the worst outcome:

	Opponent: N	**Opponent: D**
Player: N	(R, R)	(S, T)
Player: D	(T, S)	(P, P)

Table B15: Payoffs in the Discrimination Prisoner's Dilemma.

with $T > R > P > S$ and $2R > T + S$. If the game is a one-stage game, D is the dominant strategy, and hence both the Player and Opponent will discriminate, resulting in payoff P.

Beliefs and Learning: Considering now that the game will be played across multiple rounds (but each player only plays once themselves), each group h holds a belief $\mu_{t,g}$ about the probability that a member of group g discriminates. Beliefs are updated via a Beta distribution prior with mass $m = \alpha_g + \beta_g$, where α_g is interpreted as the "prior number of discriminatory acts" observed from group g, while β_g can be interpreted as the "prior number of cooperative (non-discriminatory) acts" observed from group g. After observing one new action $x \in \{0, 1\}$ (with $x = 1$ if D), the posterior mean is

$$\mu_{t+1,g} = \frac{m\mu_{t,g} + x}{m + 1}.$$

In expectation, $x = \sigma(\mu_t^g)$, where σ is the strategy of group g.

Social Preferences and Reputational Cost: Individuals care not only about their own payoff but also about the expected payoffs of future in-group members. Let $\lambda \geq 0$ be the weight on in-group welfare, $\delta \in (0, 1)$ the discount factor, and $r > 0$ the retaliation strength (how strongly the out-group responds to the discriminatory reputation of a group). The

effective reputational cost C of discriminating is

$$C \equiv \lambda \cdot \frac{\delta}{1-\delta} \cdot \frac{r}{m+1}.$$

The cost depends positively on λ, δ, r, and depends negatively on m, the strength of the prior (and hence number of observations/experiences).

Best Responses: The myopic, one-round gain from discrimination against belief p is

$$G(p) = (1-p)(T-R) + p(P-S).$$

However, if the game is played over multiple rounds, an individual discriminates if and only if $G(p) \geq C$. This defines a cutoff belief

$$p^\star = \frac{(T-R)-C}{(T-R)-(P-S)}, \qquad p^\star \in [0,1].$$

If $p > p^\star$, the player discriminates, while if $p < p^\star$ the player does not discriminate, and the player is indifferent if $p = p^\star$

Assuming that each group's discrimination probability is a linear response:$\sigma(\mu) = \beta\mu + (1-\beta)p^\star, \qquad \beta \in (0,1)$, belief dynamics then follow: $\mu_{t+1} = \frac{m\mu_t + \sigma(\mu_t)}{m+1}$.

This leads to the following proposition:

Proposition 1. *The unique symmetric steady state of the belief dynamics is $\mu^* = p^\star$. Convergence is linear with rate $\frac{m+\beta}{m+1} \in (0,1)$.*

The proof is available upon request. We therefore have the following comparative statics:

$$\frac{\partial \mu^*}{\partial \lambda} < 0, \qquad \frac{\partial \mu^*}{\partial \delta} < 0, \qquad \frac{\partial \mu^*}{\partial r} < 0, \qquad \frac{\partial \mu^*}{\partial m} > 0.$$

Hence, stronger social preferences for in-group members (λ), discount rate (δ), or retaliation (r) reduce steady-state discrimination, while stronger priors (m) increase it by making reputations less responsive to individual actions.

The *Future Rounds* treatment changed the game from a one-stage to a multiple-rounds game, and as a result introduced a non-zero discount factor. If participants did not have social preferences for in-group members ($\lambda = 0$), we would not expect treatment differences

between the *Status Quo* and *Future Rounds* treatment arms. However, we document differences between the two treatment arms, providing further support for the role of social preferences as a micro-foundation for retaliatory discrimination.

References

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

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-Dimensional Methods and Inference on Structural and Treatment Effects. *Journal of Economic Perspectives*, 28(2):29–50.

Bohren, J. A., Haggag, K., Imas, A., and Pope, D. G. (2025a). Inaccurate Statistical Discrimination: An Identification Problem. *The Review of Economics and Statistics*, pages 1–16.

Bohren, J. A., Hull, P., and Imas, A. (2025b). Systemic Discrimination: Theory and Measurement. *The Quarterly Journal of Economics*.

Bordalo, P., Burro, G., Coffman, K., Gennaioli, N., and Shleifer, A. (2024). Imagining the Future: Memory, Simulation, and Beliefs. *The Review of Economic Studies*.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, Attention, and Choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.

Buchmann, N., Meyer, C., and Sullivan, C. D. (2024). Paternalistic Discrimination. Working paper.

Coate, S. and Loury, G. (1993). Will Affirmative-Action Policies Eliminate Negative Stereotypes? *American Economic Review*, 83(5):1220–40.

Gagnon, N., Bosmans, K., and Riedl, A. (2025). The Effect of Gender Discrimination on Labor Supply. *Journal of Political Economy*, 133(3):1047–1081.

Giuliano, L., Levine, D. I., and Leonard, J. (2009). Manager Race and the Race of New Hires. *Journal of Labor Economics*, 27(4):589–631.

Glover, D., Pallais, A., and Pariente, W. (2017). Discrimination as a Self-Fulfilling Prophecy: Evidence from French Grocery Stores. *The Quarterly Journal of Economics*, 132(3):1219–1260.

Hellerstein, J. K. and Neumark, D. (2008). Workplace Segregation in the United States: Race, Ethnicity, and Skill. *The Review of Economics and Statistics*, 90(3):459–477.

Kahana, M. (2012). *Foundations of Human Memory*. OUP USA.

Kőszegi, B. and Rabin, M. (2006). A Model of Reference-Dependent Preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.

Lepage, L.-P. (2024). Experience-Based Discrimination. *American Economic Journal: Applied Economics*, 16(4):288–321.

Malmendier, U. and Wachter, J. A. (2024). Memory of Past Experiences and Economic Decisions. In *The Oxford Handbook of Human Memory, Two Volume Pack: Foundations and Applications*. Oxford University Press.

Miserocchi, F. (2023). Discrimination through Biased Memory. Working paper.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *The American Economic Review*, 83(5):1281–1302.

Rabin, M. and Schrag, J. L. (1999). First Impressions Matter: A Model of Confirmatory Bias. *The Quarterly Journal of Economics*, 114(1):37–82.

Thaler, M. (2024). The Fake News Effect: Experimentally Identifying Motivated Reasoning Using Trust in News. *American Economic Journal: Microeconomics*, 16(2):1–38.