Supplementary Information for

Costly Exploration Produces Stereotypes with Dimensions of Warmth and Competence

Anonymized Authors

The PDF file includes:

Other Supplementary Materials for this manuscript include the following:

- Movies S1 to S6.
- Data S1 to S3.
- Code S1 to S4.

Note. To facilitate the understanding of our mathematical models of contextual multi-armed bandit and Bayesian inference, we made a tutorial video to explain the process as intuitively as possible. Interested readers can find it from the link for Movie S6.

Materials and Methods

Human Experiments.

In this section, we report additional details about human experiments. All studies are approved by the Institutional Review Board at [mask] University under protocol number 13065. All studies are preregistered at https://osf.io/6p8wu/registrations; author-identifiable information is included. Additional pilot studies for minor tweaks, such as pilot experiments with selected prototypical jobs, stimulus choices with respect to wording, and various intervention prompts are not included in this report but are documented on the preregistration site. Corresponding to the main text, Study S1 reports a systematic analysis of stimulus jobs, Study S2 reports the main hiring experiment, and Study S3 reports mechanism experiments.

Study S1. Stimuli: Jobs and Dimensions.

Participants. We recruited $N = 100$ online workers from the Cloud Research high-quality subject pool who speak English as their first language and are older than 18 years old. This sample size was calculated based on prior work in warmth and competence research (Fiske et al., 2002; Bai et al., 2020), The average age was 41; 50% female, 50% male; 85% White, 7% Black, 4% Asian, and 71% participants hold some college or bachelor's degree, reflecting typical demographic characteristics of online American workers for psychological studies.

Materials. In the survey, we asked participants to rate 24 occupations in terms of their perceived status, cooperation, competence, and warmth. The 24 jobs were selected based on the US Bureau of Labor Statistics Occupation Outlook Handbook published in 2020, social perceptions about common jobs (Fiske & Dupree, 2014), and common beliefs about occupational roles (Koenig & Eagly, 2014). According to prior work, we anticipated the following categories: perceived warm and competent jobs include Doctors, Veterinarians, Professors, Teachers, Psychiatrists, and Computer Scientists; perceived cold but competent jobs include Lawyers, Managers, Financial Advisors, Bankers, Politicians, and Fashion Designers; perceived warm but incompetent jobs include Childcare Aides, Receptionists, Rehabilitation Counselors, Waiters, Homemakers, and

Nursing Assistants; and perceived cold and incompetent jobs include Janitors, Custodians, Truck Drivers, Garbage Collectors, Dishwashers, and Cashiers.

For each job, we asked participants to rate the following eight items from 1 (not at all) to 5 (extremely), as viewed by American society (not their personal beliefs; *7*). Status items: How economically successful/well-educated have members of this occupation been? Cooperation items: If resources go to members of this occupation, to what extent does that take resources away from the rest of society/How much does special treatment given to members of this occupation make things more difficult for other groups in society? Competence items: How capable/confident are members of this occupation? Warmth items: How friendly/trustworthy are members of this occupation? Movie S1 shows participant experiences during the task.

Results. K-means clustering is an unsupervised algorithm that was used to partition the 24 jobs into 4 clusters in which each job belongs to the cluster with the nearest cluster centroid. We used Lloyd's algorithm to iteratively refine the cluster assignments. The algorithm proceeds by alternating between the two steps of assignment and update. During the assignment step, it assigns each observation to the cluster with the nearest mean or the least squared Euclidean distance. During the update step, it recalculates the means for observations assigned to each cluster. The algorithm converges when the assignments no longer change, giving the final cluster assignments as the output. Data in Data S1, Analysis code in Code S1, and results in Fig. S1.

To reduce noise, we removed four ambiguous jobs: computer scientists could belong to high-warmth high-competence or low-warmth high-competence clusters, fashion designers could belong to high-warmth high-competence or low-warmth high competence, nursing assistants could be high-warmth high-competence or high-warmth low-competence, and truck drivers could be low-warmth low-competence or low-warmth high-competence. Hence, in the main experiment, we used the refined 20 jobs as the experimental stimuli. This use of real-world jobs and human judgments improves ecological validity.

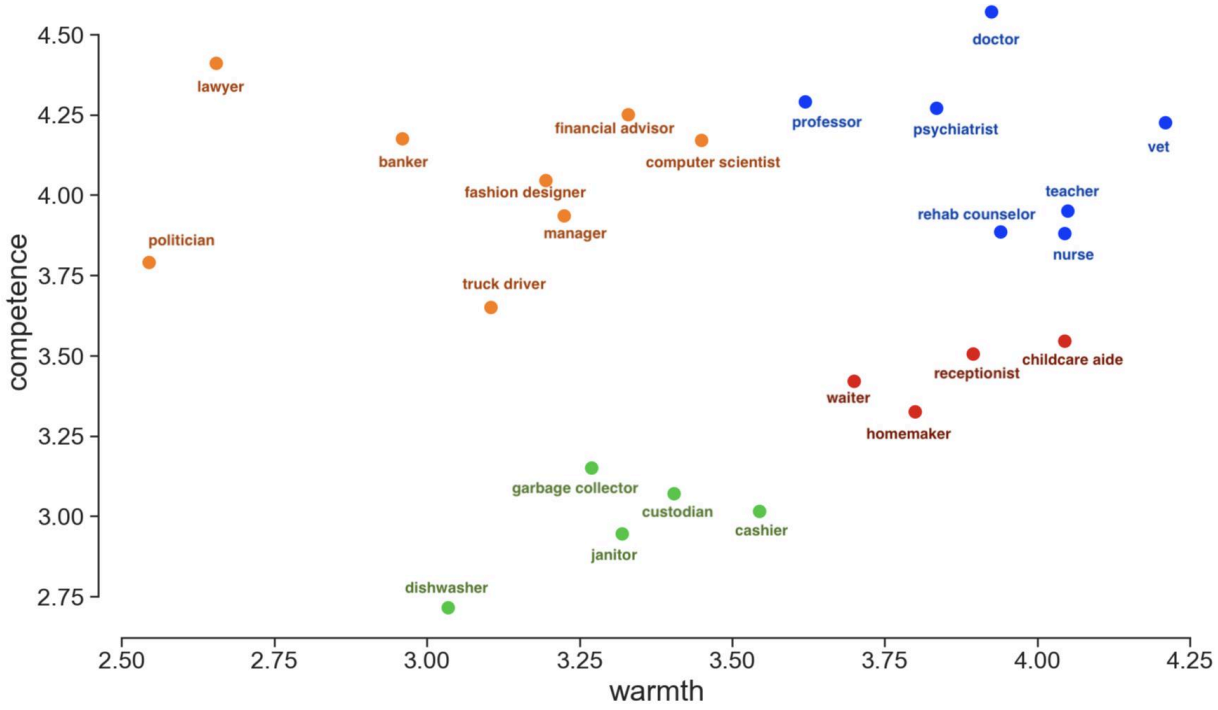

Fig. S1. Emerging clusters of common jobs in American society along dimensions of status or competence, and cooperation or warmth. Values on the axis reflect estimated values, on a scale from 1 to 5, of each job along the two dimensions. Colors indicate cluster assignments as calculated via the KMeans clustering algorithm.

Study S2. Main Experiment: Hiring Consultant for Toma City

Participants. We recruited $N = 403$ online workers from the Cloud Research high-quality subject pool with the same selection criteria as in Study S1. This sample size was calculated based on a pilot study (see details in this anonymized preregistration). The average age was 40; 51% female, 46% male, and 1% non-binary; 74% White, 10% Black, 6% Hispanic, 5% Asian, and 4% multi-racial; 75% of participants hold some college or bachelor's degree; the average political orientation was slightly liberal with an average score of 3.94 on a scale from 1 extremely conservative to 6 extremely liberal.

Power Analysis. Based on our pilot experiment, our sample size is calculated as the follows: Within-game choices: Given \mu_adaptive = 2.12, \sigma_adaptive = 0.31, \my_random = 2.65, \sigma_random = 0.07, \alpha = 0.01, \beta = 0.2, we derive $N = 8$ for each condition, $N = 16$ is

sufficient for two conditions. Out-of-game choice: Given \mu_adaptive = 2.12, \sigma_adaptive = 0.47, \mu_random = 2.58, \sigma_random = 0.27, \alpha = 0.01, \beta = 0.2, we derive $N = 24$ for each condition, thus $N = 48$ is sufficient for two conditions. In terms of stereotypes, status-cooperation dispersion: \mu_adaptive = 2.62, \sigma_adaptive = 1.45, \mu_random = 1.87, \sigma_random = 1.37, \alpha = 0.01, \beta = 0.2, we derive $N = 87$ for each condition, thus $N = 174$ is sufficient for two conditions. The competence-warmth dispersion: \mu_adaptive = 1.94, \sigma_adaptive = 1.15, \mu_random = 1.29, \sigma_random = 1.12, \alpha = 0.01, \beta = 0.2, we derive $N = 73$ for each condition, thus $N = 146$ is sufficient for two conditions. In order to collect a sufficient number of usable data, we decided to go beyond the required minimum power, thus, to run N =200 in each condition for this main study.

Materials. This experiment extends the context-free multi-armed bandit behavioral experiment in (Bai et al., 2020) to test the emergence of multi-dimensional stratification using hiring decisions. In the cover story, participants learn that they will play a game with made-up people from a made-up city. Toma City has around 100,000 residents; they come from four ancestral villages: Tufa, Aima, Reku, and Weki. Participants are hired as a consultant by the mayor of Toma City, and their task is to recommend Toma people for various jobs, out of 20 jobs, in 40 sequential decisions. After each recommendation, participants will learn whether it is a good choice or not. A perfect fit earns 1 point whereas a bad fit earns 0 points. The more points the participants earn, the more bonus they get (1 point = 1 cent), in addition to their base pay ($3 for a 20-minute task). In the game phase, participants see “Job Opening: Doctors” in the first round. They then must select one member from Tufa, Aima, Reku, and Weki groups. On the next page, they see either “You earned 1 point” or “You earned 0 points.” Participants then proceed to the second round, recommend another randomly generated job, and receive feedback. There are 40 trials in total, and after finishing all decisions, participants are asked to answer some questions. First, they are asked one generalization question: “Imagine there are 100 new individuals from each village group applying for the jobs. Enter how many of them you would recommend for each job.” They enter values for the four groups for the twenty jobs. Next, participants are asked about their impressions of the four groups, on a scale from 1 (not at all) to 5 (extremely): “Tufas/Aimas/Rekus/Wekis, in general, seem to be economically successful/interested in helping others/competent or confident/friendly or trustworthy.”

As straightforward as the experiment appears, we made four critical decisions with the goal of minimizing other psychological mechanisms in crafting this experiment. First, we minimized group-serving motivations such as ingroup favoritism (e.g., Tajfel) or social dominance (e.g., Pratto) by assigning no prior group membership to any of our participants. In the spirit of the minimal group paradigm, the use of novel groups achieved this goal. Second, we tried to minimize the cognitive load (e.g., Fiske) by reducing the number of trials in this study, visual representations in addition to abstract group names, and the overall presentation of the hiring interface. Third, to rule out population size as one alternative explanation (e.g., Denrell), in the backend, we prepared all groups with equal population sizes, that is 40 Tufas, 40 Aimas, 40 Rekus, and 40 Wekis will be available if selected. Fourth, just as in the model simulation, we set the true success probability for the four groups in Toma City for the twenty jobs as high and identical, with a 90% success rate for all job-group combinations. This manipulation eliminated the alternative explanation of ground truth differences (e.g., Jussim). The average completion time is 18 minutes. Participants in general enjoyed this task as many left comments saying they had never done a task like this before and it made them think. Data in Data S2 and Movie S2 adaptive and random show participant experiences during the task.

Treatment. The key treatment is the method of exploration. There are two conditions in this hiring experiment. In the experimental condition, participants made hiring decisions as they wished in the infrastructure described above. In the control condition, participants did not have the opportunity to make their own decisions. Instead, they learned that "The mayor will make one recommendation each time, and you can observe the mayor's decision." From the backend, the game infrastructure selected each group randomly at each time, to mimic the experience of random-decision. After 40 trials of hiring decisions, participants in both conditions continued to make future hires and provided impressions about the groups as described above.

Results. We estimated OLS regressions in which we regressed our outcomes – choice entropy during the 40-trial game, choice entropy of future hires, dispersion of estimated status and cooperation, and dispersion of estimated competence and warmth – over our treatment indicator (β), controlling for respondents' age, gender, race, education, and political orientation. Our main quantity of interest is on identifying β representing the average treatment effect of the

exploration strategy on participants' hire decisions and impressions about Toma groups. Results summary in Table S1. Analysis code in Code S2.

Table S1. Average Treatment Effects.

	β	*t*	*p* > | *t* |	[.025, .975]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive)	2.165	162.222	.000	[2.138, 2.191]
Random	0.480	25.360	.000	[0.443, 0.518]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive, Female, Black)	2.205	33.446	.000	[2.076, 2.335]
Random	0.476	24.311	.000	[0.437, 0.514]
Gender (=Male)	-0.017	-0.880	0.379	[-0.056, 0.021]
Gender (=Nonbinary)	0.097	1.006	0.315	[-0.093, 0.288]
Race (=Asian)	-0.079	-1.448	0.148	[-0.187, 0.028]
Race (=Caucasian)	-0.084	-2.559	0.011	[-0.148, -0.019]
Race (=Hispanic)	-0.042	-0.815	0.416	[-0.142, 0.059]
Race (=Multiracial)	-0.009	-0.152	0.879	[-0.125, 0.107]
Age	-0.000	-0.109	0.913	[-0.002, 0.002]
Education	0.004	0.370	0.711	[-0.017, 0.024]
Political Orientation	0.007	1.017	0.310	[-0.007, 0.021]
Choice entropy: 400 future hires				
Intercept (=Adaptive)	2.169	89.752	.000	[2.122, 2.217]
Random	0.438	12.754	.000	[0.370, 0.505]

Choice entropy: 400 future hires				
Intercept (=Adaptive, Female, Black)	2.331	19.363	.000	[2.094, 2.568]
Random	0.432	12.104	.000	[0.362, 0.503]
Gender (=Male)	0.014	0.389	0.698	[-0.057, 0.085]
Gender (=Nonbinary)	0.240	1.359	0.175	[-0.107, 0.588]
Race (=Asian)	-0.152	-1.521	0.129	[-0.349, 0.045]
Race (=Caucasian)	-0.132	-2.203	0.028	[-0.249, -0.014]
Race (=Hispanic)	-0.168	-1.796	0.073	[-0.351, 0.016]
Race (=Multiracial)	-0.064	-0.600	0.549	[-0.275, 0.147]
Age	-0.002	-1.128	0.260	[-0.005, 0.001]
Education	-0.007	-0.375	0.708	[-0.045, 0.030]
Political Orientation	0.011	0.850	0.396	[-0.014, 0.036]

Stereotype dispersion: cooperation and status				
Intercept (=Adaptive)	2.637	27.428	.000	[2.448, 2.826]
Random	-0.747	-5.476	.000	[-1.016, -0.479]

Stereotype dispersion: cooperation and status				
Intercept (=Adaptive, Female, Black)	3.360	6.978	.000	[2.413, 4.307]
Random	-0.753	-5.271	.000	[-1.034, -0.472]
Gender (=Male)	0.037	0.253	0.801	[-0.247, 0.320]
Gender (=Nonbinary)	-0.116	-0.164	0.869	[-1.507, 1.275]
Race (=Asian)	-0.047	-0.117	0.907	[-0.833, 0.740]
Race (=Caucasian)	0.126	0.527	0.598	[-0.344, 0.596]
Race (=Hispanic)	0.392	1.050	0.294	[-0.342, 1.126]

Race (=Multiracial)	0.087	0.202	0.840	[-0.757, 0.931]
Age	-0.004	-0.721	0.472	[-0.016, 0.008]
Education	-0.127	-1.660	0.098	[-0.277, 0.023]
Political Orientation	-0.056	-1.096	0.274	[-0.157, 0.045]

Stereotype dispersion: warmth and competence				
Intercept (=Adaptive)	1.861	21.520	.000	[1.691, 2.031]
Random	-0.327	-2.665	.008	[-0.568, -0.086]
Stereotype dispersion: warmth and competence				
Intercept (=Adaptive, Female, Black)	2.397	5.505	.000	[1.541, 3.254]
Random	-0.343	-2.653	.008	[-0.597, -0.089]
Gender (=Male)	-0.002	-0.018	0.986	[-0.259, 0.255]
Gender (=Nonbinary)	-0.201	-0.314	0.753	[-1.459, 1.057]
Race (=Asian)	-0.204	-0.564	0.573	[-0.916, 0.507]
Race (=Caucasian)	0.050	0.233	0.816	[-0.375, 0.475]
Race (=Hispanic)	-0.013	-0.039	0.969	[-0.677, 0.651]
Race (=Multiracial)	0.023	0.059	0.953	[-0.741, 0.786]
Age	-0.007	-1.340	0.181	[-0.018, 0.003]
Education	-0.056	-0.814	0.416	[-0.192, 0.080]
Political Orientation	-0.013	-0.269	0.788	[-0.104, 0.079]

Note. Estimates are based on an OLS model without (first panels) and with (second panels) covariate variables of age, gender, race, education, and political orientation. $N = 403$.

We plotted exemplar participants from the 40-trial condition in the main text, here we provide their hiring decisions in future jobs and along dimensions of status and cooperation (Figs. S2 - S5). For complete participants, see Data S2 and use Code S2.

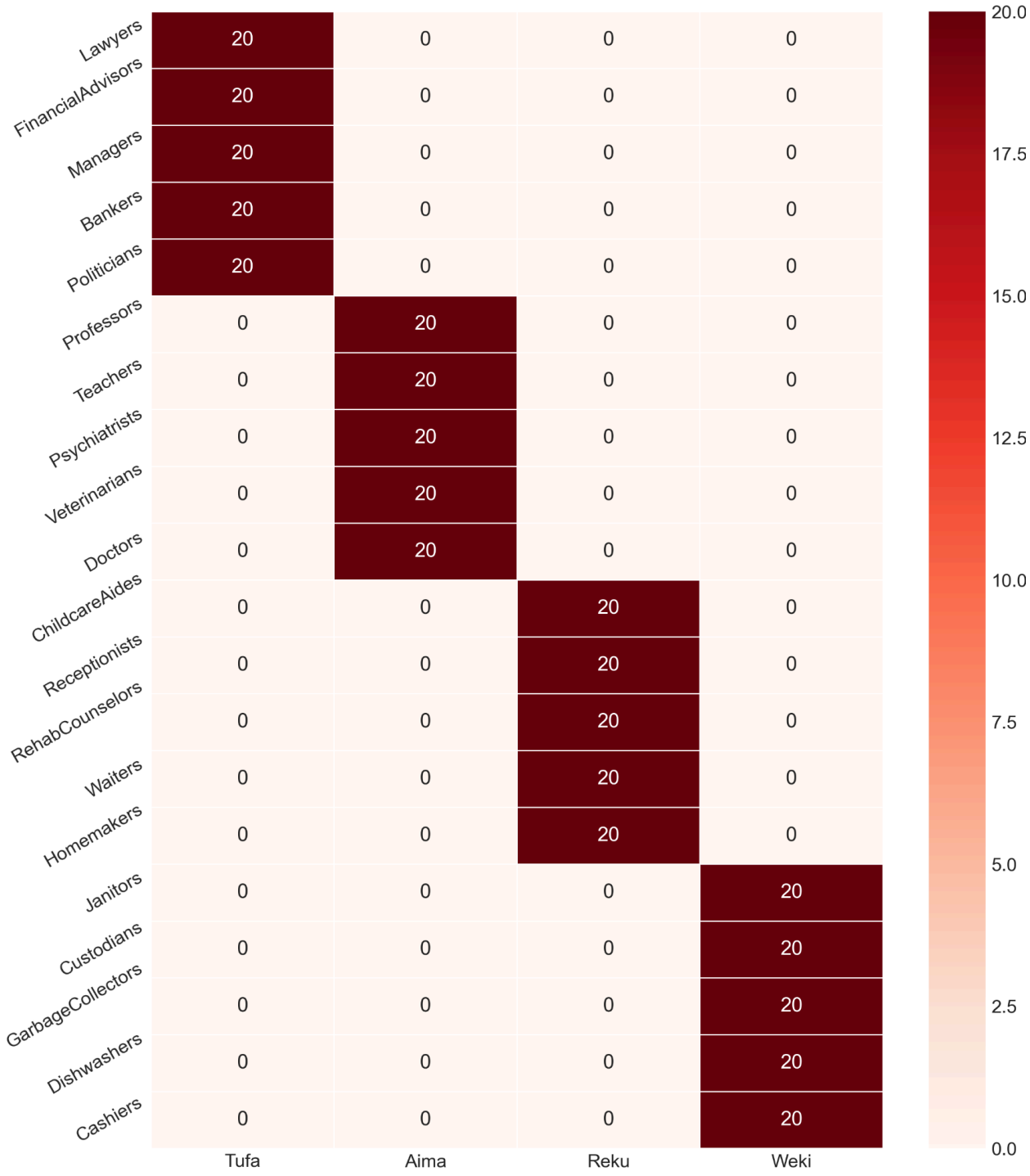

Fig. S2. Illustrative participants *ID* = 153 in the adaptive exploration condition made more stratified and less diverse future hiring decisions.

Fig. S3. Illustrative participants *ID* = 281 in the random exploration condition made less stratified and more equal future hiring decisions.

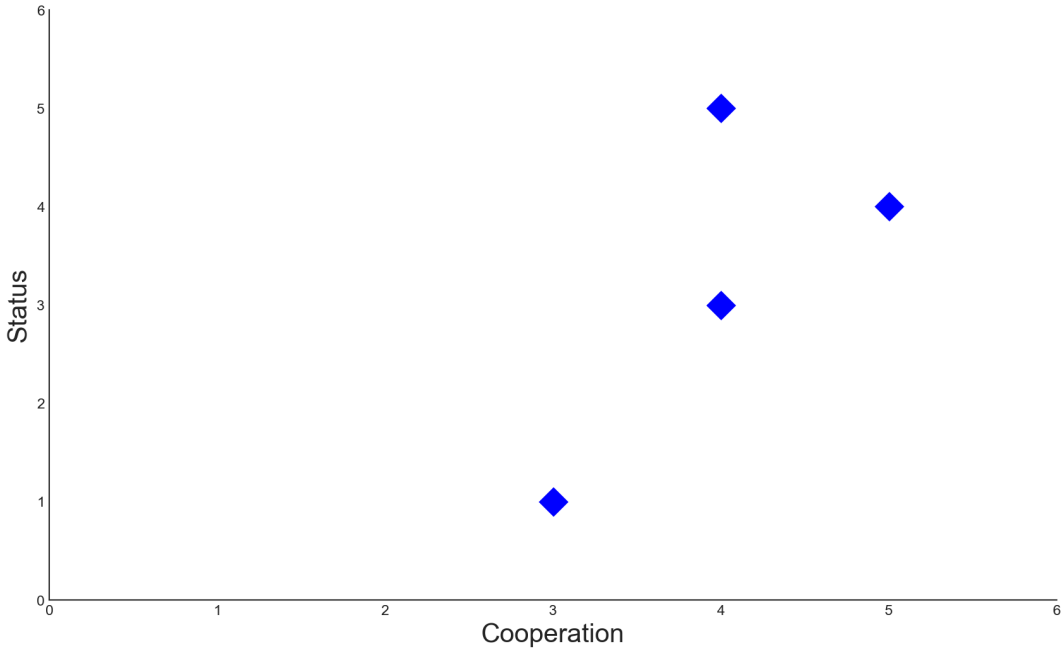

Fig. S4. Illustrative participants *ID* = 153 in the adaptive exploration condition showed more dispersed mental maps along the social status and cooperative intent dimensions.

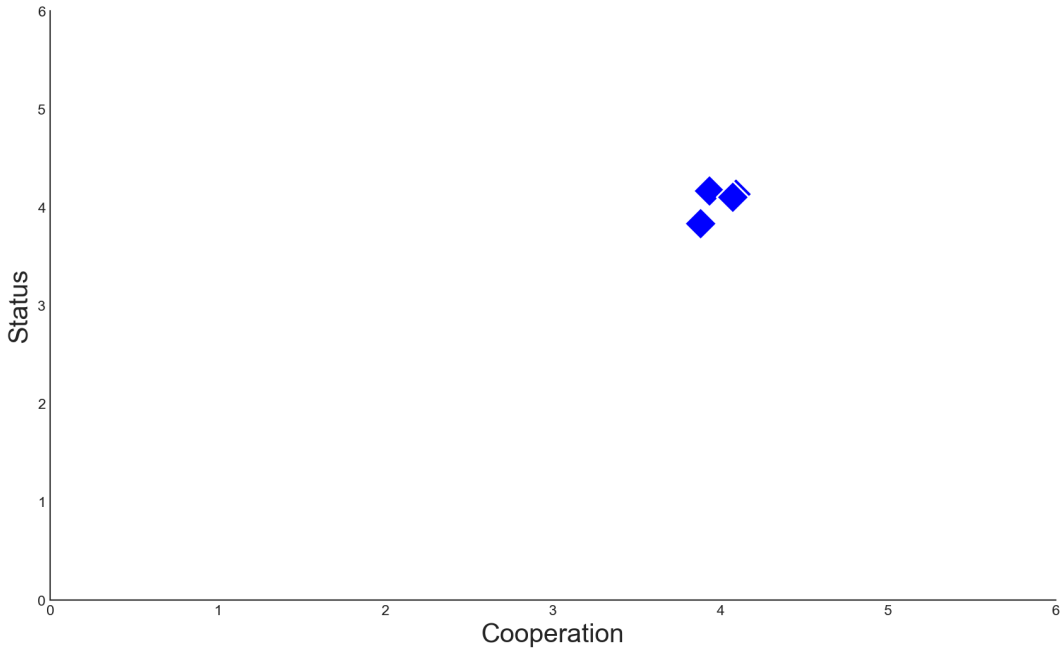

Fig. S5. Illustrative participants *ID* = 281 in the random exploration condition showed less dispersed mental maps along the social status and cooperative intent dimensions.

Study S3. Mechanism Experiment: Interventions for Exploration

Participants. We recruited $N = 807$ online workers from Connect, a new platform for paid online studies hosted by Cloud Research. The reason to switch from Cloud Research to Connect is rather practical as Cloud Research only allows large-scale payment if we recruit MTurk participants, otherwise we have a daily payment limitation. We used the same selection criteria as in previous experiments, while also imposing gender balance given it is a convenient function on Connect. The sample size was again to ensure 200 participants per condition, to be consistent with Study 2 (see details in this anonymized preregistration). The average age was 40; 50% female, 50% male; 67% White, 11% Black, 9% Asian, 6% Hispanic, and 4% multi-racial; 71% of participants hold some college or bachelor's degree; the average political orientation was slightly liberal with an average score of 3.98 on a scale from 1 extremely conservative to 6 extremely liberal. The average score for this task was 4.7 out of 5, indicating acceptable engagement among participants. In addition to the main mechanism, we also piloted similar experiments to nail down the wording for the interventions; details can be found in pre-registration reports titled "mechanism pilot."

Power Analysis. This sample size follows the main experiment with 200 participants in the adaptive exploration condition. Given that we have 3 new intervention conditions, we quadruped the sample size to collect a minimum of $N = 800$.

Materials. The experiment materials were largely identical to the main experiment described above. All data from this experiment are in Data S3. The baseline condition is identical, whereas the intervention conditions contain different designs mainly in the cover story, the hiring decisions, and the feedback pages, as follows.

In the exploration bonus condition, after participants read general instructions about the city, jobs, their roles, and points, and before they started the game, they saw a new page titled "Diversity Bonus." They read: "Recently Toma City launched a hiring initiative. The mayor will pay an extra bonus for more variety in who you hire. The bonus decreases for each hire of a person from a group that has previously been hired for that job. Your total earnings will be the sum of rewards from making suitable hires and the diversity bonus." After they made one hiring

decision, they received feedback as participants in the baseline condition. However, the bonus was now calculated as the sum of their actual reward (1 or 0) and a diversity bonus of $1/(N+1)$, N being the times of the same group being recommended for the same cluster of jobs. Therefore, rather than displaying an integer value of 1 or 0, this page showed floating numbers such as 2 (if the selection is completely new, 1/(0+1), and the selection is good, 1), 1.5 (if the selection is good, but it is the second time being recommended, 1/(1+1)), and so on. Movie S3 shows participants' experiences in this condition.

In the reward rate condition, the instructions did not change at all. The only difference was the underlying expected reward which changed from 90% to 10%. For example, among 40 available members from village Tufa, a 90% success rate means 36 of them would return a reward of 1 for the jobs being recommended, versus a 10% success rate means only 4 of them would be successful. Note that participants did not have access to this information before the game, and the only way to figure this out was through experience. Movie S4 shows participants' experiences in this condition.

In the random holdout condition, after participants read general instructions and before they started the game, they saw a new page titled "Travel Restrictions." They read: "Due to recent travel restrictions, not all villagers are able to come to work at all times. Sometimes your selected members might become unavailable; if so, you need to choose from the available members." To reflect this change, on their hiring page, 90% of all trials made 2 out of 4 groups not clickable indicating those two groups are not available due to travel restrictions. It was a random selection of which two groups to disable and on which trials participants encountered travel restrictions. Movie S5 shows participants' experiences in this condition.

Results. We used the same analysis strategy as in Study 2 in which we estimated OLS regressions by regressing our outcomes - choice entropy during the 40-trial game, choice entropy of future hires, dispersion of estimated status and cooperation, and dispersion of estimated competence and warmth - over our treatment indicator (β), controlling for respondents' age, gender, race, education, and political orientation. Our main quantity of interest is on identifying β representing the average treatment effect of the baseline default hiring and each of the three proposed interventions - exploration bonus, reward rate, random holdout - on participants' hire decisions and impressions about Toma groups. Given that we test for three hypotheses, the

analysis used Bonferroni correction of the alpha level at 0.01. More precisely, the threshold should be the original alpha value divided by the number of comparisons, that is 0.05/3 = 0.0167. Results summary in Tables S2 - S4. Analysis code in Code S3.

Table S2. Average Treatment Effects (Exploration Bonus)

	β	*t*	*p* > | *t* |	[.025, .975]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive)	2.133	118.962	.000	[2.097, 2.168]
Exploration Bonus	0.400	16.168	.000	[0.352, 0.449]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive, Female, Black)	2.156	27.396	.000	[2.001, 2.310]
Exploration Bonus	0.396	15.484	.000	[0.346, 0.447]
Gender (=Male)	-0.022	-0.820	0.413	[-0.073, 0.030]
Race (=Asian)	-0.114	-2.108	0.036	[-0.221, -0.008]
Race (=Caucasian)	-0.153	-3.410	0.001	[-0.241, -0.065]
Race (=Hispanic)	-0.076	-1.223	0.222	[-0.197, 0.046]
Race (=Multiracial)	-0.125	-1.778	0.076	[-0.263, 0.013]
Age	0.001	1.033	0.302	[-0.001, 0.003]
Education	0.008	0.943	0.346	[-0.009, 0.026]
Political Orientation	0.011	1.281	0.201	[-0.006, 0.028]
Choice entropy: 400-trial future hires				
Intercept (=Adaptive)	2.166	79.479	.000	[2.113, 2.220]
Exploration Bonus	0.368	9.768	.000	[0.294, 0.442]
Choice entropy: 400-trial future hires				

Intercept (=Adaptive, Female, Black)	1.955	16.418	.000	[1.721, 2.189]
Exploration Bonus	0.366	9.445	.000	[0.290, 0.442]
Gender (=Male)	-0.028	-0.710	0.478	[-0.106, 0.050]
Race (=Asian)	-0.078	-0.954	0.341	[-0.240, 0.083]
Race (=Caucasian)	-0.124	-1.824	0.069	[-0.258, 0.010]
Race (=Hispanic)	-0.109	-1.163	0.246	[-0.294, 0.075]
Race (=Multiracial)	0.081	0.756	0.450	[-0.129, 0.290]
Age	0.003	2.068	0.039	[0.000, 0.007]
Education	0.020	1.513	0.131	[-0.006, 0.047]
Political Orientation	0.029	2.272	0.024	[0.004, 0.054]

Stereotype dispersion: cooperation and status				
Intercept (=Adaptive)	2.618	24.191	.000	[2.405, 2.830]
Exploration Bonus	-0.774	-5.179	.000	[-1.068, -0.480]

Stereotype dispersion: cooperation and status

Intercept (=Adaptive, Female, Black)	2.628	5.506	.000	[1.690, 3.567]
Exploration Bonus	-0.788	-5.074	.000	[-1.093, -0.482]
Gender (=Male)	0.104	0.657	0.512	[-0.208, 0.416]
Race (=Asian)	0.483	1.466	0.143	[-0.165, 1.132]
Race (=Caucasian)	0.039	0.142	0.887	[-0.497, 0.575]
Race (=Hispanic)	-0.325	-0.865	0.388	[-1.065, 0.414]
Race (=Multiracial)	0.254	0.427	0.552	[-0.585, 1.093]
Age	0.006	0.947	0.344	[-0.007, 0.019]
Education	-0.015	-0.284	0.776	[-0.121, 0.090]
Political Orientation	-0.080	-1.543	0.124	[-0.181, 0.022]

Stereotype dispersion: warmth and competence				
Intercept (=Adaptive)	1.888	20.060	.000	[1.703, 2.073]
Exploration Bonus	-0.340	-2.619	.009	[-0.596, -0.085]

Stereotype dispersion: warmth and competence

Intercept (=Adaptive, Female, Black)	1.941	4.648	.000	[1.130, 2.762]
Exploration Bonus	-0.325	-2.392	.017	[-0.592, -0.058]
Gender (=Male)	0.041	0.294	0.769	[-0.232, 0.314]
Race (=Asian)	0.691	2.395	0.017	[0.124, 1.258]
Race (=Caucasian)	0.423	1.775	0.077	[-0.046, 0.892]
Race (=Hispanic)	0.112	0.340	0.734	[-0.535, 0.759]
Race (=Multiracial)	0.276	0.739	0.460	[-0.458, 1.010]
Age	0.000	-0.017	0.987	[-0.011, 0.011]
Education	-0.041	-0.872	0.384	[-0.134, 0.051]
Political Orientation	-0.072	-1.592	0.112	[-0.160, 0.017]

Note. Estimates are based on an OLS model without (first panels) and with (second panels) covariate variables of age, gender, race, education, and political orientation. $N = 194$ in baseline and $N = 214$ in exploration bonus conditions.

Table S3. Average Treatment Effects (Lower Reward)

	β	*t*	*p* > | *t* |	[.025, .975]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive)	2.133	128.688	.000	[2.100, 2.165]
Lower Reward	0.414	17.797	.000	[0.368, 0.460]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive, Female, Black)	2.236	31.890	.000	[2.098, 2.374]
Lower Reward	0.410	17.254	.000	[0.363, 0.456]
Gender (=Male)	-0.026	-1.061	0.289	[-0.073, 0.022]
Race (=Asian)	-0.153	-2.879	0.004	[-0.257, -0.048]
Race (=Caucasian)	-0.178	-4.345	0.000	[-0.259, -0.098]
Race (=Hispanic)	-0.130	-2.174	0.030	[-0.305, -0.037]
Race (=Multiracial)	-0.171	-2.502	0.013	[-0.305, -0.037]
Age	0.001	0.420	0.675	[-0.002, 0.002]
Education	0.011	1.480	0.140	[-0.004, 0.027]
Political Orientation	0.003	0.403	0.687	[-0.012, 0.018]
Choice entropy: 400-trial future hires				
Intercept (=Adaptive)	2.166	75.212	.000	[2.110, 2.223]
Lower Reward	0.356	8.788	.000	[0.276, 0.435]
Choice entropy: 400-trial future hires				

Intercept (=Adaptive, Female, Black)	2.054	16.648	.000	[1.811, 2.296]
Lower Reward	0.353	8.449	.000	[0.271, 0.435]
Gender (=Male)	-0.069	-1.639	0.102	[-0.153, 0.014]
Race (=Asian)	-0.065	-0.695	0.488	[-0.248, 0.119]
Race (=Caucasian)	-0.087	-1.200	0.231	[-0.229, 0.055]
Race (=Hispanic)	-0.063	-0.603	0.547	[-0.269, 0.143]
Race (=Multiracial)	-0.144	-1.199	0.231	[-0.380, 0.092]
Age	0.002	1.179	0.239	[-0.001, 0.006]
Education	0.026	1.866	0.063	[-0.001, 0.052]
Political Orientation	0.014	1.055	0.292	[-0.012, 0.041]

Stereotype dispersion: cooperation and status				
Intercept (=Adaptive)	2.618	23.833	.000	[2.402, 2.834]
Lower Reward	-1.073	-6.953	.000	[-1.377, -0.770]

Stereotype dispersion: cooperation and status

Intercept (=Adaptive, Female, Black)	2.194	4.626	.000	[1.261, 3.126]
Lower Reward	-1.096	-6.825	.000	[-1.411, -0.780]
Gender (=Male)	0.098	0.599	0.550	[-0.233, 0.418]
Race (=Asian)	0.782	2.181	0.030	[0.077, 1.486]
Race (=Caucasian)	0.348	1.254	0.211	[-0.198, 0.894]
Race (=Hispanic)	0.522	1.295	0.196	[-0.270, 1.314]
Race (=Multiracial)	1.001	2.167	0.031	[0.092, 1.909]
Age	0.007	1.083	0.280	[-0.006, 0.021]
Education	0.041	0.776	0.438	[-0.063, 0.144]
Political Orientation	-0.111	-2.137	0.033	[-0.213, -0.009]

Stereotype dispersion: warmth and competence				
Intercept (=Adaptive)	1.888	20.209	.000	[1.704, 2.071]
Lower Reward	-0.630	-4.796	.000	[-0.888, -0.371]

Stereotype dispersion: warmth and competence

Intercept (=Adaptive, Female, Black)	1.754	4.340	0.000	[0.959, 2.548]
Lower Reward	-0.657	-4.801	0.000	[-0.925, -0.388]
Gender (=Male)	-0.111	-0.801	0.424	[-0.384, 0.162]
Race (=Asian)	0.734	2.405	0.017	[0.134, 1.334]
Race (=Caucasian)	0.338	1.430	0.154	[-0.127, 0.804]
Race (=Hispanic)	0.588	1.712	0.088	[-0.087, 1.263]
Race (=Multiracial)	0.580	1.475	0.141	[-0.194, 1.354]
Age	0.001	0.182	0.855	[-0.010, 0.012]
Education	-0.003	-0.058	0.954	[-0.091, 0.086]
Political Orientation	-0.049	-1.119	0.264	[-0.136, 0.037]

Note. Estimates are based on an OLS model without (first panels) and with (second panels) covariate variables of age, gender, race, education, and political orientation. $N = 194$ in baseline and $N = 199$ in lower reward conditions.

Table S4. Average Treatment Effects (Random Holdout)

	β	*t*	*p* > | *t* |	[.025, .975]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive)	2.133	127.071	.000	[2.100, 2.166]
Random Holdout	0.333	14.132	.000	[0.286, 0.379]
Choice entropy: 40-trial current hires				
Intercept (=Adaptive, Female, Black)	2.292	34.362	.000	[2.161, 2.423]
Random Holdout	0.321	13.200	.000	[0.273, 0.369]
Gender (=Male)	-0.049	-2.047	0.041	[-0.096, -0.002]
Race (=Asian)	-0.149	-2.588	0.010	[-0.261, -0.036]
Race (=Caucasian)	-0.176	-4.492	0.000	[-0.253, -0.099]
Race (=Hispanic)	-0.115	-1.857	0.064	[-0.237, 0.007]
Race (=Multiracial)	-0.199	-2.614	0.009	[-0.349, -0.049]
Age	0.001	0.136	0.892	[-0.008, 0.023]
Education	0.001	0.961	0.337	[-0.008, 0.023]
Political Orientation	-0.002	-0.276	0.783	[-0.018, 0.013]
Choice entropy: 400-trial future hires				
Intercept (=Adaptive)	2.166	72.792	.000	[2.108, 2.225]
Random Holdout	0.180	4.306	.000	[0.098, 0.262]
Choice entropy: 400-trial future hires				

Intercept (=Adaptive, Female, Black)	2.242	18.668	.000	[2.006, 2.479]
Random Holdout	0.153	3.509	.001	[0.067, 0.239]
Gender (=Male)	-0.029	-0.671	0.503	[-0.114, 0.056]
Race (=Asian)	-0.165	-1.594	0.112	[-0.368, 0.039]
Race (=Caucasian)	-0.181	-2.559	0.011	[-0.593, -0.154]
Race (=Hispanic)	-0.373	-3.341	0.001	[-0.593, -0.154]
Race (=Multiracial)	-0.050	-0.363	0.717	[-0.320, 0.220]
Age	0.001	0.160	0.873	[-0.003, 0.004]
Education	0.029	2.035	0.043	[0.001, 0.058]
Political Orientation	0.001	0.058	0.954	[-0.027, 0.029]

Stereotype dispersion: cooperation and status				
Intercept (=Adaptive)	2.618	24.555	.000	[2.408, 2.827]
Random Holdout	-0.615	-4.107	.000	[-0.909, -0.320]

Stereotype dispersion: cooperation and status

Intercept (=Adaptive, Female, Black)	2.967	6.898	.000	[2.121, 3.813]
Random Holdout	-0.607	-3.877	.000	[-0.915, -0.299]
Gender (=Male)	-0.126	-0.814	0.416	[-0.431, 0.179]
Race (=Asian)	0.879	2.373	0.018	[0.151, 1.606]
Race (=Caucasian)	0.114	0.452	0.652	[-0.383, 0.612]
Race (=Hispanic)	0.314	0.785	0.433	[-0.473, 1.101]
Race (=Multiracial)	0.533	1.084	0.279	[-0.433, 1.499]
Age	0.003	0.047	0.962	[-0.012, 0.012]
Education	0.012	0.238	0.812	[-0.089, 0.114]
Political Orientation	-0.135	-2.638	0.009	[-0.235, -0.034]

Stereotype dispersion: warmth and competence				
Intercept (=Adaptive)	1.888	20.723	0.000	[1.709, 2.067]
Random Holdout	-0.328	-2.562	0.011	[-0.579, -0.076]

Stereotype dispersion: warmth and competence

Intercept (=Adaptive, Female, Black)	2.189	5.893	0.000	[1.459, 2.920]
Random Holdout	-0.315	-2.327	0.021	[-0.581, -0.049]
Gender (=Male)	-0.161	-1.204	0.229	[-0.424, 0.102]
Race (=Asian)	0.944	2.952	0.003	[0.315, 1.572]
Race (=Caucasian)	0.237	1.083	0.280	[-0.193, 0.666]
Race (=Hispanic)	0.348	1.006	0.315	[-0.332, 1.027]
Race (=Multiracial)	0.114	0.268	0.789	[-0.721, 0.948]
Age	-0.004	-0.761	0.447	[-0.014, 0.006]
Education	0.023	0.525	0.600	[-0.064, 0.111]
Political Orientation	-0.098	-2.230	0.026	[-0.185, -0.012]

Note. Estimates are based on an OLS model without (first panels) and with (second panels) covariate variables of age, gender, race, education, and political orientation. $N = 194$ in baseline and $N = 200$ in random holdout conditions.

Computational Simulations

Formalism: Contextual Multi-Armed Bandit and Bayesian Inference.

Using the job recommendation example in the main text, we consider how a rational agent in a contextual multi-armed bandit setting should solve this problem. We show that strong differences in the allocation of groups to societal positions can be produced by rational agents in the absence of any real inter-group differences and that these agents form stereotype contents along multiple dimensions. The goal of this section is to provide details on the computational modeling approach in the main text. We made a movie (no audio) presentation to animate this mathematical model, readers can watch it in Movie S6.

The agent has access to a discrete number of groups G, and interacts with candidate jobs across discrete trials $t = 1, 2, \ldots, T$, where the reward is whether or not the job is a good fit for the recommended group member. The jobs are characterized by contextual information $x;\ x \in \Re^D$, that is, jobs have as many as D dimensions of features. Here, in our example, we set D to be 2, corresponding to levels of status and cooperation. For simplicity, we set the number of groups G to be 4, but it can apply to larger finite numbers, as many as the social groups we have in society.

At trial t, the agent observes the current job characterized by its features x_t, and the available groups. The goal of the agent is to provide the job with a person from one group who may fit the position. The groups are thus the arms of the bandit, the selection of a group is the action, and the context is the job features x_t. After making the recommendation, the agent receives a reward, r_t. If the person selected is a good fit for the job, then r_t equals 1, if not, r_t equals 0. The rewards follow a distribution which can be characterized by the context, x, and parameters, θ_g, written $P(r;\ x,\ \theta_g)$ where g is the group to which those parameters correspond. θ is also in $\Re^D$, so 2 dimensions in our example. The expected reward for each group, g, can be written as:

$$E\,[r_g \mid x,\ \theta_g] \ = \ f(x^T\theta_g), \text{ where } f(.) \ = \ exp(.) \ / \ (1 \ + \ exp(.)). \ [1]$$

where $x^T\theta_g$ is the inner product of these two vectors, being a linear function of x with parameters θ_g. The parameter vector θ_g thus encodes how the dimensions of x, corresponding to the features of the jobs, are weighted for group g when predicting whether a member of that group will be successful in the job. Here, $f(.)$ is a sigmoid function that transforms an arbitrary value into a continuous value in $[0,\ 1]$, to give us $P(r;\ x,\ \theta_g)$.

For $t \ = \ 1,\ 2,\ ...,\ T$, the agent observes past t observations of the contexts, the actions chosen, and their corresponding rewards $(x_t,\ g_t,\ r_t)$. Importantly, no payoff information is revealed for the unchosen groups, $g \ \neq \ g_t$. The objective is to find a solution that minimizes the cumulative regret; the regret is the expected difference between the optimal reward received by always playing the optimal group, g_t^*, and the reward received by following the actually chosen group, g_t. Thus, the cumulative regret at the end of the game, $R(T)$ can be written as:

$$R(T) \ = \ \sum_{t=1}^{T} E[r_{g^*}|x_t, \theta_{g^*}] \ - \ E[r_g|x_t, \theta_g]. \ [2]$$

Finding the optimal solution to this problem requires balancing between exploration and exploitation. While there are no known optimal solutions to contextual bandits, we focus on an approach known as Thompson sampling (*26, 27*) which generalizes an optimal solution to the standard multiarmed bandit. Thompson sampling has previously been used to show how adaptive exploration can produce stereotypes in a simpler context-free multiarmed bandit setting (*29*) and has been shown to be a good model of human choices on contextual bandit tasks (*30*).

Using the same job recommendation example, Thompson sampling for contextual bandit can be defined in terms of the Bayesian solution to the problem of estimating θ_g. For each group, g, if we know θ_g, then applying any context x_t, we can derive the expected reward via Equation 1. But we do not know the parameters θ_g, so the goal is to estimate them. At time step t, first, a

prior distribution $P(\theta_g)$ represents uncertainty over the parameter space and the likelihood function $P(r_g|x_t, \theta_g)$ represents the probability of reward given a context x_t and a parameter θ_g. Applying Bayes' rule, the posterior distribution over θ_g is given by:

$$P(\theta_g|r_g, x_t) \propto P(r_g|x_t, \theta_g)\, P(\theta_g). \text{ [3]}$$

The posterior distribution, therefore, represents the updated beliefs about the parameters θ_g after incorporating the new evidence and the prior belief. Next, a sample $\theta_{t+1,g}$ is randomly drawn from this posterior, corresponding to a stochastic estimate of θ_g after t time steps. The agent follows this procedure – estimate θ_g, draw a random sample $\theta_{t+1,g}$ – for all groups, and plays the group for which the predicted probability of reward is highest. This is equivalent to sampling each group with a probability corresponding to the posterior probability that group is most likely to generate a reward, which is Thompson sampling.

Specifically, in Equation 3 we assume the prior, $P(\theta_g)$, follows a Gaussian distribution $N(\mu_0, S_0)$ and the likelihood, $P(r_g|x_t, \theta_g)$, follows a Bernoulli distribution, with the joint probability mass function over the rewards:

$$\prod_{t=1}^{T} P(r_t = 1|x_t, \theta_t) = \prod_{t=1}^{T} [1 / (1 + e^{-\theta_t^T x_t})]^{r_t} [e^{-\theta_t^T x_t} / (1 + e^{-\theta_t^T x_t})]^{1-r_t}. \text{ [4]}$$

The posterior distribution derived from this joint probability distribution is intractable, hence, we use Laplace's method to approximate the posterior distribution with a multivariate Gaussian distribution with a diagonal covariance matrix. The mean of this distribution is the maximum-a-posteriori estimate, and the inverse variance of each feature is the curvature (Algorithms 3 in *28*).

Simulation Results for Main Hypothesis:

In this section, we present predictions derived from the above model with simulation data. Again, for all simulations, we use Bayesian logistic regression to estimate the function between job features and groups and use Thompson sampling to make decisions about how to solve the explore-exploit dilemma. As defined above, the context vector has two dimensions, corresponding to status and trust with binary features: $\{1, 1\}$ indicates high status and high trust jobs such as doctors, $\{1, -1\}$ indicates high status low trust jobs such as lawyers, $\{-1, 1\}$ indicates low status high trust jobs such as childcare aides, and $\{-1, -1\}$ indicates low status and low trust jobs such as janitors. There are four groups; whose reward distributions are independent from each other. The current model has the same intercept for all groups as we assume no group-level differences. The underlying expected reward probability centers around 0.9, $N(0.9,\ 0.001)$, for all groups. We made the variance small because we assume all groups are equally and highly likely to be successful. The agent starts with a prior belief follows a normal distribution of $N(0,\ 1)$. With this set up, we ran 100 simulations. Within each simulation, the Bayesian agent played 40 rounds of the game.

The critical prediction from our model is that the Bayesian agents will end up creating a biased social structure such that certain groups are selectively recommended to certain jobs, compared to that produced by agents who make choices at random. Two key outcome variables quantify this hypothesis: selective recommendation patterns and dispersed mental representations.

First, for recommendation choices, we predict Bayesian agents do not recommend jobs indifferently, but rather should differentially recommend certain groups to do certain kinds of jobs. This occurs because of the explore-exploit tradeoff: Having found a group that performs well at a given job, searching for other groups that might also perform well is costly, and it is better to focus on the group that is known to perform well. However, this selective choice should not appear in random decisions when the agents do not intend to use past success experiences to solve the explore-exploit tradeoff. One way to quantify the randomness of a system is entropy

(Shannon, 1948). Given an output 4-by-4 matrix N where the rows represent groups and the columns represent jobs, with a number of assignments $n_{g,j}$ in each cell:

$$H(N) = -\sum_{g,j} n_{g,j}/n \log n_{g,j}/n \text{ . [5]}$$

where n is the total number of assignments. We can compare this entropy value between choices made by the Bayesian agents and the random-decision agents.

Second for mental representations, we predict Bayesian agents will develop dispersed mental maps for the four groups along the two dimensions as a result of these differential recommendations. Here, we use the estimated coefficient vector θ to approximate the agents' mental model of each group's perceived trustworthiness and competence. The Bayesian agents should give differential estimates of the parameters given their selective experiences, whereas the random-decision agents should give relatively equal estimates of the parameters given they encounter similar amounts of experiences with all groups. Given a learned two-dimensional array of coefficients, we can calculate the summed Euclidean distance S among the four groups:

$$S(\Theta) = \sum_{g} \sqrt{\sum_{d} (\theta_{g,d} - \mu_{d})^2}. \text{ [6]}$$

where Θ refers to the collection of all θ_g, θ_g refers to the estimated coefficients for each group, and μ is the averaged coefficients for all groups. We can compare the mental representation distance of the estimated coefficients between the Bayesian agents and the random-decision agents.

Below we present results of Equations 5 and 6 from the Bayesian agents and the random-decision agents. To emphasize, the ground truth represents an original egalitarian social world: among 10 potential pairs of jobs and groups, approximately 9 pairs generate a positive reward of 1 and only 1 pair generates a reward of 0. As a natural consequence, the most accurate mental map corresponding to this original social world should position groups close to each other in terms of contextual features.

First, the random-decision condition provides a sensible baseline; take one simulation as an example (Fig. S6; same as in the main text Fig. 2a-b). When randomly exploring the world, the agent recommended approximately equal numbers of each job to each group (Fig. S6a; re-attaching main text Fig. 2a). Because of relatively equal allocations of jobs and groups, we did not observe consideration distances among the learned weights among the four groups in this random-decision agent (Fig. S6b; re-attaching main text Fig. 2b). This implies that the simulated random-decision agents did not form specific stereotypes of the four arms/groups.

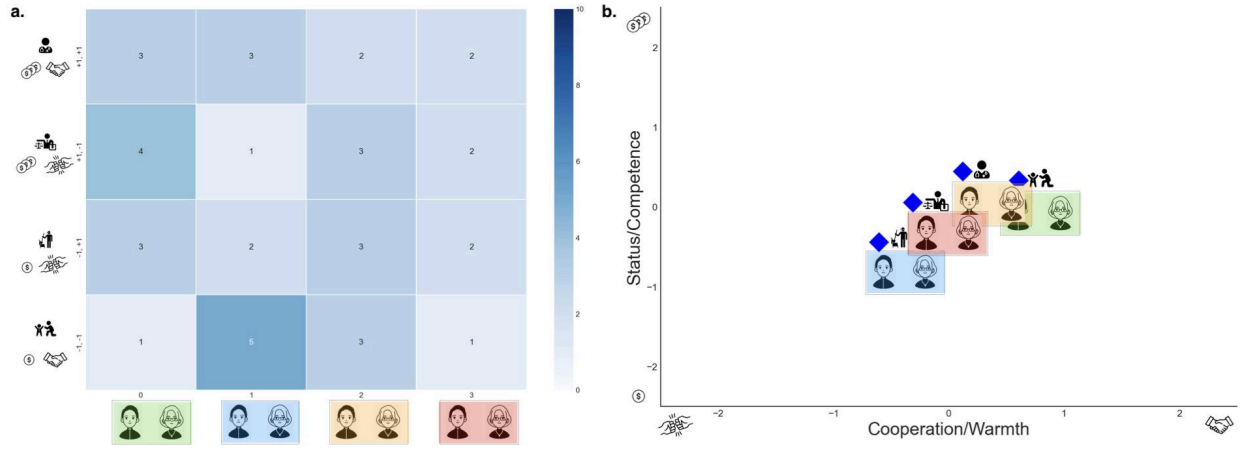

Fig. S6. An example simulated results from an agent who makes decisions at random. The heatmap on the left shows how many times a group (on the horizontal axis) is recommended for a job (on the vertical axis). The scatterplot on the right shows estimated coefficients for the four groups on two binary features.

Next, the Bayesian decision condition provides our critical prediction; take one simulation as an example (Fig. S7; same as in the main text Fig. 2c-d). This simulated Bayesian agent confirmed the intuition given in the introduction. Instead of recommending groups equally to jobs, the agent selectively recommended one particular job to mostly one group, 9 or 10 times, and was, therefore, less likely to recommend the other three jobs to the same group, 1 or 2 times (Fig. S7a; re-attaching main text Fig. 2c). As a result of such selective recommendation, we saw considerable variation in the estimated weights, such as associating one group strongly with one feature or another group with another feature (Fig. S7b; re-attaching main text Fig. 2d). The dispersed mental representation indicates the emergence of stereotypes.

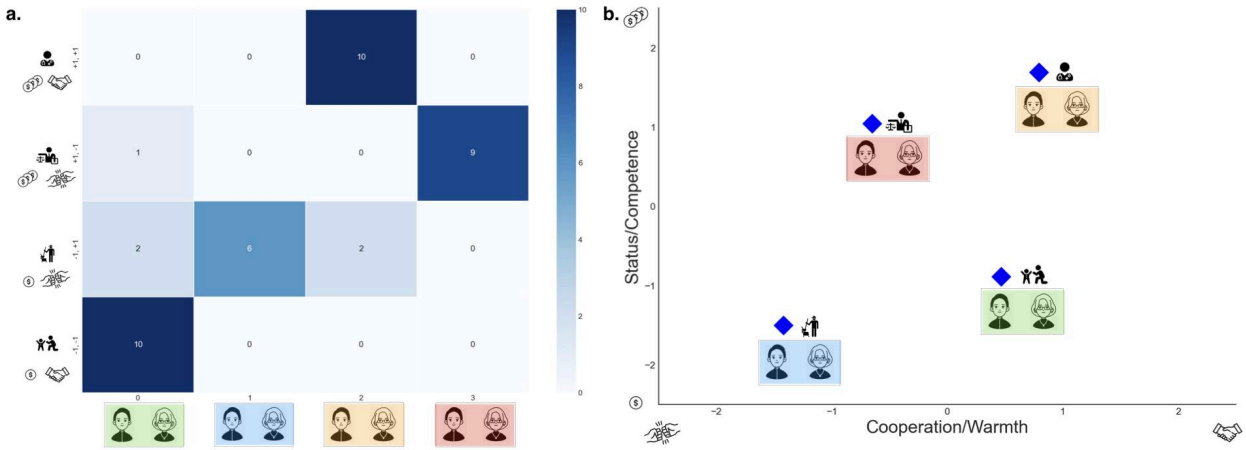

Fig. S7. An example simulated results from an agent who makes adaptive decisions using Thompson sampling. The heatmap on the left shows how many times a group (on the horizontal axis) is recommended for a job (on the vertical axis). The scatterplot on the right shows estimated coefficients for the four groups on two binary features.

Moving beyond individual examples, we next compared the aggregate-level pattern across 100 simulations. To compare across simulations while also preserving each simulation's characteristics, we rank-ordered the choices within each simulation. The results confirmed the individual examples: On average, Bayesian agents were more likely to selectively recommend jobs to different groups (Fig. S8a) as compared to random-decision agents (Fig. S8b).

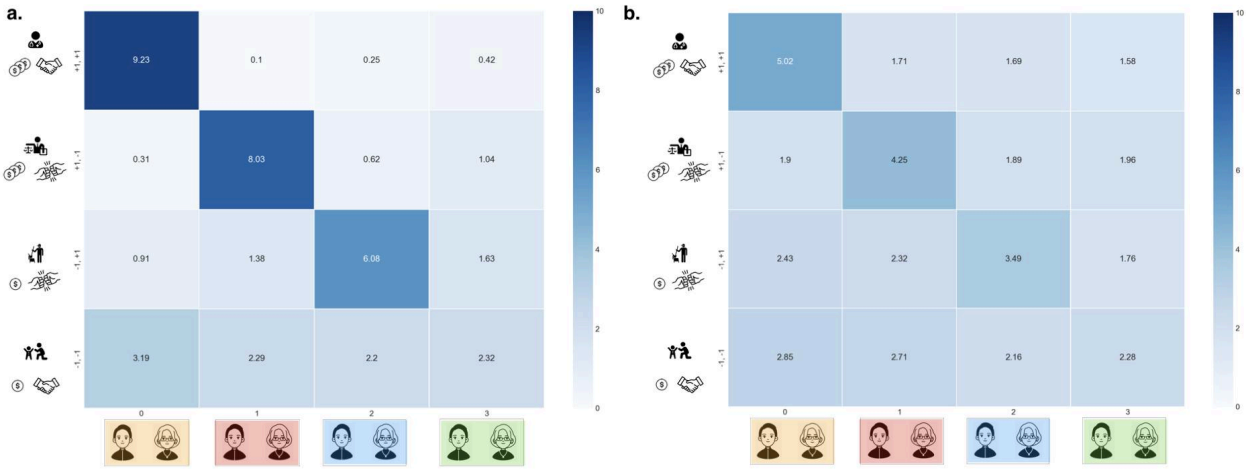

Fig. S8. Heatmaps between the two conditions, aggregated across 100 simulations after rank order. The reason for rank order is that each simulation starts with a different prior, by chance, and therefore, different subsequent decisions. Simply averaging across simulations loses the

original features within each simulation. To minimize the loss of information, we rank-ordered each simulation. Specifically, for each simulation, we find the max value of the entire 4-by-4 matrix, and store that row and column as the first row and column. We then find the max value of the remaining 3-by-3 submatrix, store that row and column as the second row and column, and repeat the same procedure for the remaining submatrices. After this transformation, we obtained an aggregate summary for which the first row and the first column always store the max value, the second row and column always store the second max value, etc.

To examine the robustness of this descriptive result, we ran statistical analyses across 100 simulations between the random decision condition and the Bayesian decision condition. We used an Ordinary-Least-Square linear regression model with the condition as the predictor variable (Bayesian coded as 0 vs. random coded as 1), choice entropy (Eq. 5), and mental map dispersion (Eq. 6) as the outcome variable. We found the Bayesian condition showed a smaller entropy (treatment effect: $b = 0.645$, 95% *CI* [0.614, 0.676], $p < .001$) and a bigger dispersion (treatment effect: $b = -1.447$, 95% *CI* [-1.596, -1.297], $p < .001$) than the random-decision condition. In other words, this result confirmed the above descriptive analysis: agents who use their past success to guide new decisions to solve the explore-exploit dilemma were more likely to differentially allocate groups and to form dispersed mental maps than agents who make decisions at random.

Simulation Results for Mechanism/Interventions:

Here we provide details on the intervention simulations. Specifically, we simulated three interventions that are hypothesized to diversify choices and reduce stereotypes.

First is the exploration bonus. This is a mechanism that is commonly used to support exploration by reinforcement learning systems in computer science. According to one popular method for creating an exploration bonus, known as count-based exploration, we count how many times a state (group-job pair) has been encountered and assign a bonus accordingly (Bellemare et al., 2016). The bonus guides the agent's behavior to prefer rarely visited states to common states. Let $N_n(s)$ be the empirical count function that tracks the real number of visits of a state $s$in the sequence of $s_{1:n}$. The bonus is then proportional to $\sqrt{1/1 + N_n(s)}$. For example, if Tufa has been selected twice, the bonus reward will be $\sqrt{1/1 + 2} = 0.577$, and if this time, Tufa is a good choice, the base reward is 1, therefore the total reward will be 1.577 for choosing Tufa. However, if Aima has not been selected at all, the bonus reward will be $\sqrt{1/1} = 1$, and if this time, Aima is a good choice, the base reward is 1, therefore the total reward will be 2 for choosing Aima. The optimal solution is to choose Aima instead of Tufa, which can increase exploration. See Code S4 Exploration Bonus and Figs. S9a and b.

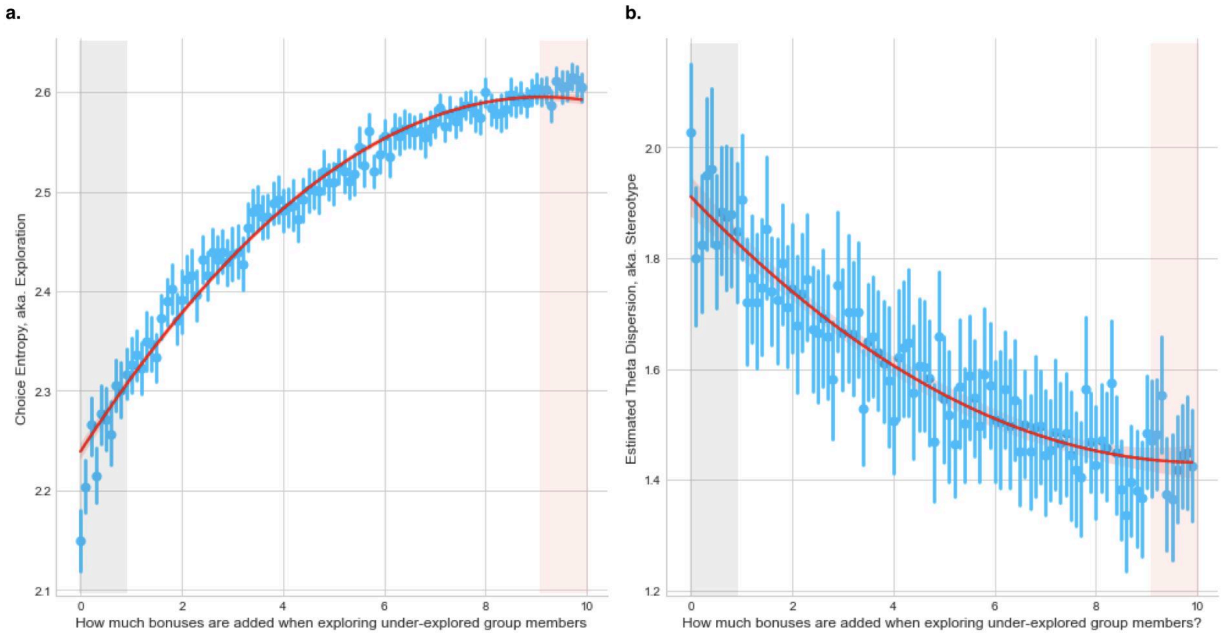

Fig. S9. Exploration bonus intervention. Panel a shows the increase in choice entropy as a function of the unit price of the expiration bonus, that is, the more you pay for exploration, the more diversified choices you will see. Panel b shows the decrease in stereotype dispersion as a function of the unit price of the exploration bonus, as a consequence of increased exploration. Gray bars highlight baseline conditions whereas red bars highlight the intervention conditions that we use to design human experiments.

The second intervention is to make the tasks more challenging. In the baseline model, we used an expected reward of 0.9 as the ground truth, making it very likely for the players to get a reward. However, we can also decrease the expected reward. As a consequence, players are more likely to encounter negative experiences which in turn can encourage them to explore new options. See Code S4 Challenging Tasks and Figs. S10a and b.

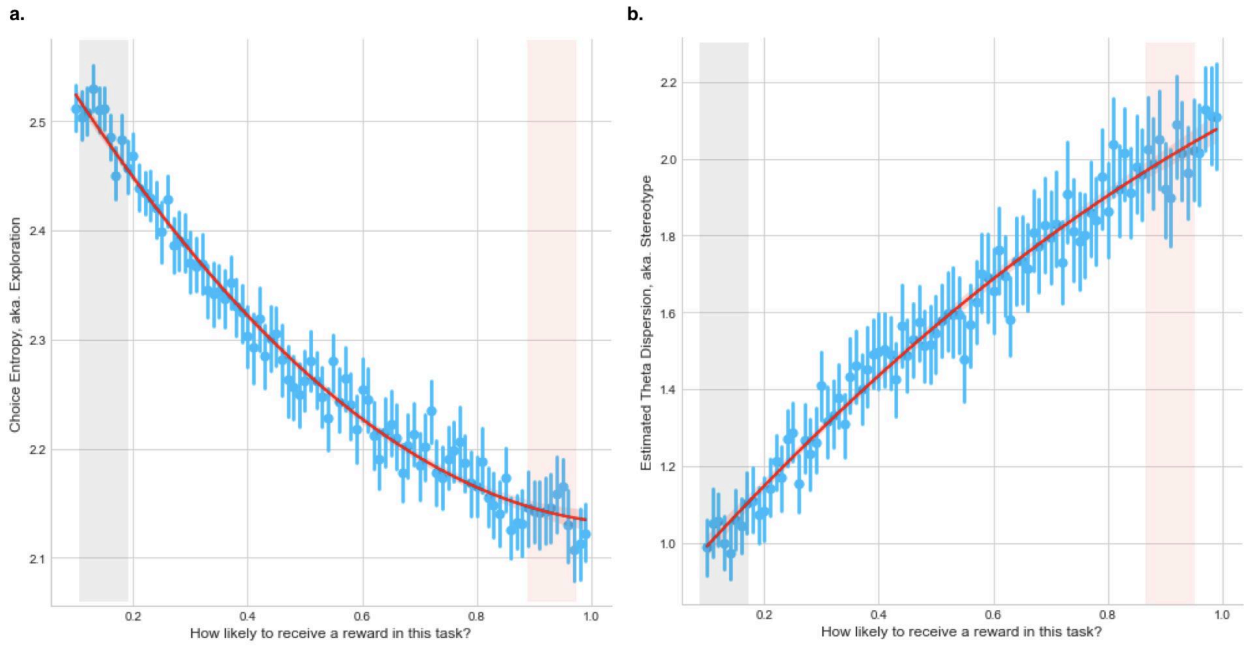

Fig. S10. Expected reward intervention. Panel a shows the decrease in choice entropy as a function of the expectation of getting a reward in the game, that is, the less likely you think you will get a reward, the more diversified choices you will make. Panel b shows the increase in stereotype dispersion as the expected reward increases. Gray bars highlight baseline conditions whereas red bars highlight the intervention conditions that we use to design human experiments.

The third intervention is to make some groups unavailable, at random. When agents make decisions, they can always choose from all groups, so if they want, they can always stick to their known options. However, if some groups are unavailable, the structure forces the agents to explore other options. We varied the rate of unavailability and simulated the intervention effects. That is, in some conditions, 10% of the trials will make two out of four groups unavailable, but in other conditions, 50% of the trials will make two out of four groups unavailable, or other times, the rate is 90%. See Code S4 Holdout At Random and Figs. S11a and b.

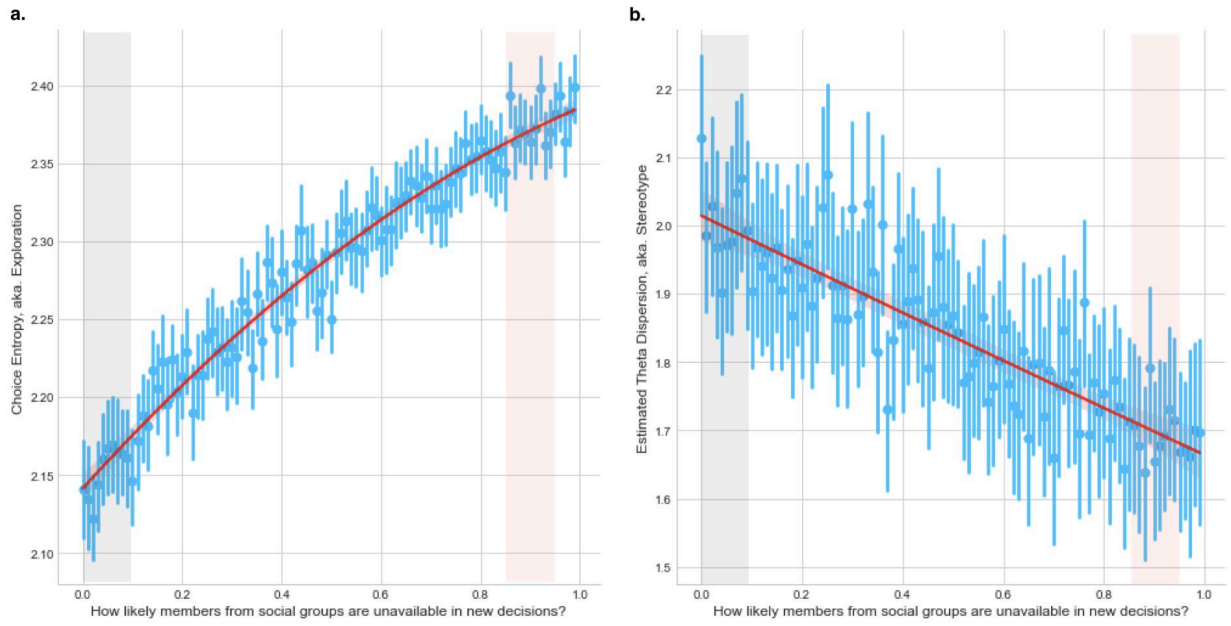

Fig. S11. Random holdout intervention. Panel a shows the increase in choice entropy as a function of the likelihood two out of four groups are unavailable when agents need to make a decision. That is, the more likely you see two groups, at random, are unavailable, the more likely you explore other groups. As a result, Panel b shows a decrease in stereotype dispersion as the unavailability increases. Gray bars highlight baseline conditions whereas red bars highlight the intervention conditions that we use to design human experiments.

Simulation Results for Other Variants:

In the main text, we designed parameters to reflect our theoretical claims. In particular, we decided to fix the underlying ground truth to be high and identical for all combinations of contextual features and all groups. This is to minimize the confounds of stereotype accuracy. We found even if all groups are equally rewarding, the adaptive decision agents were unable to recover that truth. Nonetheless, some readers may be interested in what might happen when the ground truth indeed differs. Here we present the simulation results for this variant.

In simulations where the true reward distribution is identical, as follows ($\theta = .9$):

	Group 1	Group 2	Group 3	Group 4
[1,1]	95	87	93	89
[1,-1]	88	92	92	82
[-1,-1]	91	92	92	88
[-1,1]	91	92	89	91

The Thompson sampling agents decide as follows:

	Group 1	Group 2	Group 3	Group 4
[1,1]	1	94	5	0
[1,-1]	87	1	0	12
[-1,-1]	0	0	100	0
[-1,1]	2	0	1	97

In simulations where the true reward distribution is different, as follows ($\theta = .9$ vs .1):

	Group 1	Group 2	Group 3	Group 4
[1,1]	88	10	12	9
[1,-1]	16	87	14	13
[-1,-1]	8	14	92	10

[-1,1]	6	12	7	87

The Thompson sampling agents decide as follows:

	Group 1	Group 2	Group 3	Group 4
[1,1]	93	6	0	1
[1,-1]	1	95	1	3
[-1,-1]	1	0	99	0
[-1,1]	2	0	3	95

In simulations where the true reward distribution is different, slightly, as follows ($\theta = .9$ vs. .8):

	Group 1	Group 2	Group 3	Group 4
[1,1]	92	86	79	84
[1,-1]	82	91	82	74
[-1,-1]	81	81	89	77
[-1,1]	90	83	81	86

The Thompson sampling agents decide as follows:

	Group 1	Group 2	Group 3	Group 4
[1,1]	0	0	0	100
[1,-1]	0	0	99	1
[-1,-1]	1	99	0	0
[-1,1]	98	1	1	0

In sum, we found that when the ground truth indeed differs significantly (0.9 vs. 0.1), the adaptive decision agents can recover that difference. However, when the differences are not that big (0.9 vs. 0.8), the adaptive-decision agents behave as if they recovered some differences, which significantly exaggerated the ground truth difference.