US Centre Summer Research Grant

Recipient: Chris W. Cai

Project Title: Simulated Contact, Real Empathy:

Assessing the Potential of AI Role-Play Chatbots for Reducing Anti-Immigrant Prejudice in the United States

Project Summary:

Immigration has long been central to America's strength, fuelling innovation and economic growth. Yet deepening polarization and politicization of immigration has led to growing intolerance, manifesting in discriminatory policies, social exclusion, and violence against immigrant communities. In this study, I explore the potential of using AI role-play chatbots as a means of simulating intergroup contact, and combatting anti-immigrant prejudice at scale. Preliminary results indicate that such interventions can be highly effective; on average, spending eight turns role-playing with an AI persona of a famous Chinese American historical figure, reduces participants' anti-Chinese prejudice by three-quarters of a standard deviation.

1 Introduction

Decades of empirical research have shown that immigration is, on balance, a net positive for the United States' social and economic standing in the world (Durrani, 2023; Kim and Knuckey, 2021; Costa et al., 2025). America's image as a "nation of immigrants" draws in millions of people every year, from scientists conducting cutting edge research to farm workers that put food on the nation's tables. These immigrants, in turn contribute roughly $2.1 trillion to the US economy, or roughly 18% of US economic output, despite only making up 14.3% of the country's population (Ruggles et al., 2025).

In spite of the contributions that immigrants make to their adopted country, many Americans view them with hostility and distrust, particularly immigrants who come from racially, ethnically, or religiously distinct backgrounds. Four out of ten Americans believe that immigrants pose a threat to traditional American customs and values (noa, 2023), with Republicans being disproportionately more likely (69%) to hold hostile attitudes than Democrats (17%). These anti-immigrant attitudes further license policies to limit and expel both legal and illegal immigrants, with President Donald Trump's mass deportation campaign being the latest manifestation of this problem(Chishti and Bush-Joseph, 2024; noa, 2026).

In this project, I set out to explore the possibility of using AI role-play chatbots, as a means of reducing anti-immigrant prejudice at scale. Building on intergroup contact theory—which holds that sustained, positive interactions between social groups, generally reduces prejudice between them (Allport, 1979; Paluck et al., 2021)—I hypothesise that role-playing with an AI persona of an outgroup member would reduce users' prejudice towards that outgroup, by simulating one-to-one intergroup contact.

The results from the first of two planned studies indicate that AI role-play chatbots may be highly effective at reducing prejudice; in a pre-registered, online survey experiment (N=1000), I found that participants who spent 8 turns role-playing with an AI persona of a famous Chinese American historical figure, reported a 22.8% reduction in prejudice, compared to 3.1%, for those assigned to have 8-turn conversation with a control chatbot ($d_{\text{Adj}} = 0.75$, 95% CI [0.61, 0.88], $p < .001$).

2 Methods

Following the precedent set by Costello et al. (2024), the study uses a two-cell, treatment-versus-placebo-control randomised controlled experimental design (see Figure 1A). We recruited 1,000 non-Asian American adults via Cloud Research Connect, demographically matched to U.S. census proportions for age, sex, and race/ethnicity, with additional quotas to ensure equal numbers of Democrats, Republicans, and Independents. After obtaining informed consent, participants completed a pre-treatment prejudice questionnaire, and were then randomly assigned, with equal allocation, to either the treatment or placebo control condition. Data collection ran from the 25$^{\text{th}}$ of March to the 2$^{\text{nd}}$ of April, 2026.

In the treatment condition, participants played a role-playing game, where they pretended to be a journalist conducting an important interview with a fa-

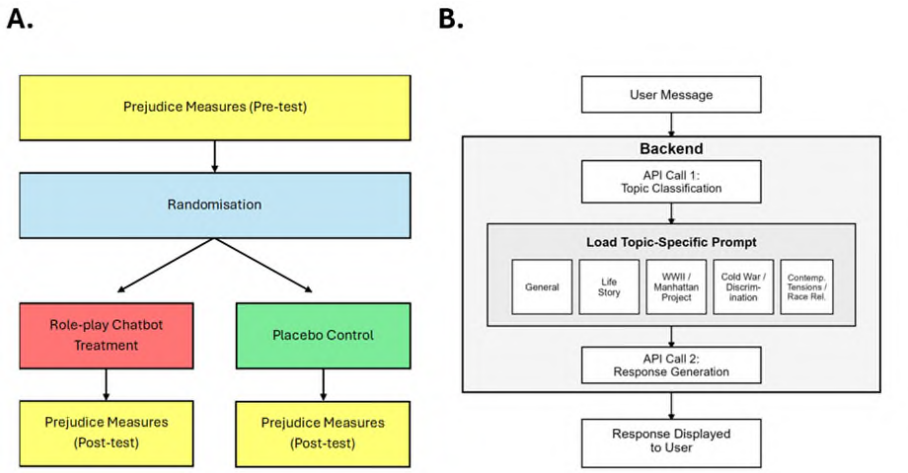


Figure 1: Research design diagram (A) and chatbot design diagram (B)

mous Chinese American historical figure. In the placebo control condition, participants chatted with a Costello et al.-style DebunkBot, which was instructed to argue persuasively in favour of return-to-office initiatives. We chose return-to-office initiatives as the topic for the placebo-control condition, because it is substantive enough to sustain a moderately taxing eight-round conversation, while being entirely unrelated to prejudice or Chinese Americans. We opted to use a placebo control over a pure control to minimise the risk of differential attrition across conditions.

To assess the effects of the intervention on participants' prejudice towards Chinese Americans, Chinese immigrants, and Chinese people in general, we administered a prejudice measurement questionnaire both before and after participants interacted with the chatbot. The questionnaire measured participants' (1) warmth, (2) acceptance, (3) acknowledgement of discrimination, (4) policy support, and (5) blatant dehumanisation, across nine items. To aid in the interpretation of the findings, we constructed a composite prejudice index by taking the mean of all nine prejudice items, each rescaled to a 0–1 range and coded such that higher values indicate greater prejudice. After completing the post-treatment questionnaire, participants in the treatment condition answered additional items measuring their empathy, willingness to befriend, and respect for the person portrayed by the AI chatbot, which we combined into an empathetic rapport index.

We framed the interaction as a *serious game* (Ritterfeld et al., 2009), rather than an open-ended chatbot conversation. Participants were told that they were playing as a journalist, working for a fictional newspaper, and that their task was to interview Professor Chien-Shiung Wu—a Chinese American nuclear physicist who was instrumental to the success of the Manhattan Project, and a founding member of the Committee of 100, a prominent Chinese American civil rights organisation.

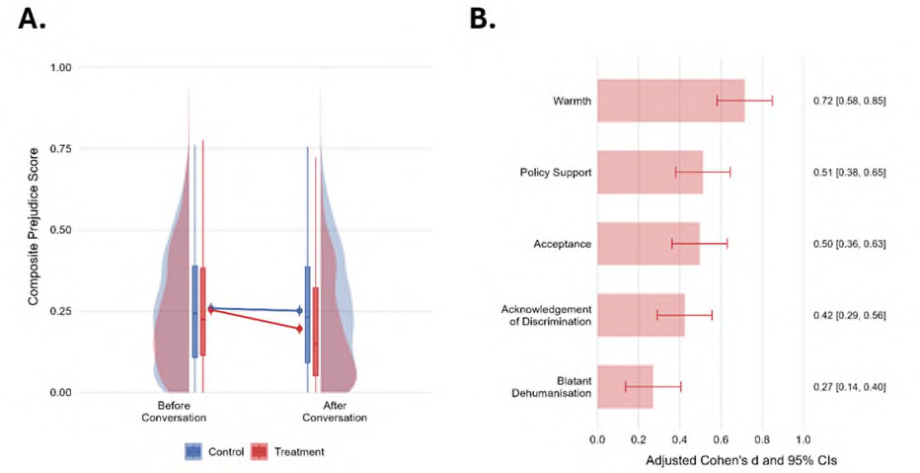


Figure 2: Main effect (A) and variable level effects (B)

3 Results

Main effect. Does role-playing with the Professor Wu chatbot reduce anti-Chinese prejudice? Yes, absolutely (see Figure 2). We see a 22.8% decrease in treatment-group participants' self-reported prejudice scores, compared to 3.1% in the control group. This translates to a 5.1-point reduction on a 0–100 scale (OLS regression with robust standard errors, controlling for composite pre-test scores; $b = -0.051$, 95% CI $[-0.060, -0.042]$, $p < .001$, $d_{\text{Adj}} = -0.75$, partial $\eta^2 = .12$). The effect holds for all five components of prejudice measured by the study—warmth, blatant dehumanisation, acceptance, acknowledgement of discrimination, and support for discriminatory policies—even with Benjamini-Hochberg corrections for false discovery rate. The largest observed effect was for warmth ($d_{\text{Adj}} = -0.72$), and the smallest was for blatant dehumanisation ($d_{\text{Adj}} = -0.27$). This pattern is consistent with the idea that floor effects diluted the overall effect size, as participants are far more likely to be willing to say that they feel cold towards Chinese immigrants, than they would be willing to say that they see Chinese people as being very uncivilised.

The treatment effect is robust to a variety of sensitivity checks for differential attrition (Lee bounds 95% CI $[-0.097, -0.005]$; intention-to-treat analysis with imputation of pre-test scores for missing post-test scores, $b = -0.045$, 95% CI $[-0.054, -0.036]$, $d_{\text{Adj}} = -0.67$), and it would take a strong unmeasured confounder to overturn the result (E-value point estimate = 3.38, $\text{CI}_{\text{Lower}} = 2.90$). Subset analyses suggest the intervention might have actually produced a much larger effect, had we recruited a purely prejudiced sample, as Costello et al. did ($d_{\text{Adj}} = -0.85$ for participants with pre-test scores > 0.10, -0.94 for > 0.25, and -1.26 for > 0.50).

Heterogeneity. The treatment effect was positive at every level of every moderator tested—age, sex, race, ethnicity, education, income, party, ideology, baseline AI attitudes, baseline chatbot attitudes, and baseline prejudice (see Figure 3). Contrary to our expectatios, the effect was *larger* for Republicans,

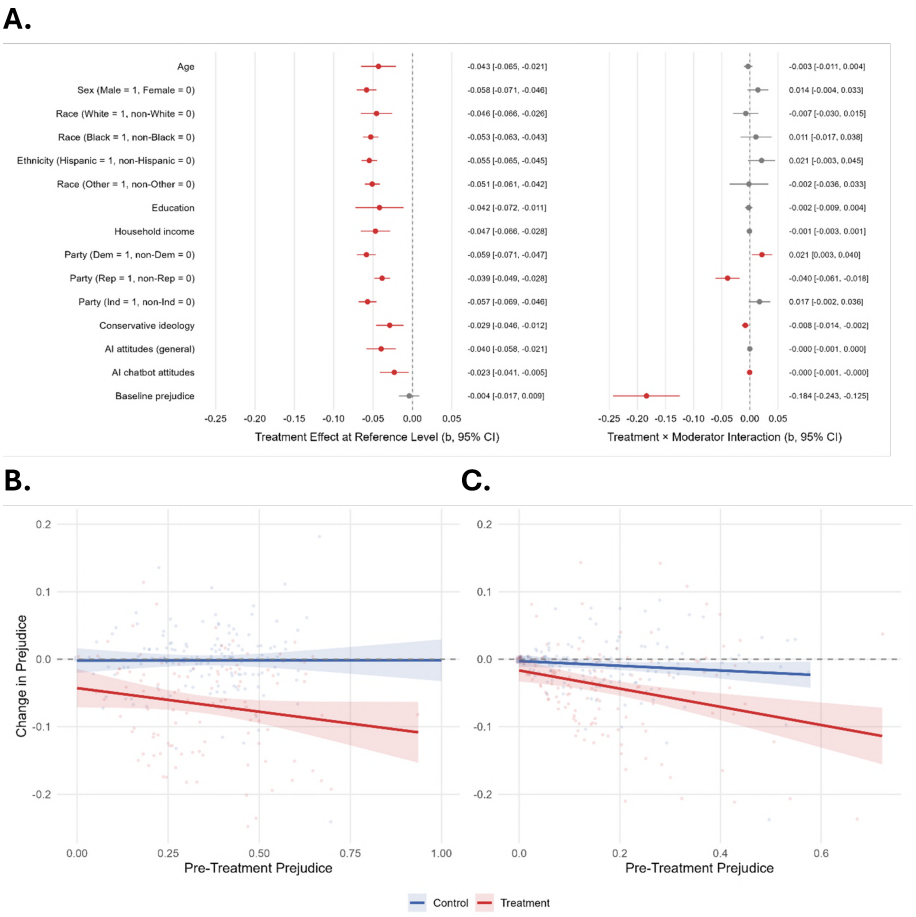


Figure 3: Conditional average treatment effects (A) HTE by party (B) & (C)

conservatives, and participants with higher baseline prejudice, than it was for other participants. We suspect that this is mostly because of floor effects, rather than high baseline prejudices genuinely making people more open to role-play interventions.

Mediation/Was it contact?/Did contact generalise? Treatment participants reported strong empathetic rapport with Professor Wu, and strength-of-rapport predicted magnitude of prejudice reduction within the treatment arm ($b = -0.13$, 95% CI $[-0.18, -0.08]$, $p < .001$). Cross-condition causal mediation **?** also indicates that approximately 53% of the treatment's effect on warmth toward *Chinese people in general* was mediated by its effect on warmth toward the proximal outgroup of Chinese Americans and Chinese immigrants ($M_{\text{ACME}} = -2.85$, 95% CI $[-3.46, -2.23]$, sensitivity $\rho = .69$). These results suggest that the interaction succeeded in simulating outgroup contact, and that participants were able to generalise the traits they perceived in Professor Wu, to the wider outgroup.

4 Discussion

The results of this first study paint a fairly compelling picture as to the potential of AI role-play chatbots for prejudice reduction. An eight-round role-play interaction with the Professor Wu chatbot reduced participants' anti-Chinese prejudice by three-quarters of a standard deviation, after adjusting for baseline prejudice ($d_{\text{Adj}} = 0.75$). This effect is comparable to the largest effects reported in the broader AI persuasion literature, most notably by Costello et al. in their 2024 study on durably reducing conspiracy beliefs using AI DebunkBots (Costello et al., 2024) (3-round conversation, treatment vs. placebo-control, study 2, $d_{\text{Adj}} = 0.79$), and it greatly exceeds the effects reported by the two closest analogues in the field—Crabtree et al.'s forthcoming paper on using AI-powered Deep Canvassing to combat transphobia **?** (6-round conversation, treatment vs. pure-control, $d_{\text{Adj}} = 0.23$), and Rand, Zazai, and Stagnaro's 2025 pre-print on using AI DebunkBots to reduce Islamophobia (Rand et al., 2025) (3-round conversation, treatment vs. placebo-control, $d_{\text{Adj}} = 0.17$).

Consistent with our exploratory hypotheses, we also observe preliminary evidence to suggest that AI role-play can effectively simulate human-to-human contact. Treatment-arm participants reported strong empathetic rapport with Professor Wu, and those who reported the strongest rapport also showed the greatest reductions in anti-Chinese prejudice.

There are, however, several limitations to bear in mind. First, we cannot isolate how much of the treatment effect was due to the affective dimensions of the role-play experience, as opposed to the informational content of the conversation. Second, the intervention was tested on a single type of prejudice, with a sample that for the most part held only weak anti-Chinese animus. While an intervention that only works for reducing weak animus, or to inoculate against prejudice, can be just as valuable as one that works for combating strong prejudices, more research is needed to determine whether the effect extends to other outgroups, stronger prejudices, and populations outside of the United States.

To address these limitations, we plan to conduct a second study within the next six months.

References

, 2023. Are Immigrants a Threat? Most Americans Don't Think So, but Those Receptive to the "Threat" Narrative Are Predictably More Anti-immigrant. Technical Report. Public Religion Research Institute. URL: https://prri.org/wp-content/uploads/2025/05/PRRI-Jan-2023-Immigration-C-1.pdf#page=20.13.

, 2026. Immigration Detention Expansion in Trump's Second Term. Technical Report. American Immigration Council. URL: https://www.americanimmigrationcouncil.org/report/immigration-detention/.

Allport, G.W., 1979. The Nature of Prejudice: 25th Anniversary Edition. Anniversary edition ed., Basic Books, New York. //categories: Intergroup conflict:9, contact hypothesis:-, intergroup contact:-, prejudice reduction:-//.

Chishti, M., Bush-Joseph, K., 2024. Comparing the Biden and Trump Deportation Records | Migration Policy Institute. Technical Report. Migration Policy Institute. URL: https://www.migrationpolicy.org/journal/policy-beat/comparing-biden-and-trump-deportation-records.

Costa, D., Bivens, J., Zipperer, B., Cid-Martinez, I., Perez, D., 2025. Immigrants and the economy. Technical Report. Economic Policy Institute. Washington, D.C.

Costello, T.H., Pennycook, G., Rand, D.G., 2024. Durably reducing conspiracy beliefs through dialogues with AI. Science 385, eadq1814. URL: https://www-science-org.lse.idm.oclc.org/doi/10.1126/science.adq1814, doi:10.1126/science.adq1814. //categories: AI for good:-, AI persuasion:-, AI risks:-, deep canvassing:-, personalisation:-//.

Durrani, S., 2023. What happens when a country bleeds soft power? Conceptualising 'Negative Watch': towards an epistemology for negative and adversarial place branding. Place Branding and Public Diplomacy 19, 456–475. URL: https://doi.org/10.1057/s41254-023-00302-9, doi:10.1057/s41254-023-00302-9.

Kim, M., Knuckey, J.O., 2021. Trump and US soft power. Policy Studies 42, 682–698. URL: https://doi.org/10.1080/01442872.2021.1950668, doi:10.1080/01442872.2021.1950668.

Paluck, E.L., Porat, R., Clark, C.S., Green, D.P., 2021. Prejudice Reduction: Progress and Challenges. Annual Review of Psychology 72, 533–560. URL: https://www.annualreviews.org/content/journals/10.1146/annurev-psych-071620-030619, doi:10.1146/annurev-psych-071620-030619. gSCC: 0000868 2025-09-24T19:22:03.402Z 3.51 //categories: prejudice reduction:9, intergroup contact:-, perspective taking:-//.

Rand, D.G., Zazai, S., Stagnaro, M.N., 2025. Conversations with a large language model improve attitudes toward Muslims and Islam without harming attitudes toward Jews. URL: https://osf.io/preprints/psyarxiv/h2s5t_v1/, doi:10.31234/osf.io/h2s5t_v1.

Ritterfeld, U., Cody, M., Vorderer, P., 2009. Introduction, in: Serious Games. Routledge. GSCC: 0000974 2026-04-30T19:24:40.645Z 1.08 Num Pages: 7.

Ruggles, S., Flood, S., Sobek, M., Backman, D., Cooper, G., Drew, J.A.R., Richards, S., Rodgers, R., Schroeder, J., Williams, K., 2025. IPUMS USA: Version 16.0. URL: https://usa.ipums.org, doi:10.18128/D010.V16.0.