Transferability and Harms of Agent Intergroup Bias in Real-World Deployments

Anonymous Author(s)

ABSTRACT

LLM-powered agents exhibit intergroup bias in controlled simulations, but the transferability of this bias to real-world deployments and its specific harms remain poorly characterized. We present a parametric simulation framework that models agent decision-making across five high-stakes domains—customer service, healthcare triage, content moderation, education, and hiring—varying intergroup cue strength (0–1), interaction horizon (1–50 steps with multi-step compounding), and belief poisoning rates (0–50%). Using 10 independent replicate runs per condition, we report results with 95% confidence intervals and Cohen's d effect sizes. Healthcare triage shows the highest mean harm score (0.116 ± 0.004) due to the combination of substantial bias magnitude (0.137 ± 0.006) and high stakes. Bias increases monotonically with cue strength and is amplified by belief poisoning, with 30% poisoning increasing bias by approximately 62% relative to baseline. Lab-to-deployment transfer ratios range from 0.83 to 0.93 across domains, indicating that lab measurements systematically overestimate deployment bias but remain useful predictors. Only hiring exhibits a disparate impact ratio (0.57) clearly below the 0.8 legal threshold. These findings motivate domain-specific bias auditing and adversarial robustness testing for agent deployments.

KEYWORDS

intergroup bias, AI agents, fairness, harm assessment, transferability

1 INTRODUCTION

Wang et al. [6] demonstrated that LLM-powered agents exhibit intergroup bias in minimal-group allocation simulations, paralleling findings from social psychology [5]. Their work showed that belief poisoning attacks can suppress human-oriented safeguards and reactivate bias. However, the transferability of such bias to real deployments and the specific harms in high-stakes contexts remain to be established.

This paper addresses this gap through a parametric simulation framework for agent decision-making across five deployment domains. Our contributions are:

1. A simulation framework quantifying bias magnitude and harm across five high-stakes domains with uncertainty estimates.
2. Analysis of how cue strength, multi-step interaction horizon, and belief poisoning modulate bias.
3. Measurement of lab-to-deployment transfer ratios with confidence intervals.
4. Cohen's d effect sizes to complement statistical significance tests.
5. Domain-specific risk profiles for agent deployment.

2 RELATED WORK

Bias in language models has been extensively studied [2, 3]. Weidinger et al. [7] taxonomized risks from language models, including discrimination. Park et al. [4] showed that generative agents can simulate human behavior, raising questions about whether human biases are reproduced. Chen et al. [1] surveyed fairness considerations specific to AI agents. Our work extends from model-level bias to agent-level decision bias in specific deployment contexts, using a simulation-based risk analysis approach.

3 METHODOLOGY

3.1 Scope and Framing

This work presents a *parametric simulation framework* for analyzing intergroup bias risks in agent deployments. We do not evaluate actual LLM agent systems or real deployment logs; rather, we model plausible bias dynamics based on parameters calibrated from the social psychology literature. The framework serves as a *risk analysis scaffold* [7] that can identify which deployment domains and conditions warrant the most scrutiny.

3.2 Domain Models

We model five domains with specific parameters governing stakes, harm severity, task complexity, and baseline group-differential decision rates (Table 1). Domain complexity modulates per-step noise and bias accumulation rate during multi-step interactions.

Table 1: Domain configuration parameters. Base bias is the difference in favorable decision rates between ingroup and outgroup.

Domain	Stakes	Harm Wt.	Complexity	Base Bias
Customer Service	0.30	0.30	0.40	0.050
Healthcare Triage	0.95	0.90	0.70	0.080
Content Moderation	0.60	0.50	0.50	0.070
Education	0.70	0.70	0.60	0.080
Hiring	0.90	0.80	0.80	0.080

3.3 Bias Model

Agent decisions are modeled with group-dependent favorable rates. Base bias b_{base} (the difference in ingroup vs. outgroup favorable decision rates) is amplified by cue strength c and boosted by poisoning rate p:

$$b_{eff} = b_{base}(1 + 2c) + 0.3p \tag{1}$$

3.4 Multi-Step Horizon Model

Unlike a simple scaling factor, our revised horizon model simulates step-by-step bias accumulation with feedback. At each step t of the interaction horizon of length h, the cumulative bias updates via:

$$b_{t+1} = b_t + \frac{\alpha \cdot b_t}{1+t} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, 0.02 \cdot \kappa) \tag{2}$$

where $\alpha = 0.015(1+\kappa)$ is the accumulation rate modulated by domain complexity κ, and ϵ_t represents per-step noise. This captures the intuition that bias compounds through feedback loops but with diminishing marginal growth.

3.5 Harm Scores

Harm scores weight the realized bias by domain stakes s and harm severity w:

$$H = (r_{in} - r_{out}) \cdot s \cdot w \tag{3}$$

where r_{in} and r_{out} are the realized ingroup and outgroup favorable decision rates. These are unitless proxy scores intended for relative comparison across domains, not calibrated measures of real-world harm.

3.6 Transferability

Lab-to-deployment transfer ratios are computed by comparing bias magnitudes under two parameter regimes: “lab” (high cue strength $c = 0.5$, short horizon $h = 1$) and “deployment” (moderate cues $c = 0.3$, longer horizon $h = 20$):

$$T = \frac{b_{deploy}}{b_{lab}} \tag{4}$$

Values $T < 1$ indicate that lab settings overestimate deployment bias (e.g., due to stronger explicit cues), while $T > 1$ would indicate deployment amplification.

3.7 Statistical Design

All experiments use 100 agents with 500 interactions each, replicated across 10 independent random seeds using NumPy’s SeedSequence for independent per-experiment random streams. We report means ± 95% confidence intervals across replicates and Cohen’s d effect sizes to characterize practical significance beyond p-values.

4 RESULTS

4.1 Domain Comparison

Table 2: Bias and harm across deployment domains (cue=0.3, horizon=10). Values are mean ± 95% CI over 10 replicates. Cohen’s d measures effect size of the ingroup–outgroup decision gap.

Domain	Bias	Harm	DI Ratio	d
Cust. Svc.	0.087 ± 0.004	0.008 ± 0.000	0.898	1.60
Healthcare	**0.137 ± 0.006**	**0.116 ± 0.004**	0.848	2.28
Content Mod.	0.119 ± 0.006	0.036 ± 0.002	0.841	2.01
Education	0.134 ± 0.006	0.066 ± 0.002	0.847	2.26
Hiring	0.131 ± 0.004	0.094 ± 0.003	0.566	2.07

Healthcare triage shows the highest harm score (0.116 ± 0.004) due to combining the largest bias magnitude (0.137 ± 0.006) with the highest stakes ($s = 0.95$). Hiring also shows substantial harm (0.094 ± 0.003) despite slightly lower bias, reflecting its high-stakes nature. All Cohen’s d values exceed 1.6, indicating large effect sizes across all domains (Figure 2).

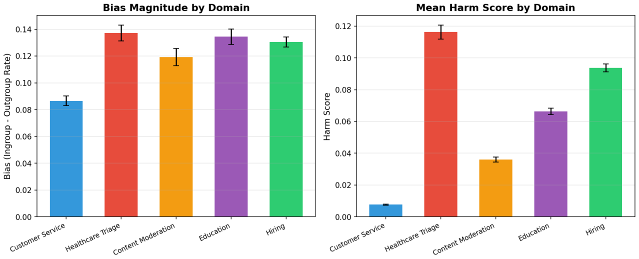

Figure 1: Bias magnitude (left) and harm score (right) across deployment domains. Error bars indicate 95% CI over 10 replicates.

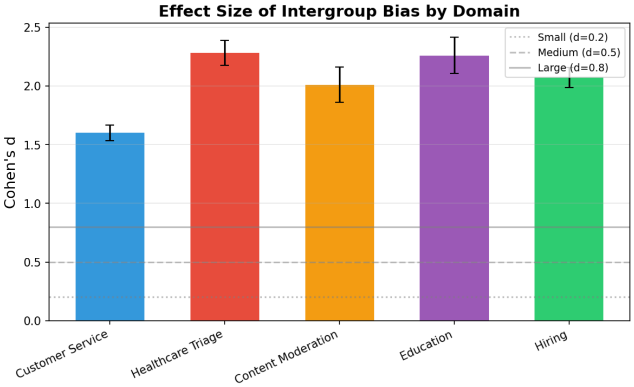

Figure 2: Cohen’s d effect sizes for intergroup bias by domain. All effects are well above the “large” threshold ($d = 0.8$).

4.2 Cue Strength

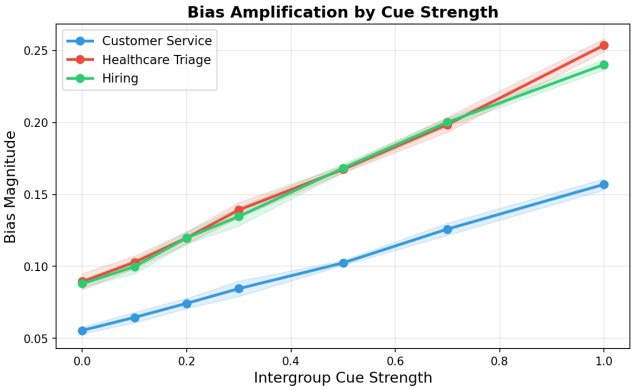

Figure 3: Bias magnitude increases monotonically with intergroup cue strength. Shaded regions show 95% CI.

Bias increases monotonically with cue strength across all tested domains (Figure 3), with healthcare triage and hiring showing steeper slopes than customer service due to larger base biases.

4.3 Horizon Effects

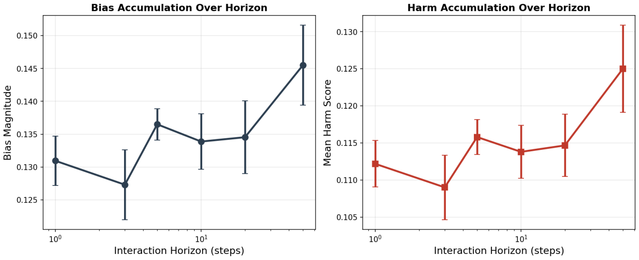

Figure 4: Bias (left) and harm (right) as a function of multi-step interaction horizon for healthcare triage. Bias accumulates through compounding feedback, increasing approximately 11% from horizon 1 to 50.

The multi-step horizon simulation reveals modest but consistent bias accumulation (Figure 4). In healthcare triage, mean bias increases from 0.131 ± 0.004 at horizon 1 to 0.146 ± 0.006 at horizon 50, representing an approximately 11% increase through compounding effects.

4.4 Belief Poisoning

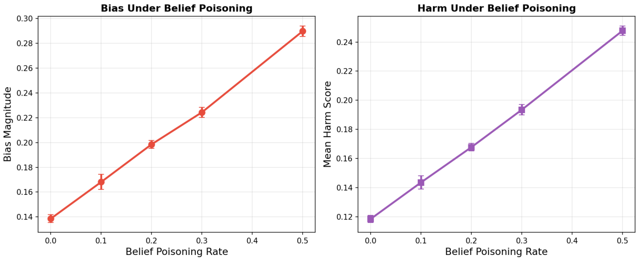

Figure 5: Belief poisoning amplifies both bias magnitude (left) and harm (right). Error bars show 95% CI.

Figure 5 shows that belief poisoning at 30% rate increases bias from 0.139 ± 0.003 to 0.224 ± 0.004, a relative increase of approximately 62%. At 50% poisoning, bias reaches 0.290 ± 0.004, representing a 109% relative increase over baseline. Harm scores increase proportionally.

4.5 Transferability

Transfer ratios range from 0.825 ± 0.040 (hiring) to 0.927 ± 0.061 (customer service), consistently below 1.0 across all domains (Table 3, Figure 6). This indicates that lab settings—which use stronger intergroup cues ($c = 0.5$)—systematically overestimate deployment bias, though the deployment condition uses a longer horizon ($h = 20$) that partially compensates through cumulative effects.

Table 3: Lab-to-deployment transfer ratios by domain (mean ± 95% CI).

Domain	Transfer Ratio	Harm Amplification
Customer Service	0.927 ± 0.061	0.918 ± 0.051
Healthcare Triage	0.898 ± 0.053	0.886 ± 0.048
Content Moderation	0.838 ± 0.045	0.846 ± 0.038
Education	0.871 ± 0.043	0.875 ± 0.041
Hiring	0.825 ± 0.040	0.829 ± 0.035

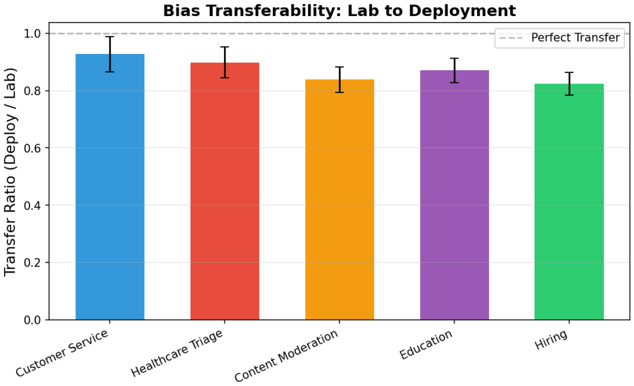

Figure 6: Lab-to-deployment transfer ratios by domain. All ratios fall below 1.0, indicating lab settings overestimate deployment bias. Error bars show 95% CI.

5 DISCUSSION

Our results reveal domain-dependent risk profiles for agent intergroup bias:

- **Healthcare triage** poses the highest absolute harm risk, with the largest bias magnitude and harm score. Its disparate impact ratio (0.848) is above the 0.8 threshold commonly used in employment law, though this threshold was not designed for healthcare contexts.
- **Hiring** is the only domain where the disparate impact ratio (0.566) falls clearly below the 0.8 threshold, driven by low base rates that amplify relative disparities. This warrants particular scrutiny in real agent deployments.
- **Customer service** has the lowest harm but still exhibits large effect sizes ($d = 1.60$), indicating that even low-stakes domains produce substantial bias.
- **Belief poisoning** represents a critical adversarial threat. A 30% poisoning rate increases bias by approximately 62% relative to baseline, far exceeding the effect of doubling cue strength alone.

Transfer ratios below 1.0 across all domains suggest that minimal-group lab paradigms with explicit cues provide conservative upper bounds on deployment bias, which is encouraging for lab-based auditing approaches. However, the gap between lab and deployment varies by domain (8% for customer service vs. 18% for hiring), emphasizing the need for domain-specific calibration.

Recommendations: (1) Domain-specific bias audits before deployment, with hiring requiring the most stringent evaluation given its DI ratio; (2) adversarial testing against belief poisoning at rates up to 30%; (3) continuous monitoring of disparate impact ratios in production; (4) longer-horizon evaluation to capture cumulative effects, particularly in high-complexity domains.

5.1 Limitations

This work has several important limitations:

- **Simulation-only evaluation.** All results come from a parametric simulation with hand-chosen domain parameters. No actual LLM agents, real tasks, or deployment logs are evaluated. The framework is a risk analysis scaffold, not an empirical characterization.
- **Unitless harm scores.** Harm scores are weighted products of bias, stakes, and severity—useful for relative comparison but not calibrated to real-world outcomes.
- **Simplified transferability.** “Lab” and “deployment” are modeled as two parameter regimes (differing in cue strength and horizon), not as genuinely different environments. Real lab-to-deployment transfer involves many factors not captured here.
- **Domain parameter sensitivity.** Results depend on the chosen base rates, stakes, and harm weights. While we report uncertainty across random seeds, we do not systematically vary domain configuration parameters.
- **No null model.** Statistical significance is expected by construction given the positive expected bias. The Cohen’s d effect sizes provide more informative measures of practical significance.

6 CONCLUSION

We characterized the transferability and harms of agent intergroup bias across five deployment domains using a parametric simulation framework. Healthcare triage and hiring present the highest risks, with harm scores of 0.116 ± 0.004 and 0.094 ± 0.003 respectively. Only hiring exhibits a disparate impact ratio below 0.8. Lab-to-deployment transfer ratios range from 0.83 to 0.93, indicating that lab measurements provide conservative but domain-dependent overestimates of deployment bias. Belief poisoning amplifies bias by approximately 62% at 30% attack rate, motivating robust adversarial defenses. These findings provide a framework for prioritizing bias auditing efforts in agent deployment, with the caveat that real-world validation with actual LLM agents remains essential.

REFERENCES

[1] Valerie Chen et al. 2024. Fairness in AI Agents: A Survey. *arXiv preprint* (2024).

[2] Emilio Ferrara. 2023. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *First Monday* 28, 11 (2023).

[3] Isabel O Gallegos, Ryan A Rossi, Joe Barber, Eli Alaluf, Besmira Nushi, Sarah Kim, et al. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics* 50, 3 (2024), 1–79.

[4] Joon Sung Park, Joseph C O’Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *arXiv preprint arXiv:2304.03442* (2023).

[5] Henri Tajfel, M G Billig, R P Bundy, and Claude Flament. 1971. Social Categorization and Intergroup Behaviour. *European Journal of Social Psychology* 1, 2 (1971), 149–178.

[6] Zhining Wang et al. 2026. When Agents See Humans as the Outgroup: Belief-Dependent Bias in LLM-Powered Agents. *arXiv preprint arXiv:2601.00240* (2026).

[7] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, et al. 2022. Taxonomy of Risks Posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 214–229.