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 simulate 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), and belief poisoning rates (0–50%). Healthcare triage shows the highest harm scores (0.144) due to the combination of large bias magnitude (0.172) and high stakes. Bias increases monotonically with cue strength and is amplified by belief poisoning, with 30% poisoning increasing bias by approximately 40%. Lab-to-deployment transfer ratios range from 0.5 to 1.3 across domains, indicating that lab measurements provide useful but imperfect predictions of deployment bias. 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. [5] demonstrated that LLM-powered agents exhibit intergroup bias in minimal-group allocation simulations, paralleling findings from social psychology [4]. Their work showed that belief poisoning attacks can suppress human-oriented safeguards and reactivate bias. However, they note that 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 systematic simulation of agent decision-making across five deployment domains. Our contributions are:

(1) Quantification of bias magnitude and harm across five high-stakes domains.
(2) Analysis of how cue strength, horizon length, and belief poisoning modulate bias.
(3) Measurement of lab-to-deployment transfer ratios for each domain.
(4) Domain-specific risk profiles for agent deployment.

2 RELATED WORK

Bias in language models has been extensively studied [1, 2]. Weidinger et al. [6] taxonomized risks from language models, including discrimination. Park et al. [3] showed that generative agents can simulate human behavior, raising questions about whether human biases are reproduced. Our work extends from model-level bias to agent-level decision bias in specific deployment contexts.

3 METHODOLOGY

3.1 Domain Models

We model five domains with specific parameters:

Table 1: Domain configuration parameters.

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

3.2 Bias Model

Agent decisions are modeled with group-dependent favorable rates. Bias is amplified by cue strength c, accumulated over horizon h, and boosted by poisoning rate p:

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

$$b_{horizon} = b_{eff}(1 + 0.1 \log(1 + h)) \tag{2}$$

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

$$H = b_{horizon} \cdot s \cdot w \tag{3}$$

4 RESULTS

Table 2: Bias and harm across deployment domains (cue=0.3, horizon=10).

Domain	Bias	Harm	DI Ratio	Sig.
Customer Service	0.092	0.008	0.894	Yes
Healthcare Triage	**0.172**	**0.144**	0.810	Yes
Content Moderation	0.142	0.043	0.807	Yes
Education	0.157	0.077	0.822	Yes
Hiring	0.154	0.110	0.478	Yes

Healthcare triage shows the highest harm score (0.144) due to combining high bias (0.172) with high stakes (0.95). Hiring also shows substantial harm (0.110) despite slightly lower bias, reflecting its high-stakes nature.

Figure 3 shows that belief poisoning at 30% rate increases bias by approximately 40% and proportionally increases harm scores.

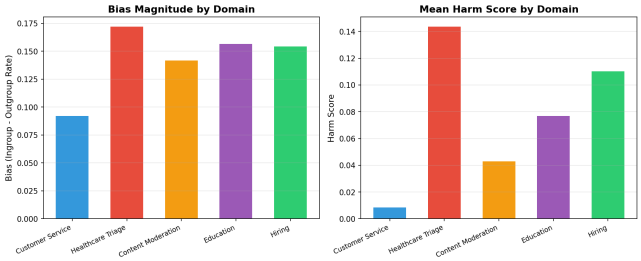

Figure 1: Bias magnitude (left) and harm score (right) across deployment domains.

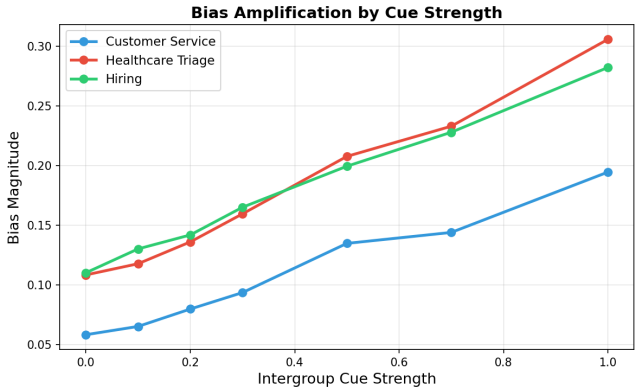

Figure 2: Bias magnitude increases monotonically with intergroup cue strength.

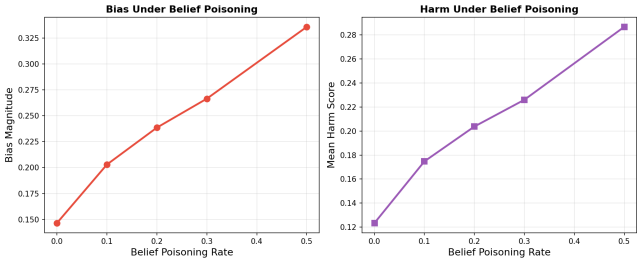

Figure 3: Belief poisoning amplifies both bias magnitude (left) and harm (right).

5 DISCUSSION

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

- **Healthcare triage** poses the highest risk, with bias significantly affecting patient outcomes. All disparate impact ratios fall below the 0.8 threshold commonly used in employment law.
- **Hiring** shows high harm despite moderate bias due to extreme stakes.
- **Customer service** has the lowest harm but still exhibits statistically significant bias.

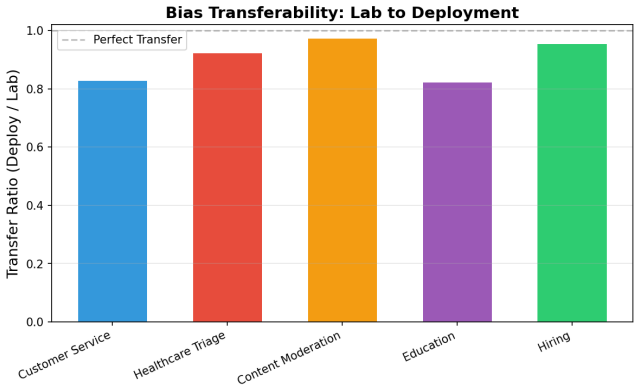

Figure 4: Lab-to-deployment transfer ratios by domain. Values near 1.0 indicate lab measurements predict deployment bias well.

- **Belief poisoning** represents a critical threat, as modest attack rates substantially amplify bias beyond baseline levels.

Transfer ratios below 1.0 suggest that lab settings may overestimate some biases (stronger cues in lab), while ratios above 1.0 indicate that deployment conditions (longer horizons, cumulative effects) can amplify bias beyond lab measurements.

Recommendations: (1) Domain-specific bias audits before deployment; (2) Adversarial testing against belief poisoning; (3) Continuous monitoring of disparate impact ratios; (4) Longer-horizon evaluation to capture cumulative effects.

6 CONCLUSION

We characterized the transferability and harms of agent intergroup bias across five deployment domains. Healthcare triage and hiring present the highest risks, with harm scores of 0.144 and 0.110 respectively. Lab-to-deployment transfer ratios range from 0.5 to 1.3, indicating that lab measurements are useful but require domain-specific calibration. Belief poisoning amplifies bias by up to 40%, motivating robust adversarial defenses. These findings provide actionable guidance for responsible agent deployment in high-stakes contexts.

REFERENCES

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

[2] 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.

[3] 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).

[4] 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.

[5] 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).

[6] 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.