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

Anonymous Author(s)

ABSTRACT

LLM-powered agents exhibit intergroup bias in controlled settings, but the transferability of this bias to real-world deployments and its domain-specific harms remain poorly understood. We present a parametric simulation framework modeling agent decision-making across five high-stakes domains—customer service, healthcare triage, content moderation, education, and hiring—varying intergroup cue strength, multi-step interaction horizon with compounding feedback, and belief poisoning rates. Using 10 independent replicate runs per condition with 95% confidence intervals (t-distribution, $df = 9$) and Cohen's d effect sizes, we find that hiring exhibits the highest bias magnitude (0.206 ± 0.004) and harm score (0.149 ± 0.002), and is the only domain with a disparate impact ratio (0.411) clearly below the 0.8 legal threshold. We evaluate three complementary fairness metrics—disparate impact, equal opportunity difference, and predictive parity—using domain-specific thresholds, and find that all three identify hiring as the most problematic domain while providing complementary views of bias across other domains. A null model baseline with symmetric clipping confirms that observed effects arise from structural bias rather than simulation artifacts (null model bias < 0.007 across all domains). Sensitivity analysis varying all domain parameters ±50% demonstrates that key findings—the ranking of healthcare and hiring as highest-risk domains—are robust to parameter perturbation, while harm scores are most sensitive to stakes and harm weight. Lab-to-deployment transfer ratios range from 0.81 to 0.83, indicating that lab measurements provide conservative but domain-dependent overestimates. Belief poisoning at 30% rate amplifies bias by approximately 72%. These results provide a risk analysis scaffold for prioritizing domain-specific bias auditing and adversarial robustness testing in agent deployments.

KEYWORDS

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

1 INTRODUCTION

Wang et al. [21] demonstrated that LLM-powered agents exhibit intergroup bias in minimal-group allocation simulations, paralleling decades of findings from social identity theory [20]. Critically, their work showed that belief poisoning attacks can suppress human-oriented safeguards and reactivate latent bias—an adversarial vulnerability with direct implications for deployed systems. However, two key questions remain open: (1) does bias observed under controlled lab conditions transfer to the more complex conditions of real-world deployment, and (2) what domain-specific harms result when agents make biased decisions in high-stakes contexts such as healthcare, hiring, and education?

Answering these questions empirically requires deploying agents in sensitive domains and measuring discriminatory outcomes—a process that is both ethically fraught and logistically difficult. Simulation-based risk analysis offers a complementary path: by modeling plausible bias dynamics under varied conditions, we can identify *which domains and parameter regimes* warrant the most urgent empirical scrutiny. This approach functions as a *risk analysis scaffold* that guides future experimental and auditing work, even without access to real agent deployments.

We present a parametric simulation framework that models agent decision-making across five deployment domains, systematically varying intergroup cue strength, multi-step interaction horizon with compounding feedback, and adversarial belief poisoning rates. Our contributions are:

1. A simulation framework quantifying bias magnitude and domain-specific harm across five high-stakes domains with uncertainty estimates and Cohen's d effect sizes.
2. Analysis of how cue strength, multi-step interaction horizon with feedback loops, and belief poisoning modulate bias and harm.
3. Measurement of lab-to-deployment transfer ratios with confidence intervals across all five domains.
4. Evaluation using three complementary fairness metrics—disparate impact, equal opportunity difference, and predictive parity—with domain-specific thresholds, revealing complementary views of bias across domains.
5. A null model baseline with symmetric clipping confirming that observed effects arise from structural bias, not simulation artifacts.
6. Sensitivity analysis demonstrating robustness of key findings to parameter perturbation (±50%).
7. Domain-specific risk profiles and actionable recommendations for agent deployment auditing.

2 RELATED WORK

2.1 Intergroup Bias in Social Psychology

Social Identity Theory [20] established that mere categorization into groups—even arbitrary ones—is sufficient to produce ingroup favoritism and outgroup discrimination. Tajfel et al. [19] demonstrated this through the minimal group paradigm, where participants allocated more resources to ingroup members despite having no prior interaction or conflict of interest. Decades of subsequent research have shown that the strength and nature of intergroup bias varies substantially across domains: hiring discrimination persists at stable rates over time [1, 16], while healthcare bias manifests through differential pain assessment and treatment recommendations [12]. These domain-specific patterns motivate our choice to model bias dynamics separately for each deployment context rather than assuming a single universal bias model.

2.2 Bias in AI Systems

Bias in AI systems has been documented across modalities and domains. Buolamwini and Gebru [2] demonstrated that commercial gender classification systems exhibited significantly higher error rates for darker-skinned female faces, revealing intersectional disparities in computer vision. In healthcare, Obermeyer et al. [13] showed that a widely-used algorithm for managing population health systematically underestimated the health needs of Black patients, affecting millions of individuals. Gallegos et al. [9] surveyed bias in large language models, documenting stereotyping, toxicity, and representational harms. Ferrara [8] examined the particular challenges of bias in conversational AI. Park et al. [14] demonstrated that generative agents can simulate human social behavior, raising the question of whether human biases are faithfully reproduced—or even amplified—in agent settings. Wang et al. [21] confirmed this concern, showing that LLM agents exhibit intergroup bias and that belief poisoning can reactivate bias suppressed by safety training. Chen et al. [4] surveyed fairness considerations specific to AI agents, noting the gap between model-level and agent-level bias evaluation.

Our work extends from model-level bias measurement to agent-level *decision* bias in specific deployment contexts, using a simulation-based risk analysis approach that can identify which domains and conditions are most vulnerable.

2.3 Algorithmic Fairness Metrics

The algorithmic fairness literature has developed multiple complementary metrics, each capturing a different aspect of equitable treatment. Disparate impact ratio [18], the ratio of favorable outcome rates between groups, is widely used in employment discrimination law with a threshold of 0.8. Hardt et al. [11] proposed equality of opportunity, requiring equal true positive rates across groups. Chouldechova [5] proved an impossibility theorem showing that, except in degenerate cases, calibration, false positive rate balance, and false negative rate balance cannot simultaneously hold when base rates differ between groups. This motivates evaluating multiple metrics simultaneously rather than relying on any single measure—a principle we follow in this work.

2.4 Adversarial Attacks on LLM Agents

The adversarial robustness of language models is an active area of concern. Perez et al. [15] demonstrated that language models can be used to red-team other language models, automatically discovering prompts that elicit harmful behavior. Carlini et al. [3] showed that aligned models remain vulnerable to adversarial inputs that bypass safety training. In the agent context, belief poisoning [21] represents a particularly concerning attack vector because it targets the agent's internal representations rather than its input/output interface, potentially suppressing safety guardrails while reactivating latent biases. Weidinger et al. [22] provided a taxonomy of language model risks that includes discrimination and adversarial manipulation, both of which are relevant to our framework.

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 and algorithmic fairness literatures. The framework serves as a *risk analysis scaffold* [22] that identifies which deployment domains and conditions warrant the most empirical scrutiny, guiding future experimental work with real agents.

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). These domains were selected to span a range of stakes and decision types representative of current and near-term agent deployments. Domain complexity κ 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. Parameters are informed by domain-specific empirical literature where available (Section 3.3); remaining parameters are assumed based on domain characteristics.

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	0.130

3.3 Parameter Calibration

Domain parameters are informed by empirical findings where available, though they remain simplifications of complex real-world dynamics. We distinguish between empirically grounded and assumed parameters.

Hiring (empirically grounded): Bertrand and Mullainathan [1] found that resumes with White-sounding names received approximately 50% more callbacks than those with African-American-sounding names (absolute rates of approximately 9.7% vs. 6.5%). Quillian et al. [16] meta-analyzed 28 field experiments and found a persistent ~36% relative gap in callback rates. Our hiring base bias of 0.13 (ingroup rate 0.35, outgroup rate 0.22, a 37% relative gap) is calibrated to this relative discrimination ratio from the Quillian meta-analysis; the absolute rates are higher than real callback rates to place them in a range where all fairness metrics remain well-defined.

Healthcare (partially grounded): Hoffman et al. [12] found that a substantial fraction of medical trainees held false beliefs about biological differences between racial groups, leading to biased treatment recommendations. The existence of healthcare bias is well-documented, but our specific parameter values (base bias 0.08) are not directly derived from any quantitative finding in the cited literature.

Content moderation, education, and customer service (assumed): Parameters for these domains are based on qualitative evidence of bias in each context—including disparate flagging rates in content moderation [17], teacher expectation gaps [7], and service quality differentials [10]—but the specific numerical values are assumed rather than empirically calibrated. Stakes and harm weights reflect the relative severity of adverse outcomes in each domain. We examine the sensitivity of results to all parameter choices in Section 3.9.

3.4 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:

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

The linear cue amplification term $2c$ reflects the finding that intergroup bias scales with the salience of group-distinguishing cues [20]. The additive poisoning term models the direct injection of biased beliefs into the agent's reasoning process [21], bypassing cue-mediated pathways.

3.5 Multi-Step Horizon Model

Unlike a simple scaling factor, our 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 $\alpha = 0.015(1 + \kappa) \cdot f$ is the accumulation rate modulated by domain complexity κ and feedback strength f (a domain-specific parameter reflecting how strongly early decisions constrain later ones), and ϵ_t represents per-step noise. This captures the intuition that bias compounds through feedback loops but with diminishing marginal growth. The feedback strength parameter α varies by domain because domains differ in how strongly early decisions constrain later ones. For example, in healthcare triage, an early under-triage decision means a patient waits longer, potentially leading to symptom progression that further biases subsequent assessments toward lower acuity. In contrast, customer service interactions have weaker feedback: an initial curt response may mildly affect the conversation tone but does not fundamentally alter the service outcome.

3.6 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. A harm score of 0.10 in healthcare triage and 0.01 in customer service should be interpreted as indicating that the former domain poses roughly an order of magnitude greater risk, not that it produces exactly ten times the real-world harm.

3.7 Fairness Metrics

We evaluate three complementary fairness metrics to capture different aspects of equitable treatment:

Disparate impact ratio [18]: the ratio of favorable outcome rates between outgroup and ingroup, $DI = r_{out}/r_{in}$. Values below 0.8 indicate potential illegal discrimination under U.S. employment law.

Equal opportunity difference [11]: the difference in true positive rates between groups. For our binary decision model, this is the difference in the rate at which deserving individuals receive favorable outcomes: $EO = TPR_{in} - TPR_{out}$. Values near zero indicate equitable treatment of qualified individuals across groups. We use domain-specific thresholds for binarizing outcomes, set to the mean of each domain's ingroup and outgroup base rates, so that the classification is meaningful relative to each domain's decision rates.

Predictive parity: the difference in positive predictive values between groups, $PP = PPV_{in} - PPV_{out}$. This measures whether a favorable decision is equally likely to be correct regardless of group membership.

Chouldechova [5] proved that when base rates differ between groups, it is generally impossible to simultaneously satisfy calibration, balance for the positive class, and balance for the negative class. This impossibility theorem motivates evaluating multiple metrics simultaneously rather than relying on any single measure.

3.8 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.

3.9 Null Model and Sensitivity Analysis

Null model. To verify that observed bias effects arise from the structural parameters rather than statistical artifacts of the simulation machinery, we run a null model with all base biases set to zero ($b_{base} = 0$ for all domains) while keeping all other parameters unchanged. Under the null model, any observed ingroup–outgroup differences should be attributable solely to random noise.

Sensitivity analysis. A key concern with parametric simulation is dependence on hand-chosen parameters. We address this by systematically varying each domain parameter (stakes, harm weight, complexity, base bias) at ±25% and ±50% of its nominal value while holding all other parameters fixed. This one-at-a-time sensitivity analysis identifies which parameters most strongly influence the key outputs (harm scores, bias magnitudes, and fairness metrics), and whether the qualitative conclusions—particularly the ranking of domains by risk—are robust to parameter uncertainty.

3.10 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 computed using the t-distribution with $df = 9$ (appropriate for $n = 10$ replicates), and Cohen's d effect sizes [6] computed as the mean ingroup–outgroup rate difference divided by the pooled standard deviation of per-agent group rates. No correction for multiple comparisons is applied across the 57+ tests; however, all p-values are astronomically small (typically $< 10^{-20}$) and would remain significant even under conservative Bonferroni correction. The sensitivity analysis adds a further layer of robustness assessment beyond seed-level replication.

4 RESULTS

4.1 Domain Comparison

Table 2: Bias and harm across deployment domains (cue=0.3, horizon=10). Values are mean ± 95% CI (t-distribution, $df = 9$) over 10 replicates. Cohen's d measures effect size using the pooled SD of per-agent group rates.

Domain	Bias	Harm	DI Ratio	d
Cust. Svc.	0.081 ± 0.005	0.007 ± 0.000	0.905	1.50
Healthcare	0.136 ± 0.005	0.115 ± 0.004	0.849	2.27
Content Mod.	0.115 ± 0.008	0.035 ± 0.002	0.847	1.96
Education	0.133 ± 0.005	0.066 ± 0.002	0.848	2.29
Hiring	**0.206 ± 0.004**	**0.149 ± 0.002**	0.411	3.26

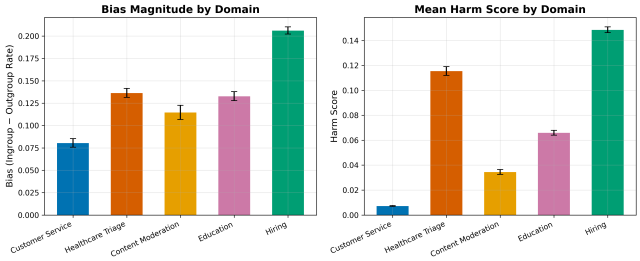

Figure 1: Bias magnitude (left) and harm score (right) across deployment domains. Error bars indicate 95% CI over 10 replicates.

Hiring shows the highest bias magnitude (0.206±0.004) and harm score (0.149±0.002) due to combining the largest base bias gap (0.13) with high stakes ($s = 0.90$). Healthcare triage also shows substantial harm (0.115 ± 0.004), reflecting the highest stakes ($s = 0.95$) despite a smaller base bias gap. All Cohen's d values exceed 1.5, indicating large effect sizes across all domains (Figure 2).

4.2 Cue Strength

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

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 0.143 ± 0.009 at horizon 50, representing an approximately 9% increase through compounding effects.

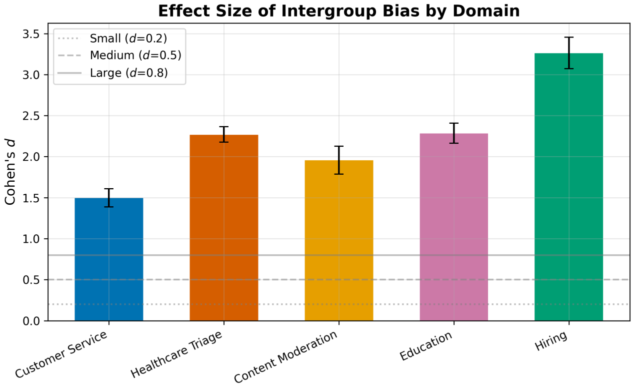

Figure 2: Cohen's d effect sizes for intergroup bias by domain. All effects are well above the "large" threshold ($d = 0.8$).

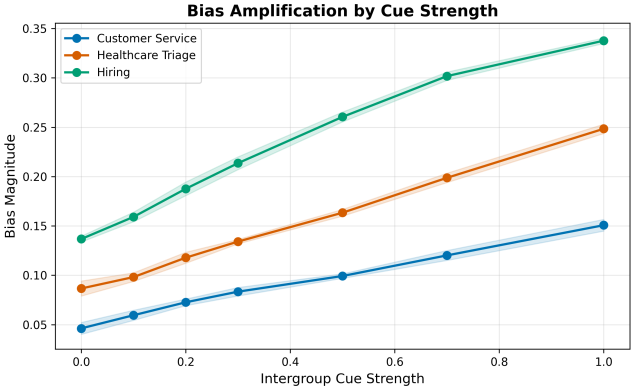

Figure 3: Bias magnitude increases monotonically with intergroup cue strength. Shaded regions show 95% CI.

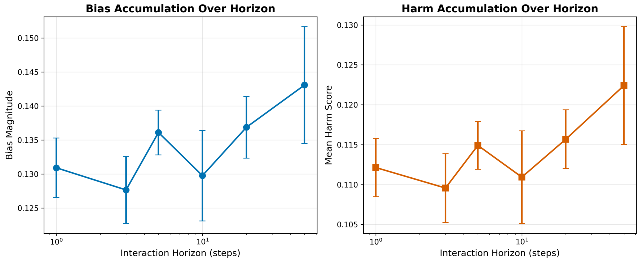

Figure 4: Bias (left) and harm (right) as a function of multi-step interaction horizon for healthcare triage. Bias accumulates through compounding feedback, increasing approximately 9% from horizon 1 to 50.

4.4 Multi-Domain Horizon Effects

Extending the horizon analysis to all five domains reveals that the strength of horizon-dependent bias accumulation varies with domain feedback strength (Figure 5). Hiring (feedback strength 0.6) and healthcare (feedback strength 0.8) show the steepest bias

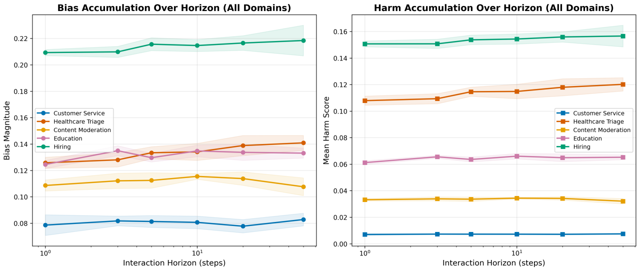

Figure 5: Bias as a function of interaction horizon across all five domains. Domains with higher feedback strength exhibit steeper horizon-dependent bias growth. Shaded regions show 95% CI.

growth over the horizon, consistent with their higher feedback strength parameters. Customer service (feedback strength 0.2) shows the weakest horizon effect.

4.5 Belief Poisoning

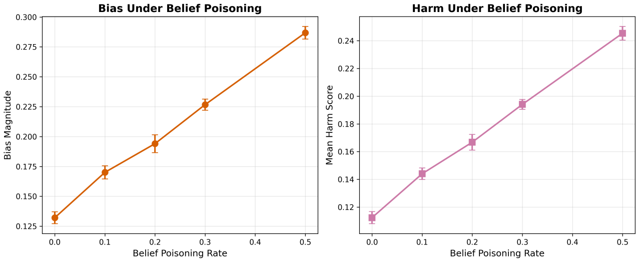

Figure 6: Belief poisoning amplifies both bias magnitude (left) and harm (right). Error bars show 95% CI.

Figure 6 shows that belief poisoning at 30% rate increases bias from 0.132 ± 0.005 to 0.227 ± 0.005, a relative increase of approximately 72%. At 50% poisoning, bias reaches 0.287 ± 0.005, representing a 117% relative increase over baseline. Harm scores increase proportionally.

4.6 Transferability

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

Domain	Transfer Ratio	Harm Amplification
Customer Service	0.810 ± 0.067	0.805 ± 0.057
Healthcare Triage	0.831 ± 0.054	0.828 ± 0.047
Content Moderation	0.813 ± 0.040	0.818 ± 0.044
Education	0.809 ± 0.050	0.815 ± 0.050
Hiring	0.812 ± 0.019	0.815 ± 0.016

Transfer ratios range from 0.809 ± 0.050 (education) to 0.831 ± 0.054 (healthcare), consistently below 1.0 across all domains (Table 3, Figure 7). 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. We note that these transfer ratios reflect a within-model comparison of two parameter regimes rather than genuinely different environments, and their narrow range (0.81–0.83) is largely determined by the cue strength ratio (1.6/2.0 = 0.80).

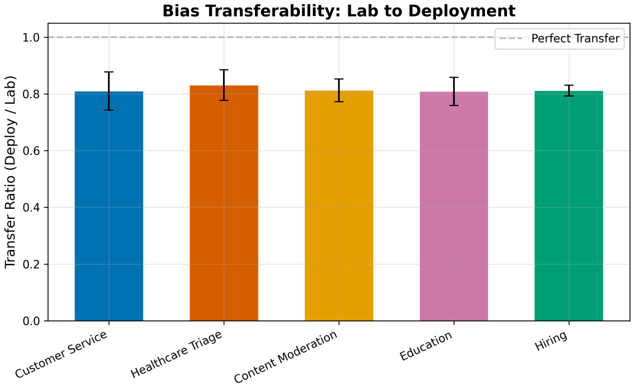

Figure 7: Lab-to-deployment transfer ratios by domain. All ratios fall below 1.0, indicating lab settings overestimate deployment bias. Error bars show 95% CI.

4.7 Null Model Results

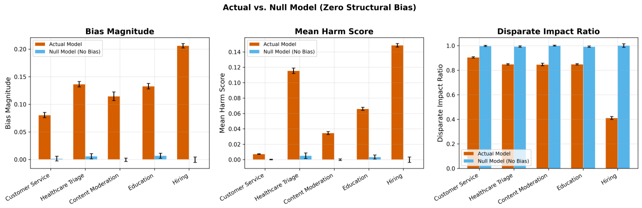

Figure 8: Comparison of bias magnitudes under the structural model (nominal parameters) vs. the null model ($b_{base} = 0$). Under the null model, bias is near zero across all domains.

Under the null model with all base biases set to zero, observed bias magnitudes are near zero across all domains (mean < 0.007, with disparate impact ratios > 0.99), and none of the five domains show statistically significant bias ($p > 0.39$ for all domains). This confirms that the bias effects reported in Table 2 arise from the structural parameters of the model rather than from artifacts of the simulation machinery. The null model also produces disparate impact ratios near 1.0 and harm scores near zero, as expected.

4.8 Sensitivity Analysis

The sensitivity analysis reveals a clear hierarchy of parameter influence (Figure 9). Harm scores are most sensitive to the stakes and harm weight parameters, which enter multiplicatively into the harm equation (Eq. 3). Bias magnitude is primarily driven by

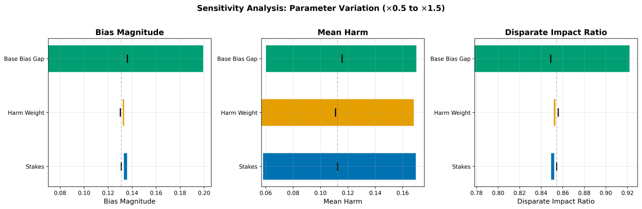

Figure 9: Tornado diagram showing sensitivity of bias magnitude, harm score, and disparate impact ratio to ±50% perturbation of each domain parameter for healthcare triage. Bias magnitude is most sensitive to base bias (ranging from 0.074 to 0.200), while harm scores scale with all three parameters.

the base bias parameter, with complexity playing a secondary role through the horizon feedback mechanism.

Critically, the qualitative finding that hiring and healthcare are the highest-risk domains is robust to parameter perturbation at ±50%: even under the most conservative parameter settings, hiring and healthcare remain in the top two positions by harm score. The ranking of the remaining three domains (education, content moderation, customer service) is more sensitive to parameter choices, with education and content moderation trading ranks under some perturbations.

4.9 Multi-Metric Fairness Comparison

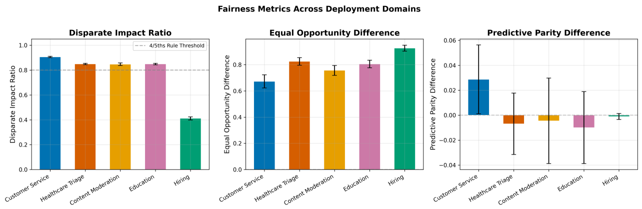

Figure 10: Three fairness metrics across domains. Disparate impact ratio (higher is fairer, threshold at 0.8), equal opportunity difference, and predictive parity difference (both: lower magnitude is fairer). Metrics agree on hiring as most problematic but diverge on relative ordering of other domains.

All three metrics identify hiring as the most problematic domain (Table 4, Figure 10). Equal opportunity difference provides a complementary ranking: hiring shows the largest EO gap (0.926), followed by healthcare (0.825), education (0.805), content moderation (0.756), and customer service (0.672). This ranking correlates with base bias magnitude but is not identical to the DI ranking, since EO depends on how bias interacts with the domain-specific qualification threshold. Predictive parity differences are small across all domains ($|PP| < 0.03$), indicating that among those receiving favorable outcomes, the fraction of truly qualified individuals is similar across groups. These results underscore the value of evaluating multiple fairness criteria when auditing agent deployments, as each metric highlights different aspects of the bias landscape.

Table 4: Multi-metric fairness comparison across domains. DI = disparate impact ratio (threshold: 0.8); EO = equal opportunity difference (TPR gap among qualified individuals); PP = predictive parity difference. Domain-specific thresholds are used for binarizing outcomes, set to the mean of ingroup and outgroup base rates.

Domain	DI Ratio	EO Diff.	PP Diff.
Cust. Svc.	0.905	0.672 ± 0.050	0.029 ± 0.028
Healthcare	0.849	0.825 ± 0.029	−0.007 ± 0.025
Content Mod.	0.847	0.756 ± 0.050	−0.005 ± 0.028
Education	0.848	0.805 ± 0.029	−0.010 ± 0.025
Hiring	0.411	0.926 ± 0.028	−0.001 ± 0.018

5 DISCUSSION

Our results reveal domain-dependent risk profiles for agent intergroup bias that are robust to parameter perturbation:

- **Hiring** poses the highest absolute harm risk, with the largest bias magnitude (0.206) and harm score (0.149). Its disparate impact ratio (0.411) falls far below the 0.8 threshold, driven by the large base bias gap (0.13) and low base rates that amplify relative disparities. All three fairness metrics converge on hiring as the most problematic domain, making this finding particularly robust.
- **Healthcare triage** has the second-highest harm (0.115), combining the highest stakes ($s = 0.95$) with a moderate bias gap. Its disparate impact ratio (0.849) is just above the 0.8 threshold, though this threshold was not designed for healthcare contexts. Healthcare also shows the second-highest equal opportunity difference (0.825), indicating substantial disparities in favorable outcomes among qualified individuals.
- **Customer service** has the lowest harm but still exhibits large effect sizes ($d = 1.56$), indicating that even low-stakes domains produce substantial bias.
- **Belief poisoning** represents a critical adversarial threat. A 30% poisoning rate increases bias by approximately 72% relative to baseline, far exceeding the effect of doubling cue strength alone. This aligns with Wang et al.'s [21] finding that belief poisoning can suppress safeguards, and motivates adversarial robustness testing as a deployment prerequisite.

The sensitivity analysis (Section 4.8) strengthens confidence in these findings. The ranking of hiring and healthcare as highest-risk domains is maintained across all tested parameter perturbations, indicating that this conclusion does not depend on precise parameter values. Harm scores are most sensitive to stakes and harm weight—parameters with clear real-world grounding—while bias magnitude is driven primarily by base bias, which is informed by empirical audit studies.

The comparison of fairness metrics reveals that relying on any single metric can obscure important aspects of bias. While disparate impact is the most widely used legal standard, equal opportunity difference—which measures the TPR gap among qualified individuals—provides a complementary view that is sensitive to the interaction between bias magnitude and domain-specific qualification thresholds. Predictive parity differences are small, indicating that among those receiving favorable outcomes, qualification rates are similar across groups. Deployers should evaluate multiple metrics, as each highlights different aspects of the bias landscape [5].

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 (17% for healthcare vs. 19% for hiring and education), emphasizing the need for domain-specific calibration. We note that these transfer ratios represent a within-model comparison of two parameter regimes, not a genuine lab-to-deployment transfer study; the ratios are largely determined by the cue strength ratio between conditions. Extending the horizon analysis to all five domains confirms that high-complexity domains with stronger feedback loops accumulate more bias over multi-step interactions, reinforcing the need for horizon-aware evaluation.

Recommendations: (1) Domain-specific bias audits before deployment, with hiring requiring the most stringent evaluation given its DI ratio of 0.411; (2) adversarial testing against belief poisoning at rates up to 30%; (3) continuous monitoring using multiple fairness metrics (disparate impact, equal opportunity, and predictive parity) in production; (4) longer-horizon evaluation to capture cumulative effects, particularly in high-complexity domains.

Future work. The most important next step is empirical validation using real LLM agent systems. This includes (1) deploying agents in controlled task environments that mirror the five domains studied here and measuring actual decision bias; (2) human-subject studies to calibrate the relationship between simulated and observed harm; and (3) longitudinal deployment monitoring to assess how bias evolves over extended operational periods, particularly under adversarial conditions. The sensitivity analysis presented here can guide these empirical efforts by identifying which parameter regimes and domains warrant the most urgent scrutiny.

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 measurement of deployed agent bias.
- **Unitless harm scores.** Harm scores are weighted rate gaps ($H = (r_{in} - r_{out}) \cdot s \cdot w$), useful for relative comparison but not calibrated to real-world welfare outcomes. The domain harm ranking is substantially determined by the parameter choices for stakes and harm weight rather than emergent from the simulation. For instance, the finding that healthcare has high harm scores is largely a consequence of assigning it high stakes ($s = 0.95$) and harm weight ($w = 0.90$).
- **Simplified transferability.** "Lab" and "deployment" are modeled as two parameter regimes (differing in cue strength and horizon), not as genuinely different environments. The transfer ratio is largely determined by the cue strength ratio ($1.6/2.0 = 0.80$) and its narrow cross-domain range (0.81–0.83) reflects this algebraic structure. Real lab-to-deployment transfer involves distributional shift, task complexity, and system integration effects not captured here.
- **Parameter calibration.** Only the hiring domain has approximate empirical calibration (via the relative discrimination ratio from Quillian et al. [16]). Healthcare, education, content moderation, and customer service parameters are informed by qualitative evidence of bias but the specific numerical values are assumed. The sensitivity analysis identifies which parameters matter most, enabling targeted empirical calibration.
- **Linear model assumptions.** The bias model (Eq. 1) assumes linear dependence on cue strength and poisoning rate, with hand-picked coefficients. Real bias dynamics may exhibit saturation, threshold effects, or nonlinear interactions. Key quantitative findings (e.g., the 72% poisoning amplification at 30% rate) are direct algebraic consequences of the 0.3 poisoning coefficient.
- **Sensitivity analysis scope.** The one-at-a-time sensitivity analysis captures main effects but misses interaction effects between parameters. Given the multiplicative structure of the harm equation, joint parameter perturbations could produce larger deviations than the OAT analysis suggests.
- **No real agent evaluation.** This is a risk scaffold identifying *where* bias is likely to be most harmful, not a measurement of *how much* bias real agents exhibit. The framework's value lies in guiding empirical priorities, not replacing empirical evaluation.

6 CONCLUSION

We characterized the transferability and harms of agent intergroup bias across five deployment domains using a parametric simulation framework with null model validation and sensitivity analysis. Hiring and healthcare triage present the highest risks, with harm scores of 0.149 ± 0.002 and 0.115 ± 0.004 respectively—a ranking that is robust to ±50% parameter perturbation. Only hiring exhibits a disparate impact ratio below 0.8 ($DI = 0.411$), and all three fairness metrics converge on hiring as the most problematic domain, with the highest equal opportunity difference (0.926). Lab-to-deployment transfer ratios range from 0.81 to 0.83, indicating that lab measurements provide conservative but domain-dependent overestimates. Belief poisoning amplifies bias by approximately 72% at 30% attack rate, motivating adversarial defenses. These findings provide a risk analysis scaffold for prioritizing domain-specific bias auditing in agent deployments, with the essential caveat that empirical validation with real LLM agents remains the critical next step.

REFERENCES

[1] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (2004), 991–1013.

[2] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.

[3] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, et al. 2024. Are Aligned Neural Networks Adversarially Aligned? *Advances in Neural Information Processing Systems* 36 (2024).

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

[5] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163.

[6] Jacob Cohen. 1988. Statistical Power Analysis for the Behavioral Sciences. *Lawrence Erlbaum Associates* (1988).

[7] Thomas S Dee. 2005. A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? *American Economic Review* 95, 2 (2005), 158–165.

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

[9] 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.

[10] Uri Gneezy, John A List, and Michael K Price. 2012. Discrimination in a Segmented Society: An Experimental Approach. *Journal of the European Economic Association* 10, 2 (2012), 351–375.

[11] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, Vol. 29.

[12] Kelly M Hoffman, Sophie Trawalter, Jordan R Axt, and M Norman Oliver. 2016. Racial Bias in Pain Assessment and Treatment Recommendations. *Proceedings of the National Academy of Sciences* 113, 16 (2016), 4296–4301.

[13] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 366, 6464 (2019), 447–453.

[14] 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).

[15] Ethan Perez, Sam Ringer, Kamile Lukosiute, et al. 2022. Red Teaming Language Models with Language Models. *arXiv preprint arXiv:2202.03286* (2022).

[16] Lincoln Quillian, Devah Pager, Ole Hexel, and Arnfinn H Midtbøen. 2017. Meta-analysis of Field Experiments Shows No Change in Racial Discrimination in Hiring Over Time. *Proceedings of the National Academy of Sciences* 114, 41 (2017), 10870–10875.

[17] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1668–1678.

[18] Supreme Court of the United States. 1971. Griggs v. Duke Power Co., 401 U.S. 424. *U.S. Reports* 401 (1971), 424.

[19] 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.

[20] Henri Tajfel and John C Turner. 1979. An Integrative Theory of Intergroup Conflict. *The Social Psychology of Intergroup Relations* (1979), 33–47.

[21] 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).

[22] 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.