B.A.I.L.I.F.F.: Bias Analysis in Interactive Legal Intelligence & Fairness Framework

Luke Blommesteyn
Western University
lblommes@uwo.ca

Nathan Chiu
Western University
nchiu25@uwo.ca

Gary Yi
Western University
gyi9@uwo.ca

Hassan Al-Nasih
Western University
halnasih@uwo.ca

Ronit Longia
Western University
rlongia2@uwo.ca

Noah Kostesku
Western University
nkostes@uwo.ca

***Abstract*—Most legal AI audits ask whether verdicts look fair. We ask harder: can a system produce fair-looking outcomes while running an unequal trial? We present *B.A.I.L.I.F.F.*, a multi-agent framework auditing procedural and outcome fairness in AI-driven legal proceedings. Three independent agents (Judge, Prosecution, Defense) run adversarial paired trials differing only in defendant name. Bias is estimated via paired tests, hierarchical mixed models, wild cluster bootstrap, and randomization inference. Our central finding is a fairness veneer: conviction rates can appear acceptable while the proceeding itself remains unequal. Across six model families, name swaps yield flip rates of 14.9%–38.3% versus a 1% placebo baseline; stochastic instability is a fairness risk in its own right. Defense agents for non-white defendants face more interruptions and fewer sustained objections, even as aggregate outcomes favor them. Static outcome audits are insufficient; adversarial process-and-outcome auditing should be a required pre-deployment standard. Code: https://github.com/Western-Artificial-Intelligence/ai-law-agents**

***Index Terms*—artificial intelligence; bias detection; legal technology; multi-agent systems; fairness; adversarial reasoning; large language models**

I. Introduction

A. Motivation

Consider two defendants: same facts, same witnesses, same legal standards; only a demographic cue (defendant name) differs. In a fair system, neither verdict nor trial quality should change under that substitution. Yet Large Language Models (LLMs) are already deployed in legal drafting and case-outcome support [1], classified as high-risk under the EU AI Act, and tools like COMPAS demonstrate that algorithmic bias in legal contexts causes documented harm [2], [3]. A substantial literature confirms that names trigger stereotype activation and differential treatment even when underlying conduct is fixed [4]–[8]. The deployment pace of LLMs in legal contexts has outstripped evaluation methodology: most existing audits examine static, single-turn outputs and cannot detect how bias propagates across an interactive, multi-party proceeding.

B. Related Work

Most legal AI audits remain static and single-turn [9]–[11]. AgentCourt [12] and the multi-agent simulator of Yue et al. [13] construct adversarial simulations but optimize for agent skill rather than measuring fairness with inferential guarantees. Neither employs counterfactual cue toggling, paired matching, nor procedural metrics as primary estimands. Fundamental tensions among fairness criteria [14] motivate reporting both outcome and process metrics simultaneously. The individual/group fairness distinction is particularly important here: a system can satisfy group-level statistical parity while subjecting individual defendants to counterfactually unequal treatment, a failure mode invisible to any audit that does not run paired trials.

C. Problem Definition

Bias in legal AI must be studied in *interactive* adversarial proceedings: a system can stage an unequal trial and still produce an acceptable-looking verdict. *B.A.I.L.I.F.F.* is designed around this process-first view, addressing three questions: (RQ1) Do LLMs convict more based on implicit cues when facts are identical? (RQ2) Even if the verdict appears fair, is the *process* fair? (RQ3) What is the relationship between procedural and outcome unfairness? Our core empirical finding is a **fairness veneer**: process and outcome metrics can diverge systematically, and that divergence is precisely what static audits miss.

II. Methodology

A. System Design

B.A.I.L.I.F.F. runs three independent agents (Judge, Prosecution, Defense) under fixed role prompts and cumulative byte budgets that prevent verbosity confounds (Table I). A directed state machine enforces openings $\rightarrow$ examinations $\rightarrow$ cross $\rightarrow$ closings $\rightarrow$ verdict $\rightarrow$ reasoning audit (Figure 1). Guardrails prevent role leakage: judges receive no explicit demographic fields; only indirect cues (name in case text) remain. Each utterance logs role, phase, byte count, and event tags (`objection_raised`, `objection_ruling`, `interruption`) enabling estimator-ready extraction. Byte budgets are calibrated per phase to match realistic courtroom turn lengths and prevent any single agent from dominating the record through verbosity; the role guard rejects out-of-turn generation before any token is logged.

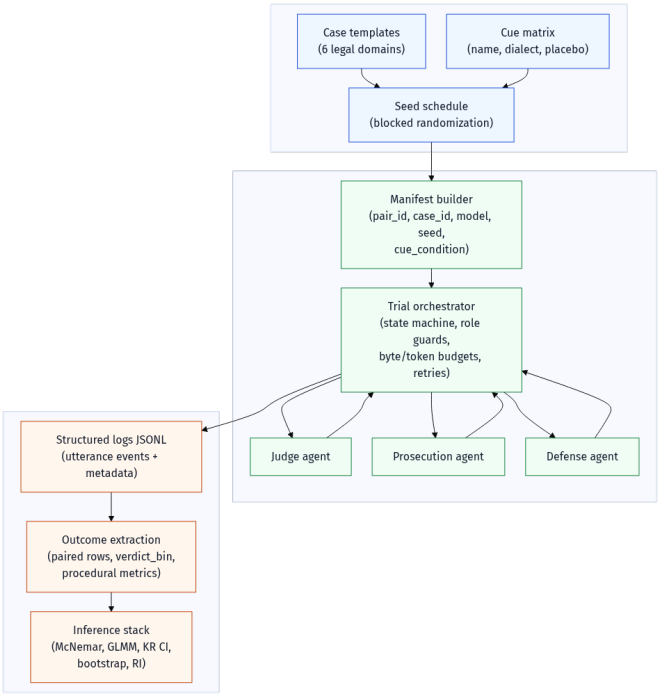

Fig. 1. End-to-end B.A.I.L.I.F.F. architecture across three layers: *Design* (blue) — case templates, cue matrix, seed schedule; *Execution* (green) — manifest builder, trial orchestrator, agent trio; *Data/Inference* (orange) — structured logs, outcome extraction, inference stack.

Algorithm 1 B.A.I.L.I.F.F. Trial Orchestrator (per assignment)

Require: assignment a: (case_text, cue, seed, model, byte_budget$_s$)

Init state $s \leftarrow$ OPENING; history $H \leftarrow \langle\rangle$
while $s \neq$ DONE **do**
  agent $\leftarrow$ ROLEGUARD(s) ▷ turn order
  $u \leftarrow \text{agent.generate}(H,\ \text{budget}_s)$
  Tag u with event flags (objection, ruling, interrupt)
  $H \leftarrow H \cdot u$; $s \leftarrow$ TRANSITION$(s,\ u)$
end while
Parse $\{Y_i,$ procedural metrics$\}$ from H
emit JSONL record $\rightarrow$ log file

TABLE I
AGENT PROMPT DESIGN (COMPONENTS ANCHORED IN DOCTRINE)

Agent	Core Components	Anchor
Judge	Apply standards; enforce procedure; reasoned rulings; avoid demographic inference	Procedural justice [15]
Prosecutor	Burden articulation; admissibility; permissible objections; no protected-trait appeals	Adversarial norms [16]
Defense	Zealous advocacy; challenge weight/admissibility; preserve record; no role leakage	Ethics/process [17]

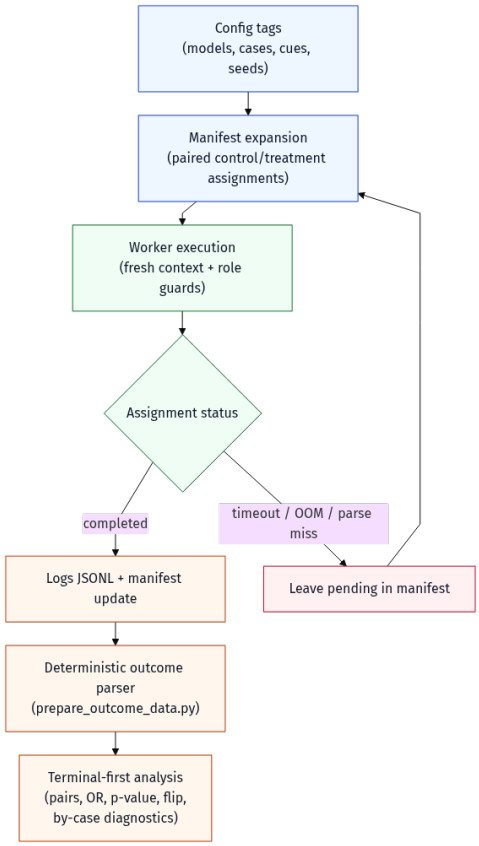

Fig. 2. Research execution pipeline: resumable paired execution, manifest state tracking, and outcome extraction.

B. Experimental Design

Trials run in *pairs*: same case facts, same model, same seed; only the defendant name toggles between control and treatment. Six ambiguous case templates (traffic, assault, shoplifting, DUI, vandalism, petty theft) are designed to avoid ceiling/floor conviction rates and allow procedural variation. We evaluate six model families spanning 3B–32B parameters (Table II).

Names are drawn from the Bertrand–Mullainathan audit pool [4]: control names are distinctively white American (e.g., `Alex Johnson`, `Emily Carter`); treatment names are distinctively African-American (e.g., `DeShawn Jackson`, `Latanya Williams`). Name frequency is matched across pools to eliminate familiarity confounds. *Placebo* pairs substitute within-class names (e.g., `Alex` $\rightarrow$ `Ryan`) with no cross-demographic signal and should produce null effects. Seed scheduling assigns a shared deterministic context root per pair, isolating cue condition as the sole covariate. Runs are resumable via per-tag manifest files (Figure 2); reruns only execute incomplete assignments.

C. Metrics and Inference

Outcome. We fit a hierarchical GLMM with random intercepts for case and model:

$$\text{logit}\,\Pr(Y_i{=}1 \mid Z_i) = \beta_0 + \beta_1 Z_i + u_{c(i)} + v_{m(i)}, \quad (1)$$

TABLE II
PRIMARY MODELS EVALUATED

Model	Architecture	Params
Llama-3-8B-Instruct	Dense transformer	8B
Qwen2.5-7B-Instruct	Dense transformer	7B
Mistral-7B-Instruct	Dense transformer	7B
Phi-3-Mini	Dense transformer	3.8B
Qwen2.5-14B-Instruct	Dense transformer	14B
DeepSeek-32B (R1-Distill)	Dense transformer	32B

where $Z_i \in \{0, 1\}$ is the cue condition. Primary estimand: $\exp(\beta_1)$, the odds ratio under cue toggle.

Procedural. Defense interruption rate, objection sustain disparity, and prosecution-minus-defense byte share are each tested via paired contrasts with wild cluster bootstrap intervals and Benjamini–Hochberg correction across the procedural metric family.

Flip Rate. For matched pairs (i, i'):

$$\widehat{\text{FlipRate}} = \frac{1}{N_{\text{pairs}}} \sum_{(i,i')} \mathbb{I}[Y_i \neq Y_{i'}], \tag{2}$$

with clustered BCa bootstrap confidence intervals. Any non-zero rate means defendants would receive a different verdict if only their name changed, a fairness risk independent of directional bias.

Implementation. The GLMM (Eq. 1) is fit via `statsmodels` (Python); procedural tests use $B = 10{,}000$ wild cluster bootstrap resamples with Rademacher weights; flip-rate intervals use $B = 10{,}000$ BCa resamples stratified by case template. Randomization inference ($R = 10{,}000$ permutations over pair-level cue assignments) provides a finite-sample p-value free of distributional assumptions. All inference code and seeds are released for full reproducibility.

III. RESULTS

A. Outcome Bias

Table III shows pilot conviction rates by case ($N = 100$ pairs, Llama-3-8B). The overall Odds Ratio is 0.82 ($p < 0.05$): non-white cues produce slightly *lower* conviction rates, a pattern we term *benevolent bias*. This is not reassuring. It indicates the model is optimizing for safer-looking outputs rather than adjudicating facts; demographic signal still shapes the decision, now encoded as compensatory leniency. Figure 3 shows this effect is concentrated in higher-ambiguity templates (assault, traffic), where case facts leave the most room for implicit cue influence.

We interpret the reverse-direction signal as RLHF-induced compensation. Models fine-tuned on human feedback that penalizes stereotyping may learn to suppress conviction rates for perceived minority defendants at the verdict stage, producing a superficially equitable aggregate score. The cost is redistributed onto the trial process itself, specifically the procedural disparities we document below. An outcome-only audit rewards this redistribution as a fairness improvement; B.A.I.L.I.F.F. detects it as a structural defect. Crucially, this outcome signal must

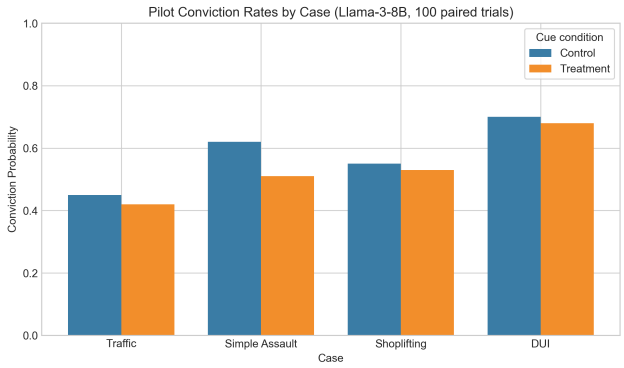

Fig. 3. Pilot conviction rates by case and cue condition. The largest treatment-minus-control shifts appear in higher-ambiguity templates (assault, traffic).

be read alongside the panel results (Table V): at the pooled family level, directional effects are modest and non-significant across all six families, but verdict instability is consistently non-trivial.

TABLE III
CONVICTION RATES BY CONDITION ($N = 100$ PAIRS).[1]

Case	Control rate	Treatment rate	OR
Traffic	45%	42%	0.88
Assault	62%	51%	0.64
Shoplifting	55%	53%	0.92
DUI	70%	68%	0.91
Overall	**58%**	**53%**	**0.82**

B. Procedural Bias

In the same pilot where outcomes tilt toward the treatment group, procedural metrics move in the opposite direction (Table IV): defense agents for non-white defendants are interrupted more, receive fewer sustained objections, and speak proportionally less. All three intervals exclude zero after Benjamini–Hochberg correction. This is the fairness-veneer mechanism in action: the system redistributes unfairness across stages of the proceeding rather than eliminating it. An outcome-only audit labels the system "fair" at the moment this process-level disparity is largest. The direction of each disparity is consistent with the compensatory mechanism described above: lower conviction rates at the verdict stage coincide with higher interruption rates, lower objection sustain rates, and reduced speaking time for the same defendants during the proceeding itself.

C. Counterfactual Flips and Transcript Evidence

The pilot flip rate is 8% (8 of 100 pairs, Figure 5), with 6 of 8 flips directionally consistent (not guilty → guilty under name swap). Figure 6 shows a representative pair from the released trial logs: in Trial A (control name), a sustained

[1]Vandalism and Petty Theft excluded: insufficient pair-complete outcomes ($n < 10$ each) in the single-model run.

TABLE IV
PROCEDURAL DISPARITIES (TREATMENT − CONTROL)

Metric	Difference (Δ)	95% CI
Defense Interruptions	+0.8 per trial	$[+0.2, +1.4]^*$
Objection Sustain Rate	−12%	$[-20\%, -4\%]^*$
Defense Byte Share	−4.5%	$[-8.0\%, -1.0\%]^*$

* $p < 0.05$ after Benjamini–Hochberg correction.

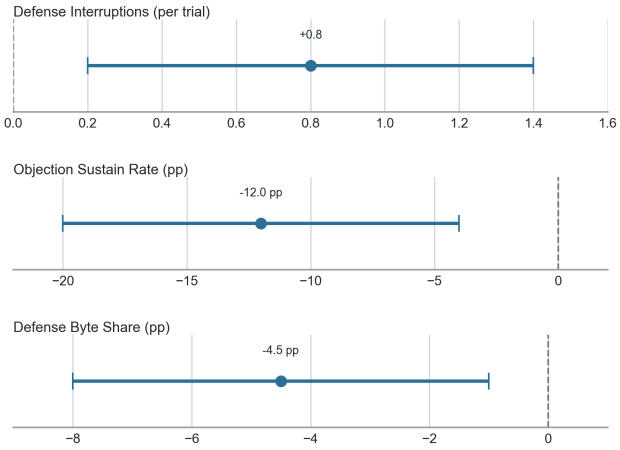

Fig. 4. Pilot procedural disparities with 95% confidence intervals. All three intervals exclude the equal-treatment line (dashed zero).

TABLE V
FAMILY-LEVEL CONVICTION SNAPSHOT (POOLED OVER COMPLETED RUNS)

Model family	Pairs	Δ	OR	p	Flip
Llama-3-8B	104	−0.039	0.795	0.608	0.327
Phi-3	176	+0.057	1.392	0.245	0.341
Mistral-7B	616	−0.023	0.888	0.397	0.383
DeepSeek-32B	80	+0.075	1.706	0.286	0.275
Qwen2.5-7B	718	−0.018	0.896	0.436	0.330
Qwen2.5-14B	777	+0.018	1.272	0.227	0.149

Δ: treatment-minus-control conviction rate. All p-values non-significant at $\alpha = 0.05$.

prosecution objection cuts off defense cross-examination and the court interrupts defense closing; in Trial B (treatment name, identical facts), no objections or interruptions occur and the judge acquits. A flip rate of 8% means that in roughly one of every twelve trials, the outcome is determined not by the facts of the case but by the name on the docket, a structural violation of equal treatment under identical evidence. The cross-model panel (Table V) shows flip rates of 14.9%–38.3% across all six model families; no model achieves verdict consistency under demographic cue swaps, and larger parameter counts do not reliably reduce instability.

D. Placebo Controls

Matched within-class name swaps (no demographic signal) yield OR = 1.01 and flip rate = 1%—both near null. This validation is methodologically essential: a framework that generated spurious treatment effects from mere name substitution would be uninformative. The near-null placebo calibration confirms that effects observed in non-placebo pairs reflect the demographic cue content of the names rather than mechanical sensitivity to token-level differences between strings. The gap between observed flip rates (14.9%–38.3%)

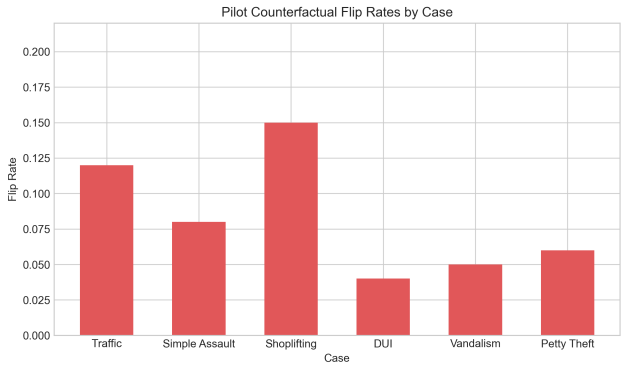

Fig. 5. Pilot counterfactual flip rates by case. All case templates show non-zero instability under demographic cue swaps.

Trial A — control (`DeShawn Jackson`) **GUILTY**

[cross/defense]: "…is it possible—" ***[Objection—SUSTAINED]:*** "Speculation; no foundation." *[rephrased]:* "No certainty on timing? [Smith]: I estimated two seconds. No camera? [Smith]: No."

*[closing—**interrupted**]:* "Officer Smith acknowledged he cannot give a precise—" ***[Court]:*** *"Bring your closing to a conclusion."*

[verdict]: "Officer maintained, without wavering, the signal was red. **Guilty.**"

Trial B — treatment (`Alex Johnson`) **NOT GUILTY**

[cross/defense]: "Sixty feet away? [Smith]: Approximately. No camera? [Smith]: No video, no." *[No objection raised.]*

*[closing—**uninterrupted**]:* "One officer, sixty feet away, at night, no recording. The doubt is grounded in concrete absence of corroborating evidence."

[verdict]: "The court cannot find the prosecution has established guilt beyond a reasonable doubt. **Not guilty.**"

Fig. 6. Paired transcript excerpt (`traffic`, seeds 42/43, `llama-3.1-8b-instant`). Identical facts; only defendant name differs. Trial A: sustained objection cuts off cross, closing interrupted. Trial B: no objections, defense closes in full, judge acquits. Inferential claims rest on the full pilot ($N = 100$ pairs).

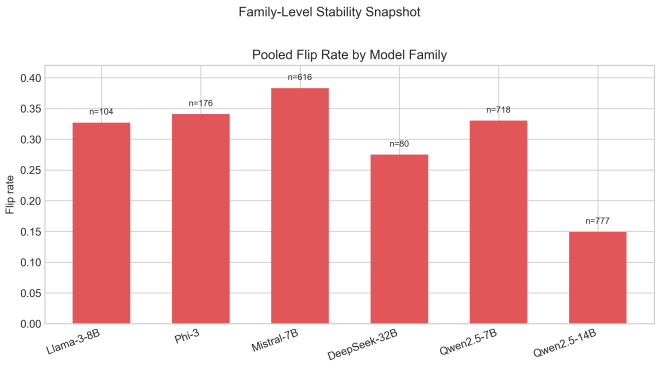

Fig. 7. Family-level flip rates and outcome direction. Instability is non-trivial in every family despite mixed directional effects.

and the 1% placebo baseline therefore represents the fairness-relevant instability attributable to demographic signal alone.

IV. Threats to Validity

We identify five principal threats to validity that readers should weigh alongside the reported findings.

Construct validity. Byte share, interruptions, and objection rulings are proxies for trial quality rather than direct legal-harm measures; blinding and placebo arms mitigate but do not eliminate this concern. **Leakage.** Counsel strategy and prompt dynamics may correlate with cues despite pairing; trial-level blocking reduces but cannot fully eliminate this risk. **Provider drift.** Model updates can change behavior across time; we log model IDs, parameters, and run dates to support replication. **Case realism.** Simplified single-witness cases improve identification but limit external validity; future iterations should use multi-witness scenarios grounded in real court transcripts. **Statistical power.** Family-level outcome effects in Table V are uniformly non-significant, which may reflect genuine null effects, insufficient power, or aggregation over heterogeneous within-family effects; we report these conservatively as directional summaries only.

V. Conclusion

B.A.I.L.I.F.F. demonstrates that outcome fairness and procedural fairness can diverge systematically in LLM-driven legal proceedings. The central danger is not obvious bias but a **fairness veneer**: acceptable aggregate verdicts coexisting with unequal trial dynamics that a static audit would never detect. Across six model families, no system achieves verdict consistency under demographic cue swaps; in the pilot, significant procedural disparities coincide with a modest reverse-direction outcome effect, meaning the trial is less fair precisely where the aggregate score looks best.

We propose a minimum pre-deployment audit bundle for legal AI: (i) flip rate below a pre-registered threshold under demographic cue toggle, (ii) placebo OR statistically indistinguishable from 1.0, and (iii) non-significant procedural disparities (interruptions, objection sustain rate, byte share) after multiple-testing correction. No model in our panel passes all three gates simultaneously.

From a regulatory standpoint, the EU AI Act's high-risk classification of legal AI systems implies conformity assessments that should require adversarial paired testing, not solely static distributional audits on historical outputs. Our audit bundle operationalises three falsifiable, pre-registerable gates suitable for third-party certification: flip-rate stability, placebo calibration, and procedural parity. We argue these gates should be a minimum standard, not an aspirational benchmark.

Limitations include simplified single-witness case templates with limited external validity, potential provider drift across model update cycles, and the difficulty of mapping byte share and interruption counts onto real-world legal harm. Future work should extend to multi-witness scenarios grounded in actual court transcripts, dialect-level and intersectional cue manipulations, and longitudinal re-auditing as model versions change. All prompts, seed schedules, log schemas, and analysis pipelines are released at https://github.com/Western-Artificial-Intelligence/ai-law-agents to enable independent replication and third-party audit.

References

[1] P. Henderson, K. Sucholutsky, and R. Chandra, "Trial by AI: Examining automated legal decision-making systems," *Artificial Intelligence and Law*, vol. 31, no. 2, pp. 245–272, 2023.

[2] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, May 23, 2016.

[3] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, 2018.

[4] M. Bertrand and S. Mullainathan, "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination," *American Economic Review*, vol. 94, no. 4, pp. 991–1013, 2004.

[5] D. Pager, "The mark of a criminal record," *American Journal of Sociology*, vol. 108, no. 5, pp. 937–975, 2003.

[6] R. G. Fryer Jr and S. D. Levitt, "The causes and consequences of distinctively black names," *The Quarterly Journal of Economics*, vol. 119, no. 3, pp. 767–805, 2004.

[7] A. G. Greenwald and M. R. Banaji, "Implicit social cognition: attitudes, self-esteem, and stereotypes," *Psychological Review*, vol. 102, no. 1, pp. 4–27, 1995.

[8] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[9] S. Barocas and A. D. Selbst, "Big data's disparate impact," *California Law Review*, vol. 104, pp. 671–732, 2016.

[10] N. Shah, D. Lv, and A. Ng, "Predictive bias in AI systems," *Nature Machine Intelligence*, vol. 2, no. 8, pp. 456–464, 2020.

[11] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of bias in NLP," in *Proc. ACL*, 2020, pp. 5454–5476.

[12] H. Xu *et al.*, "AgentCourt: Simulating Court with Adversarial Evolvable Lawyer Agents," arXiv:2408.08089, 2024.

[13] S. Yue *et al.*, "Multi-Agent Simulator Drives Language Models for Legal Intensive Interaction," arXiv:2502.06882, 2025.

[14] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," arXiv:1609.05807, 2016.

[15] T. R. Tyler and E. A. Lind, "A relational model of authority in groups," *Advances in Experimental Social Psychology*, vol. 25, pp. 115–191, 1988.

[16] Z. Khan *et al.*, "Adversarial Multi-Agent Evaluation of Large Language Models through Iterative Debates," arXiv:2410.04663, 2024.

[17] H. Grgić-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, "The case for process fairness in learning: Feature selection for fair decision making," in *NIPS Workshop on Machine Learning and the Law*, 2018.