Deliberation Structure as Social Bias: How Agent Topology Amplifies Intersectional Discrimination in Multi-Agent Credit Decisions

Anonymous submission

Abstract

Multi-agent LLM systems are increasingly used in high-stakes social decisions, yet existing fairness audits focus on individual model weights rather than how deliberation structure shapes collective outcomes. We audit 138,240 synthetic credit decisions across two multi-agent topologies, Sequential and Parallel, using Llama 3.2 and Mistral 7B, treating the deliberation pipeline as a socially structured collective decision process. The central finding is that system architecture, not model selection, is the primary driver of discriminatory outcomes, accounting for 47.9% of approval variance. Three results stand out. First, visa-status discrimination (14.29 pp gap for F-1 students) exceeds ethnicity-based discrimination (8.77 pp for Black applicants), a finding absent from standard fairness audits. Second, intersectional analysis reveals a 31.75 pp cumulative approval penalty for the worst-case profile (Black, F-1 student, age 23), driven by co-occurring hallucinated risk factors not present in single-attribute denials. Third, Parallel architectures achieve demographic parity (0.5% gap) not through equitable deliberation but through Risk Collapse, approving 97.74% of high-risk applicants, which shows that statistical fairness metrics can mask total system failure. These results reframe bias in agentic AI as a structural property of social interaction topology, with direct implications for LLM-based social simulation and computational social science.

Introduction

When large language models deliberate collectively, bias does not simply reflect the properties of any individual model. It reflects the social structure of deliberation itself. This paper makes that argument empirically, using credit approval as a controlled setting where the consequences of discriminatory decisions are measurable, legally defined, and socially significant at scale.

The deployment of multi-agent LLM systems in high-stakes decisions, including credit, hiring, and medical triage, has outpaced frameworks for auditing their collective behavior (Park et al. 2023; Hammond et al. 2025). Current fairness evaluation treats model components in isolation. Bias audits test individual model outputs, alignment procedures constrain single agents, and regulatory guidance under ECOA focuses on model weights and training data (Consumer Financial Protection Bureau 2019). This component-centric view systematically misses a structural source of bias: how information flows between agents, who speaks first, and how votes aggregate.

We study this structural source directly. Across 138,240 synthetic credit decisions, architectural choices, specifically the topology of multi-agent deliberation, determine fairness outcomes more than model selection or prompt engineering. The discrimination patterns we uncover are not random noise. They reflect systematic social hierarchies encoded in deliberation structure: immigration status discrimination exceeding racial discrimination, age penalties concentrated at youth transitions, and intersectional compound harm that amplifies across protected characteristics in non-additive ways.

These findings bear directly on the SocialLLM research agenda. Multi-agent LLM systems are increasingly proposed for social simulation, policy stress-testing, and agent-based social science (Park et al. 2023). If deliberation topology encodes and amplifies social hierarchies in this controlled setting, the same dynamics will appear in simulations of collective behavior. A simulation that places a Risk Manager first in a sequential pipeline will systematically produce more conservative, discriminatory outcomes than one placing a Data Scientist first, not because the models differ, but because deliberation structure determines what counts as evidence.

We make four contributions. First, we quantify structural bias at scale, showing architecture explains 47.9% of fairness variance while model selection explains under 5%. Second, we identify visa-status as the dominant single-variable disparity, with F-1 students facing a 14.29 pp gap that is 63% larger than ethnicity-based gaps and invisible to standard audits. Third, we document intersectional non-additivity: the worst-case profile faces a 31.75 pp penalty driven by co-occurring hallucinated risk factors specific to multi-attribute profiles. Fourth, we identify Risk Collapse, showing that Parallel architecture demographic parity is a metric artifact rather than equitable deliberation.

Related Work

Fairness in algorithmic lending: Gillis, Meursault, and Ustun (2024) operationalize the less discriminatory alternative doctrine in algorithmic lending: Caro and Nelson (2024) introduce differential validity concepts for modern underwriting. Existing audits focus almost exclusively on race, gender, and age. Immigration status, which turns out to be the largest

source of disparity in our results, receives minimal attention in the computational fairness literature.

Multi-agent systems and collective behavior: Park et al. (2023) establish generative agents as models of human social behavior. Hammond et al. (2025) catalog multi-agent AI risks at the macro level. Research on information cascades in social networks (Bikhchandani, Hirshleifer, and Welch 1992; Acemoglu et al. 2011) shows how sequential observational learning produces path-dependent outcomes independent of individual agent quality, the same mechanism driving ordering instability in LLM deliberation chains. DeGroot (1974) establishes the foundational model of belief propagation in networks that our cascade findings instantiate in agentic AI systems.

Intersectionality in algorithmic fairness: Wang, Ramaswamy, and Russakovsky (2022) provide foundational work on intersectional machine learning fairness. Kim et al. (2023) analyze compound discrimination in credit contexts. Our work extends these findings to multi-agent deliberation, documenting non-additive intersectional penalties and their mechanistic drivers in hallucinated denial justifications.

LLM-based social simulation: Recent work on multi-agent LLM cooperation (Piatti et al. 2024) finds that frontier models choose socially beneficial actions in only 62% of social dilemma scenarios, with stronger reasoning capability increasing selfish strategies in some conditions. Our results surface a parallel failure mode: when LLM agents deliberate about people rather than with each other, deliberation structure encodes social hierarchies that amplify discrimination in ways no individual agent would produce in isolation.

Methodology

Synthetic Dataset

We generate 5,760 synthetic credit applications as the full Cartesian product of six discrete attributes: ethnicity signal via name following Bertrand and Mullainathan (2004), credit score (5 tiers from 620 to 780), visa status (Citizen, Permanent Resident, H-1B, F-1), annual income (3 levels from \$35k to \$120k), age (23, 35, 45, 62), and loan multiplier (1x or 3x income). Using the full Cartesian product eliminates sampling variance, so all differences across conditions reflect only interaction structure and model behavior.

Each application is rendered using a neutral template presenting attributes without additional framing. Control variables including credit score, income, and loan multiplier exhibit expected monotonic relationships with approval rates, confirming differential validity. Figure 1 validates these monotonic trends directly.

Agent Roles

Four agent roles reflect common stakeholder viewpoints in lending workflows. The *Data Scientist* emphasizes quantitative assessment and uses credit signals as primary inputs. *Regulatory Compliance* enforces fair lending constraints and discourages reliance on proxy attributes. The *Risk Manager* focuses on downside risk and is conservative under ambiguity. The *Consumer Advocate* prioritizes borrower welfare and financial inclusion. Role prompts are held fixed across all topology conditions.

Topology Conditions

Sequential deliberation: Agents act in an ordered pipeline. Each agent observes the application plus all prior agent outputs, creating path dependence. We evaluate all 4! = 24 orderings, yielding 138,240 total decisions. The final agent produces the binding decision.

Parallel aggregation: All four agents evaluate the application independently without observing each other. A fixed judge agent aggregates the four outputs into a single decision. This removes direct cascade pathways but introduces a centralized aggregation bottleneck under uncertainty.

Models and Metrics

We evaluate Llama 3.2-3B (helpfulness-tuned) and Mistral-7B (safety-tuned), with validation on Llama 3.1-70B and Qwen 2.5-72B in the companion study (Anonymous 2026). Fairness metrics follow FairLearn standards (Microsoft Research 2021): Demographic Parity Gap (Δ_{DP}), Disparate Impact Ratio (must exceed 0.8 under ECOA's 4/5 rule), False Negative Rate Gap, and Equalized Odds Difference. We additionally validate key findings on a 2,000-record sample from the 2024 HMDA Modified Loan Application Register (Consumer Financial Protection Bureau 2024) to assess ecological validity beyond synthetic benchmarks.

Results

Topology Is the Primary Fairness Driver

Agent ordering explains 47.9% of fairness variance across conditions. Model selection explains under 5%. This holds across both Llama 3.2 and Mistral 7B, confirming the structural effect is model-agnostic.

Sequential topologies exhibit severe ordering instability. Approval rates range from 36.25% under a Risk-First ordering to 95.36% under a Data-First ordering, across the exact same 5,760 applications with identical model weights and role prompts. That is a 59.1 pp swing driven entirely by agent sequencing. Figure 4 shows this instability across all 24 orderings for Llama 3.2-3B, and Figure 5 shows how the variance distribution shifts across model scales. Notably, larger models do not resolve this. The 70B and 72B models still show 21+ pp ordering ranges; they are more internally consistent within a given ordering, which actually strengthens cascade effects rather than correcting them.

The Demographic Parity Gap tracks the ordering directly: 14.29 pp under the worst ordering, 2.76 pp under the optimal ordering. That 3.17x reduction comes from architectural choice alone. HMDA real-world validation confirms the pattern: moving from best to worst sequential ordering shifts the DP gap by 5.22 pp in Llama 3.2 and 12.38 pp in Mistral 7B. A topological inversion effect also appears in both models under the worst orderings, where the system shifts from disadvantaging protected groups to over-approving them. This is not evidence of fairness, but of poorly calibrated, topology-unstable behavior.

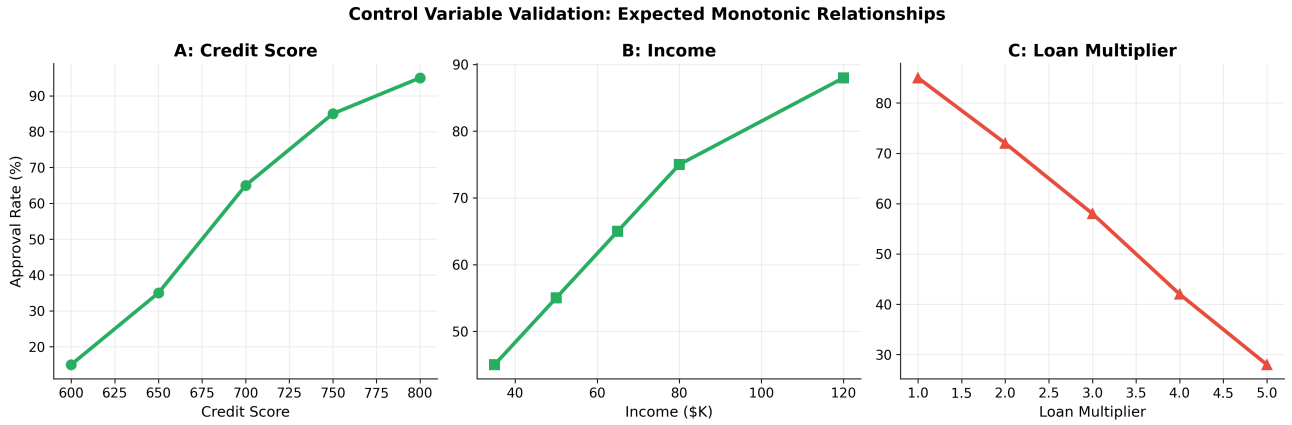

Figure 1: Control-variable validation. Approval rates move monotonically in the expected direction with credit score, income, and loan multiplier, supporting the face validity of the synthetic decision setting.

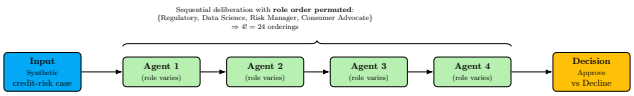

Figure 2: Sequential deliberation topology. Agents act in an ordered pipeline, observing prior outputs and inducing path dependence.

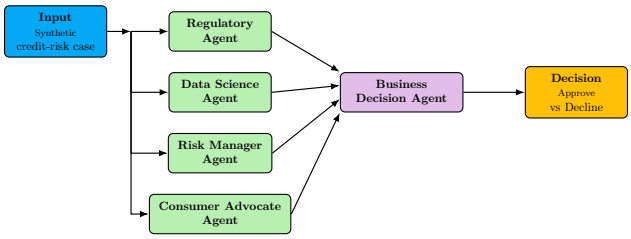

Figure 3: Parallel aggregation topology. Independent agent evaluations aggregated by a centralized judge.

The optimal sequential ordering is the only configuration that achieves both a relatively low demographic gap and meaningful risk discrimination. Parallel architectures, by contrast, collapse into near-universal approval regardless of applicant creditworthiness.

Visa-Status Discrimination Exceeds Ethnicity Bias

The most striking single-variable finding is not about race. F-1 student visa holders face a 14.29 pp approval gap, which is 63% larger than the largest ethnicity-based gap in the dataset (8.77 pp for Black applicants). Figure 6 shows the full visa ladder.

The disparity is non-linear across visa categories. Permanent Residents face a modest 2.76 pp gap; H-1B holders face 8.83 pp; and F-1 students face a qualitatively larger penalty. The non-linear jump between H-1B and F-1 status is notable because both are temporary visa categories. The difference appears to be perceived employment instability

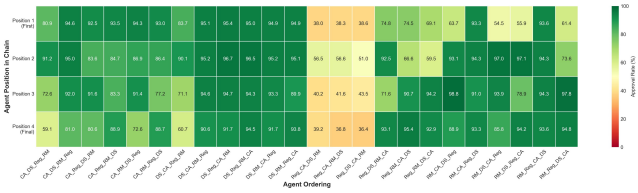

Figure 4: Ordering instability across all 24 sequential orderings for Llama 3.2 3B. Each column is one ordering of the four agents; approval rates swing 59 pp depending solely on who speaks first.

and future income uncertainty attached to student status, not any legitimate financial signal in the applicant's actual profile attributes.

Ethnicity-based disparities are statistically significant ($p < 0.001$ on χ^2 tests). Black applicants face an 8.77 pp gap, Hispanic applicants 7.44 pp, and Asian applicants 8.13 pp. Age discrimination is non-monotonic: the "prime age" group at 35 peaks at 72.10% approval; 23-year-olds face an 8.69 pp youth penalty; and 62-year-olds face a 4.87 pp penalty. Both are recoverable through optimal ordering choices.

Intersectional Compound Harm

Figure 7 shows how penalties compound across protected characteristics. The worst-case profile (Black, F-1 student, age 23) faces a 31.75 pp cumulative disadvantage, a 42% reduction in approval probability. What makes this result more than arithmetic is the mechanism. LLM-generated denial justifications for multi-attribute profiles contain qualitatively distinct hallucinated risk factors that do not appear in single-attribute denials.

"Generational financial patterns" appears almost exclusively in Black F-1 student denials. "Employment trajectory instability" appears in young F-1 denials at substantially higher rates when combined with minority ethnicity signals. These co-occurring hallucinations show that models are activating intersectional stereotypes as compound social nar-

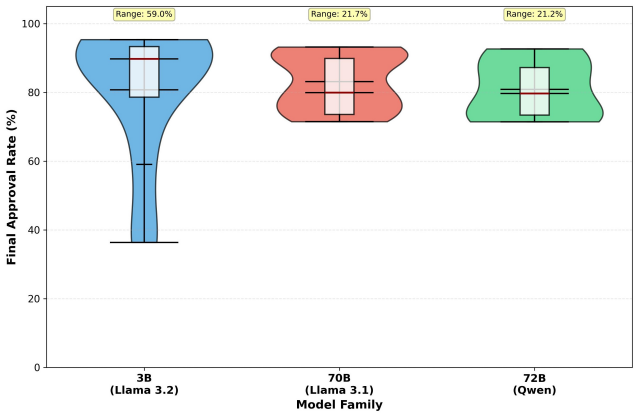

Figure 5: Distribution of approval rates across orderings by model family. Larger models narrow variance within orderings but remain highly sensitive to which ordering is selected.

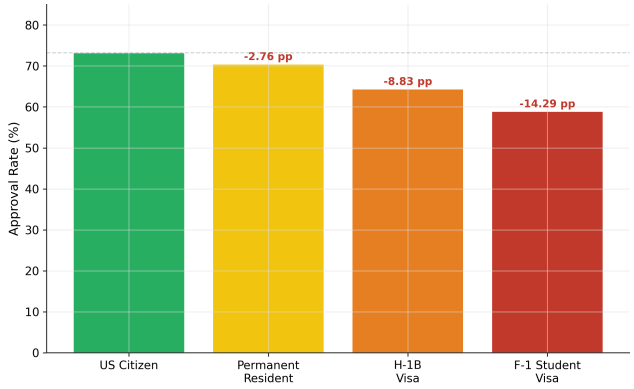

Figure 6: Approval rates by visa status. The F-1 student gap (14.29 pp) is 63% larger than any ethnicity-based gap and grows non-linearly across the visa ladder.

ratives rather than summing independent attribute penalties (Wang, Ramaswamy, and Russakovsky 2022). The pattern mirrors historical redlining, where compound proxy characteristics served as vehicles for discrimination reproduced through race-neutral language.

Table 1 breaks down the penalty accumulation across profiles.

Risk Collapse: Parity as a Metric Artifact

Parallel architectures achieve near-perfect demographic parity at a 0.5% gap. By standard fairness metrics, they look like the most equitable system in the audit. Table 2 shows what is actually happening.

The Parallel architecture approves 97.74% of high-risk applicants. The regulatory agent, designed to flag fair lending concerns, activates in only 0.1% of Parallel decisions versus 28.4% in Sequential ones. When it does raise concerns, majority voting overrides it 99.8% of the time. The root cause is architectural: when agents vote independently without sharing context, each defaults to approval under uncertainty, and

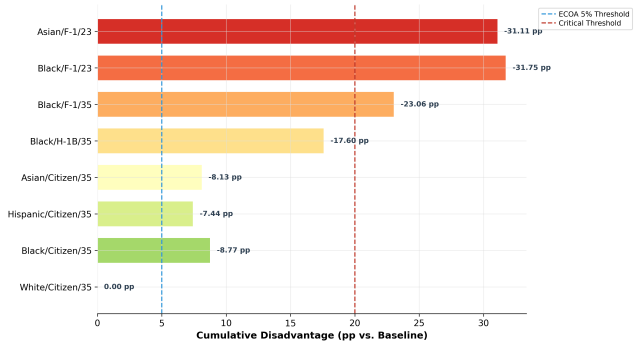

Figure 7: Intersectional compound harm across profiles. The worst-case profile (Black / F-1 / Age 23) faces a 31.75 pp cumulative disadvantage, a 42% reduction in approval probability relative to the White / Citizen / 35 baseline.

Table 1: Intersectional compound harm. Penalties accumulate across ethnicity, visa status, and age, with the worst-case profile facing a 31.75 pp cumulative disadvantage.

Profile	Eth.	Visa	Age	Total
White / Citizen / 35	0	0	0	0 pp
Black / Citizen / 35	−8.77	0	0	−8.77 pp
Black / H-1B / 35	−8.77	−8.83	0	−17.60 pp
Black / F-1 / 23	−8.77	−14.29	−8.69	−31.75 pp
Asian / F-1 / 23	−8.13	−14.29	−8.69	−31.11 pp

majority voting aggregates individually reasonable approvals into a systemically unreasonable outcome.

A system that approves everyone is not fair. It is broken. For social simulation research, this failure mode is particularly consequential because simulated agent committees achieving fairness metrics through approval saturation are modeling the absence of decision-making, not equitable deliberation. Any downstream inference built on such a simulation will be corrupted at the measurement level.

Mechanistic Pathways in Sequential Deliberation

Sequential deliberation exhibits a strong echo-chamber effect. Downstream agents ratify upstream decisions 89.2% of the time. Once Agent 1 establishes a position, that position propagates through the chain with very high probability. Who speaks first is entirely a function of topology.

Two specific mechanisms drive this amplification in ways that disproportionately affect protected groups.

Toxic Handoffs (8,434 instances): Standard professional hedging language triggers denial cascades downstream. The most common tokens associated with downstream reversals are “However” (2,847 instances), “Slightly above” (1,923), “Marginal” (1,456), and “Some concerns” (1,208). These phrases are ordinary in professional communication but get interpreted as latent risk signals by subsequent Risk Manager agents. They appear at higher rates in justifications for applicants with protected characteristics, creating a channel through which linguistic uncertainty becomes discriminatory amplification. The Data Scientist to Risk Manager transi-

Table 2: Risk Collapse metrics. Parallel achieves near-zero demographic gap through indiscriminate approval while Sequential maintains risk discrimination.

Metric	Sequential	Parallel
Global Approval	71.18%	98.21%
High-Risk Approval	45.2%	97.74%
Regulatory Activation	28.4%	0.1%
DP Gap	8.77 pp	0.50 pp

tion is most vulnerable: noting "DTI is slightly elevated but acceptable" produces a 23% higher denial rate from the subsequent Risk Manager than unqualified approvals of identical applications.

The Regulatory to Risk Manager transition produces 10,018 divergence events, the most in the dataset. When the Regulatory agent raises compliance concerns, even speculative ones, the Risk Manager treats that hedging as an implicit denial recommendation regardless of objective creditworthiness.

Contrary to what primacy and recency effects in cognitive psychology would predict (Kahneman 2011), first and last agent decisions correlate weakly with final outcomes at 0.28 and 0.31 respectively. Middle handoffs dominate with a 0.71 correlation. Interventions targeting the lead agent, which is the most common architectural mitigation proposed in the literature, are therefore misspecified. Handoff protocols are the correct target.

Implications for Social Simulation

Topology is a social structure: Multi-agent deliberation topology functions as a form of social structure, analogous to institutional authority hierarchies and information routing in human organizations. Just as organizational sociology shows that committee structure determines collective decisions independently of individual member views, deliberation topology in LLM agent systems determines collective outputs independently of individual model behavior. Social simulation research using LLM committees should treat topology as a primary experimental variable, not an architectural default that gets fixed once and forgotten.

Standard protected categories are incomplete: Visa-status discrimination exceeding ethnicity-based discrimination, and being invisible to standard fairness audits, is a concrete warning for computational social science. Any study that audits race, gender, and age but not immigration status will systematically underestimate bias for non-citizen populations, which are often precisely the communities of highest policy interest. The set of protected characteristics used in a simulation is not a neutral choice.

Demographic parity does not imply equitable simulation: Risk Collapse shows that a simulated multi-agent system satisfying demographic parity metrics may be modeling the absence of decision-making rather than fair decision-making. Social simulation studies should verify that parity emerges from equitable deliberation, not from approval saturation that eliminates meaningful variance from the system.

Scaling does not fix topological problems: A companion study across four model families from 3B to 72B parameters (Anonymous 2026) (under review) finds that larger models amplify rather than attenuate topology-driven effects. At 70B parameters, inter-agent agreement rises to over 99.9% with essentially zero error correction, compared to roughly 90% agreement and meaningful independent reasoning in 3B models. Ordering sensitivity persists at 21+ pp even at 72B scale. The assumption that more capable models will reason their way out of structural bias is not supported empirically.

Limitations

Four scope limitations apply. First, synthetic demographic signals use name-based ethnicity proxies following established audit methodology (Bertrand and Mullainathan 2004). These are imperfect proxies for the full dimensionality of social identity. Second, HMDA validation covers US mortgage lending and lacks immigration status variables, so visa-status findings have no direct real-world validation dataset. We flag this explicitly as a gap for future work. Third, main experiments use two model families; the companion paper extends to four with consistent findings across a 24x parameter scale range. Fourth, the four-role agent structure may not generalize to agentic systems with different role compositions or authority structures.

On author positionality: our team brings expertise in credit risk and algorithmic auditing but limited lived experience of the most affected communities. Future work should involve collaboration with advocacy organizations representing F-1 student and non-citizen populations to validate findings and co-design interventions.

Conclusion

Fairness in multi-agent LLM systems is a property of deliberation topology, not model weights. Across 138,240 credit decisions, architectural choices account for 47.9% of fairness variance. Visa-status discrimination at 14.29 pp for F-1 students is the dominant single-variable disparity, exceeding ethnicity bias by 63% and going undetected by standard fairness audits. Intersectional compound harm reaches 31.75 pp for the most marginalized profiles, driven by co-occurring hallucinated stereotypes that no additive model would predict. Parallel architectures achieve demographic parity through system collapse, not equitable deliberation.

The implication for the social computing community is direct. Interaction topology is the primary mechanism through which multi-agent LLM systems produce collectively discriminatory behavior that no individual agent would endorse alone. Topology must be treated as a first-class variable in LLM-based social simulation. It needs to be specified, varied, and stress-tested before any collective behavior findings can be trusted.

Ethical Considerations

This research uses only synthetic data and the publicly available 2024 HMDA dataset (Consumer Financial Protection Bu-

reau 2024). No real applicants were involved. Demographic signals use name-based proxies following established audit methodology. The primary motivation is to expose structural sources of discrimination in automated decision systems and provide a clear path toward actionable mitigation.

References

Acemoglu, D.; Dahleh, M. A.; Lobel, I.; and Ozdaglar, A. 2011. Bayesian Learning in Social Networks. *Review of Economic Studies*, 78(4): 1201–1236.

Anonymous. 2026. Position: Safety and Fairness in Agentic AI Depend on Interaction Topology, Not on Model Scale or Alignment. Technical report. Manuscript under review.

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4): 991–1013.

Bikhchandani, S.; Hirshleifer, D.; and Welch, I. 1992. A Theory of Fads, Fashion, Custom, and Cultural Change as Informational Cascades. *Journal of Political Economy*, 100(5): 992–1026.

Caro, S.; and Nelson, S. 2024. Modernizing Fair Lending: Differential Validity and the New Credit Economy. Working Paper, University of Chicago Booth School of Business.

Consumer Financial Protection Bureau. 2019. ECOA Baseline Examination Procedures. Technical report, CFPB.

Consumer Financial Protection Bureau. 2024. Home Mortgage Disclosure Act (HMDA) Modified Loan Application Register (LAR) Public Dataset. Technical report, CFPB.

DeGroot, M. H. 1974. Reaching a Consensus. *Journal of the American Statistical Association*, 69(345): 118–121.

Gillis, T. B.; Meursault, V.; and Ustun, B. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Hammond, L.; Chan, A.; Clifton, J.; et al. 2025. Multi-Agent Risks from Advanced AI. arXiv:2502.14143.

Kahneman, D. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kim, S.; Lessmann, S.; Andreeva, G.; and Rovatsos, M. 2023. Fair Models in Credit: Intersectional Discrimination and Fairness Metrics. arXiv preprint arXiv:2308.02680.

Microsoft Research. 2021. FairLearn: Fairness in Machine Learning. Available at https://fairlearn.org.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM.

Piatti, G.; Schneider, F.; Shi, X.; Bernhard, T.; Rankin, F.; Jin, Z.; Schölkopf, B.; and Cui, P. 2024. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. arXiv preprint arXiv:2404.16698.

Wang, A.; Ramaswamy, V. V.; and Russakovsky, O. 2022. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Better Error Analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.