Beyond Automation: Understanding Fairness, Ethics, and Human Discretion in AI-driven Societal Decisions

Gaurab Pokharel
Virginia Tech
Alexandria, VA
gaurab@vt.edu

Abstract

My doctoral research investigates the ethical readiness and societal implications of deploying AI in high-stakes resource allocation. This work bridges empirical analysis of AI capabilities with computational modeling of human decision-making and theoretical explorations of long-term fairness. Through a multi-method approach, I first evaluate the reliability of Large Language Models (LLMs) in a real-world homelessness services context, revealing critical inconsistencies. I then use interpretable machine learning to model the sophisticated discretion of human caseworkers, demonstrating that their non-formulaic judgments are systematic and effective. Finally, I use agent-based simulations to show how repeated, ostensibly "fair" allocations can entrench group-level inequities over time. Collectively, these findings argue for caution in autonomous AI deployment, highlight the value of human-in-the-loop systems, and call for a more dynamic understanding of fairness in sociotechnical systems.

Introduction and Motivation

My parents have always preferred working with a bank teller rather than an ATM because they trust a human to sense their needs and *respond with care*. This simple observation inspired my first research project: training a neural network to recognize human emotions. The model performed well on its test set but failed completely when I pointed the camera at my own face. The reason was stark. Its training data consisted almost entirely of Caucasian faces. This led me to a formative realization early on in my academic career: AI systems, even when designed with the best intentions, are not abstract mathematical objects. They are sociotechnical artifacts reflecting the limitations of their data and the values of the complex social systems in which they are embedded.

This realization anchors my doctoral research, which evaluates the ethical readiness of AI for real-world scarce resource allocation in high-stakes domains. As AI systems promise unprecedented efficiency, their use in ethically sensitive domains (Maitra et al. 2025; Taylor 2024; Bender and Hanna 2025) compels us to confront fundamental questions: Can machines adjudicate the complex moral trade-offs on which human well-being depends? Will automated decisions alleviate — or entrench — existing social inequities? And how can we reconcile algorithmic speed with the indispensable contextual knowledge of human discretion? While much of the technical literature focuses on static fairness metrics that assess outcomes at a single point in time (Hardt, Price, and Srebro 2016; Dwork et al. 2012), these snapshots often fail to capture the long-term, dynamic effects of algorithmic processes (Gohar et al. 2025; Corbett-Davies et al. 2023). To address this gap, my work develops methods for identifying and mitigating algorithmic harms by interrogating both the statistical properties of algorithms and the dynamic, institutional contexts in which they operate, advancing AI toward more just and trustworthy deployment.

Central Research Questions

To operationalize this inquiry, I have structured my doctoral research program around three interconnected questions that bridge empirical analysis with theoretical investigation:

1. **AI Judgments vs. Human Values:** To what extent can large-scale AI models replicate nuanced human moral judgments in resource allocation, and what ethical risks emerge when they fail?
2. **Modeling Human Discretion:** How do "street-level bureaucrats" exercise discretion beyond prescriptive rules, and what can computational models of their behavior teach us about designing effective and ethical human-AI systems?
3. **Temporal Fairness Dynamics:** How do merit-based fairness criteria evolve over repeated allocations, and under what conditions might seemingly fair processes generate entrenched group-level inequities over time?

Research and Findings

My research follows a clear arc from the empirical to the theoretical. I begin by identifying the limitations of current AI systems, model the human expertise they lack, and finally explore the long-term consequences of automated decisions.

Street-Level AI (Pokharel et al. 2025): Reliability of LLMs in Homelessness Decision-Making LLMs have captured public attention and have been broadly touted for their potential to streamline human work. In a forthcoming paper in AIES 2025, we empirically evaluate the reliability of leading LLMs (e.g., Llama 3 (AI@Meta 2024), and

DeepSeek (DeepSeek-AI 2024)) for homelessness resource allocation, a high-stakes domain where frontline caseworker judgment is paramount. To simulate a realistic "off-the-shelf" use case, we applied zero-shot prompting to the models, tasking them with prioritizing clients from a real-world dataset from St. Louis's Homeless Management Information System. We then compared the LLM-generated rankings against established vulnerability scoring tools (e.g., VI-SPDAT (OrgCode Consulting Inc. and Community Solutions 2015)) and the decisions made by human caseworkers.

Results show a severe lack of consistency: model outputs varied widely across runs and models, with very low rank correlation (Spearman's ρ (Spearman 1904)). LLM rankings further diverged from both bureaucratic benchmarks and expert judgments, despite superficial agreement in controlled pairwise tests. These findings provide strong evidence that off-the-shelf LLMs are not yet suitable for autonomous deployment in systems requiring robustness, consistency, and value alignment.

Discretionary Trees (Pokharel, Das, and Fowler 2024): Capturing Human Discretionary Judgment To understand the human expertise that current AI models lack, our "Discretionary Trees" project computationally quantifies the sophisticated judgment of caseworkers. Also using data from St. Louis at a time when homelessness resource assignment was not strictly formulaic, we developed an interpretable machine learning framework to model caseworker deviations from rule-based scoring. By first learning simple policy heuristics as short decision trees (using SER-DT (Souza et al. 2022))and then isolating discretionary decisions (i.e., cases the simple rules get wrong), we find that human caseworkers apply non-formulaic judgment systematically, not randomly. Crucially, our findings showed that caseworkers selectively override baseline rules for households they deem less vulnerable, and these discretionary overrides yield significantly greater marginal benefits for clients than random deviations. This finding underscores the subtle, value-rich expertise embedded in human discretion, a capability that purely automated systems might overlook, and which is essential for effective and equitable outcomes.

EvoFair: Evolution of Meritocracy (Ongoing) This project extends beyond static evaluations to investigate a core tension in algorithmic fairness: how processes that are 'fair' at the individual level can generate and entrench inequality at the group level over time. By modeling repeated, merit-based selections (e.g., college admissions) with stochastic growth and admission thresholds, preliminary results show that even symmetric, individually fair, selection rules generate persistent inequality at the group level when small random advantages compound over time. We are deriving theory to formalize the conditions that produce this drift and to design policies that can proactively counteract it.

Contributions and Significance

The work that I have done so far contributes to AI, ethics, and societal decision-making in three principal ways. First, it **provides empirical evidence of the ethical risks** of current LLMs in high-stakes allocation, grounding abstract alignment concerns in a concrete social domain and calling for caution before deploying these models in autonomous roles. Second, it **underscores the value of human discretion** by computationally demonstrating how street-level bureaucrats use contextual knowledge to improve outcomes, arguing for human-in-the-loop systems over full automation. Third, it **develops a theoretical framework for dynamic fairness**, revealing mechanisms by which static fairness criteria can generate unintended long-term inequities and suggesting interventions to foster robust and sustained equity.

Future Research Directions

Future research will build on these studies. First, as a direct follow-up to "Discretionary Trees," I am developing a mathematical model to formalize the principles of caseworker discretionary judgment, moving from identifying that discretion exists to understanding its operation. Concurrently, to address the reliability issues identified in my "Street-Level AI" paper, I am designing experiments to evaluate whether a "committee of experts" approach, using ensembles of LLMs, can produce more robust and consistent decisions than a single model. Furthermore, I am in the preliminary stages of a theoretical project investigating how different forms of uncertainty in automated systems propagate to affect decision outcomes. *Critically*, the most important part of my research agenda is to tie these technical threads together through direct stakeholder engagement. This involves initiating a series of workshops with caseworkers, policymakers, and community advocates, not simply to present findings, but to collaboratively define what 'fair' and 'effective' AI assistance means within their institutional contexts. The goal is to develop frameworks for human-AI collaboration that are not just technically robust but are grounded in and accountable to the communities they serve.

Conclusion

From the simple preference for a human bank teller to the complex discretion of a social caseworker, my research underscores a unifying theme: the critical need for systems that can *respond with care*. My work demonstrates that current AI systems often fail on this front, exhibiting technical brittleness and a disconnect from established social values. In contrast, human experts consistently apply nuanced, context-aware judgment that improves outcomes. The challenge, therefore, is not to build a more perfect automaton, but to design tools that augment human expertise. Through such principled hybrid designs, we can leverage the strengths of AI without sacrificing the human-centered values essential for a just and equitable society.

Acknowledgments

This work was supported by the NSF (1939677, 2127752, 2533162) and Amazon through an NSF FAI award. We thank our community partners for defining challenges in homeless service delivery and supporting local families.

References

AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Bender, E. M.; and Hanna, A. 2025. On the Very Real Dangers of the Artificial Intelligence Hype Machine. *Literary Hub*.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. 24(1).

DeepSeek-AI. 2024. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. Cambridge Massachusetts: ACM. ISBN 9781450311151.

Gohar, U.; Tang, Z.; Wang, J.; Zhang, K.; Spirtes, P. L.; Liu, Y.; and Cheng, L. 2025. Long-Term Fairness Inquiries and Pursuits in Machine Learning: A Survey of Notions, Methods, and Challenges. (arXiv:2406.06736). ArXiv:2406.06736.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. NIPS'16, 3323–3331. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.

Maitra, S.; Sleep, L.; Henman, P.; Fay, S.; and Conversation, T. 2025. AI is being used in social services—but we must make sure it doesn't traumatize clients. *Phys Org*.

OrgCode Consulting Inc.; and Community Solutions. 2015. Vulnerability Index–Service Prioritization Decision Assistance Tool (VI-SPDAT): Prescreen Triage Tool for Single Adults. https://everyonehome.org/wp-content/uploads/2016/02/VI-SPDAT-2.0-Single-Adults.pdf. Accessed May 17, 2025.

Pokharel, G.; Das, S.; and Fowler, P. 2024. Discretionary Trees: Understanding Street-Level Bureaucracy via Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22303–22312.

Pokharel, G.; Farabi, S.; Fowler, P. J.; and Das, S. 2025. Street-Level AI: Are Large Language Models Ready for Real-World Judgments? (arXiv:2508.08193). ArXiv:2508.08193.

Souza, V. F.; Cicalese, F.; Laber, E.; and Molinaro, M. 2022. Decision Trees with Short Explainable Rules. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 12365–12379. Curran Associates, Inc.

Spearman, C. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1): 72.

Taylor, J. 2024. AI ban ordered after child protection worker used ChatGPT in Victorian court case. *The Guardian*.