Ethical Decision-making with AI: Value Alignment and the Role of Reasoning

Samarth Khanna

The Pennsylvania State University
samarth.khanna@psu.edu

Abstract

My research investigates how to evaluate and enhance large language models' (LLMs) alignment with human values in collective decision-making scenarios. I focus on three interrelated aspects of this challenge: (i) normative alignment, (ii) procedural competence, and (iii) personalization.

Introduction

The rapid integration of large language models (LLMs) into decision-making workflows raises pressing questions about their ethical reliability. From medical triage to resource allocation and hiring, these systems increasingly make, or strongly influence, choices with real-world consequences. While LLMs excel at pattern recognition and natural language reasoning, their ability to reflect human moral judgments and follow principled procedures remains uncertain.

Ethical decision-making requires more than surface-level accuracy. It requires (i) *normative alignment*, i.e. capturing fairness norms and moral intuitions endorsed by society, and (ii) *procedural competence*, i.e. the ability to apply transparent, principled rules for aggregating preferences. Misalignment in either dimension undermines trust, accountability, and fairness. A further requirement is (iii) *personalization*, since moral reasoning varies across individuals and contexts. AI systems must not only reflect majority judgments but also capture individual-level diversity to ensure both personalized and collective ethical reliability.

Normative Alignment

This aspect of my work focuses on both evaluating and enhancing human-AI alignment in terms of the **principles** and **rules** that govern decision-making, with an emphasis on **resource allocation**.

Distributive Fairness

Resource allocation provides a natural testbed for normative alignment, where philosophy, economics, and computer science converge. Outcomes are judged by fairness axioms such as **envy-freeness** (Foley 1966) and **equitability** (Dubins and Spanier 1961), that are often conflicting and computationally intractable. Additionally, human perceived fairness depends on various contextual factors (Konow 2003). A central question is whether LLMs reflect the preferences of humans when faced with the same fairness trade-offs.

LLMs as Social Planners. In (Hosseini and Khanna 2025), we evaluate LLMs in the role of centralized decision-makers allocating goods among individuals with known valuations. We find that LLMs rarely generate equitable (perfectly equal) allocations, in contrast to humans who frequently prioritize them (Herreiner and Puppe 2007). Interestingly, when selecting among pre-computed allocations, LLMs disproportionately favor equitable outcomes, revealing a *value–action gap*. Closer analysis suggests difficulty in computing equitability: models struggle when explicitly instructed to produce such allocations, though they succeed more often with envy-freeness or welfare maximization. Sensitivity to semantic and non-semantic prompt variations further highlights limitations in their distributive reasoning.

LLMs as Recipients. In (Hosseini et al. 2025), we examine fairness from the recipient's perspective. In a human study, participants judged whether their assigned bundle was "fair," under conditions varying in information availability, fairness notion satisfied, and framing. Replicating these experiments with LLMs, we find that models set a higher threshold for fairness, requiring a larger share to classify allocations as fair. They also show a stronger preference for envy-freeness compared to humans. Fine-tuning improves alignment with majority human responses but fails to enhance an understanding of nuanced fairness preferences.

Future work: Reinforcement learning for fairness. Reinforcement learning from human feedback (RLHF) has proven effective for steering models (Ouyang et al. 2022). I propose the use of reward models that explicitly encode fairness criteria, with resource-efficient methods such as GRPO (Shao et al. 2024), to (i) enforce specific notions (e.g., equitability), (ii) balance multiple fairness objectives, and (iii) improve alignment on expanded fairness benchmarks.

Moral Value Alignment

I am also studying alignment in morally charged, high-stakes contexts, such as organ donation.

Kidney Allocation. In (Dickerson et al. 2025), we examined decisions about allocating a single kidney between two patients varying in age, health, drinking habits, and number of dependents. LLMs align with humans on age and drinking habits but display inconsistency on dependents and diverge in prioritizing attributes. Unlike humans, they express indecision far less often, choosing decisively even in morally ambiguous cases. Fine-tuning improves aggregate alignment but fails to capture various nuances of human preferences.

Future work: Responsibility. Ongoing work evaluates whether LLMs align with human reasoning while attributing *responsibility* to harmful behaviors (smoking, drinking), disease occurrence, or deprivation of medical resources, drawing on findings from Chan et al. (2024).

Procedural Competence

Even if values are aligned, outcomes depend on faithfully applying aggregation procedures. We ask whether LLMs can reason with **ranked preferences** and **execute established algorithms** to reach desirable solutions.

Matching Markets. In (Hosseini, Khanna, and Singh 2025), we focus on the domain of two-sided matching markets. We find that performance, in terms of generating, improving, and evaluating solutions, degrades rapidly with input size, largely due to hallucinations when reasoning about preferences. Fine-tuning on synthetic reasoning traces improves accuracy for smaller instances but failed to generalize to larger cases. This highlights a gap between local pattern learning and scalable procedural reasoning.

Future work: Preference reasoning benchmark. I am extending this evaluation to domains such as voting (single- and multi-winner), participatory budgeting, resource allocation (with or without endowments), and coalition formation (hedonic games). Additional tasks include counterfactual and strategic reasoning, as well as generating explanatory examples. These benchmarks will include more realistic preferences such as those with ties and incompleteness, enabling a more exhaustive evaluation of algorithmic fidelity.

Personalization

Moral judgments vary across individuals and contexts. Beyond majority alignment, AI systems must capture individual **moral diversity** to provide both personalized ethical advice and richer collective decisions.

Future work: Preference elicitation from noisy responses. Human preferences are often inconsistent or costly to elicit. Building on recent works using LLMs for preference elicitation (Soumalias et al. 2025), I plan to investigate whether LLMs can (i) detect inconsistencies in natural language preference reports and (ii) ask clarifying questions to recover stable underlying values.

Future work: Personalized ethical advice. Recent work shows that people often rate LLM-generated moral advice as superior to professional ethicists (Howe et al. 2023). However, such advice is generic and not tailored to individual moral values. I propose leveraging LLMs' linguistic capabilities to elicit individuals' moral values and provide *personalized* advice aligned with their own ethical compass.

Conclusion

By integrating insights from computer science, economics, and philosophy, this research aims to inform the design of AI systems that are not only capable and efficient, but also ethically reliable, procedurally sound, and responsive to human moral diversity.

References

Chan, L.; Sinnott-Armstrong, W.; Borg, J. S.; and Conitzer, V. 2024. Should Responsibility Affect Who Gets the Kidney? In Davies, B.; Marco, G. D.; Levy, N.; and Savulescu, J., eds., *Responsibility and Healthcare*, 35–60. Oxford University Press USA.

Dickerson, J. P.; Hosseini, H.; Khanna, S.; and Pierce, L. 2025. Who Gets the Kidney? Human-AI Alignment, Indecision, and Moral Values. *arXiv preprint arXiv:2506.00079*.

Dubins, L. E.; and Spanier, E. H. 1961. How to Cut A Cake Fairly. *The American Mathematical Monthly*, 68(1): 1–17.

Foley, D. K. 1966. *Resource allocation and the public sector*. Yale University.

Herreiner, D.; and Puppe, C. 2007. Distributing Indivisible Goods Fairly: Evidence from a Questionnaire Study. *Analyse & Kritik*, 29(2): 235–258.

Hosseini, H.; Kavner, J.; Khanna, S.; Sikdar, S.; and Xia, L. 2025. Bridging Theory and Perception in Fair Division: A Study on Comparative and Fair Share Notions. *arXiv preprint arXiv:2505.10433*.

Hosseini, H.; and Khanna, S. 2025. Distributive Fairness in Large Language Models: Evaluating Alignment with Human Values. *arXiv preprint arXiv:2502.00313*.

Hosseini, H.; Khanna, S.; and Singh, R. 2025. Matching Markets Meet LLMs: Algorithmic Reasoning with Ranked Preferences. *CoRR*, abs/2506.04478.

Howe, P. D. L.; Fay, N.; Saletta, M.; and Hovy, E. 2023. ChatGPT's advice is perceived as better than that of professional advice columnists. *Frontiers in Psychology*, Volume 14 - 2023.

Konow, J. 2003. Which Is the Fairest One of All? A Positive Analysis of Justice Theories. *Journal of Economic Literature*, 41(4): 1188–1239.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Soumalias, E.; Jiang, Y.; Zhu, K.; Curry, M.; Seuken, S.; and Parkes, D. C. 2025. LLM-Powered Preference Elicitation in Combinatorial Assignment. *arXiv preprint arXiv:2502.10308*.