Jiashi Gao | Research Statement

My research interests lie at the intersection of computer science, economics, and sociology, focusing on fairness in AI- and human-involved systems—including data markets, federated learning, and human–AI collaboration. I investigate fairness-related issues arising from human factors such as heterogeneous data resources, behaviors, cognitive capacities, and demographic characteristics. My academic goal is to empower AI systems to be trustworthy, efficient, and fair in practice.

Artificial Intelligence (AI) has become an integral part of our daily lives, providing efficient, large-scale support for decision-making across various domains, such as medical diagnoses, financial risk assessments, and sentencing judgments. Given its close relation to human well-being, it is imperative to examine whether AI acts as an ethical assistant; for example, will AI assistance exacerbate information inequality among diverse human demographics—particularly those with varying expertise due to unequal access to advanced knowledge? Could AI systems propagate subtle biases and mislead humans into making unfair decisions?

During my PhD studies, I actively engaged in examining and solving fairness-related issues in systems involving both humans and AI, including but not limited to data markets [1], collaborative model training represented by federated learning [2, 3, 4], and human-AI collaboration [5]. Each scenario presents distinct challenges to fairness across different dimensions. Specifically, ① Budget-constrained consumers seek to purchase datasets that most effectively enhance fairness in AI model training, yet the ability of datasets to improve fairness before purchase remains unknown [1]; ② There is a fundamental trade-off between allocating model performance fairly among participants and achieving optimal overall model performance [2]; ③ Fair model performance allocation can disadvantage advantaged participants, prompting them to withdraw and potentially causing the collaborative training system to be unstable [3]; ④ Widely used fairness-aware mechanisms in collaborative model training can be stealthily bypassed by malicious participants, undermining fairness without reducing accuracy [5]; ⑤ Humans' decision-making gains from AI assistance are highly sensitive to humans' cognitive capacity, resulting in unfairness issues in AI-assisted decision-making systems [4].

Below, I describe my representative work and future research plan.

Data Acquisition Strategy for Fair Model Training

The emergence of data markets (e.g., Dawex, Bloomberg, LexisNexis) has provided researchers with access to high-quality, non-public datasets, offering a critical resource for training domain-specific AI models. However, datasets from different providers may carry biases toward specific populations (e.g., race, gender, age, marital status, LGBTQ+).

Biased datasets, if acquired by data consumers for AI model training, may result in unexpectedly unfair decisions in their downstream tasks. Drawing upon my early experience in the digital assets market [6, 7], I address this issue at its root by exploring how rational consumers can acquire unbiased yet high-quality data effectively. Designing data acquisition strategies within the context of data markets is particularly challenging due to the lack of transparency: consumers cannot inspect data before purchase to evaluate potential biases. Acquiring and assessing all available datasets is impractical, especially as consumers are often constrained by limited budgets. My solution is a **knowledge-mining and knowledge-utilization** data acquisition strategy. In the knowledge-mining phase, I derive the minimum sample size required so that the subset can approximately estimate the fairness (using statistical disparity metrics) and accuracy of any unseen dataset. Building upon the estimated knowledge of unseen datasets, I address the unbiased data acquisition problem under a limited budget as a variant of the nonlinear knapsack problem (NLKP), with the solution being both natural and interpretable.

This work takes a first step toward supporting fair AI development through rational data acquisition in data markets and has been published in *AAAI/ACM Conference on AI, Ethics, and Society (AIES-24)*.

Fairness in Collaborative Model Training

Promoting fairness for humans—especially for disadvantaged populations—has always been the central goal and meaningful purpose of my work. In the context of collaborative model training, i.e., federated learning, disadvantaged populations refer to those who, due to historical inequalities in accessing resources, lack sufficient data to effectively engage with AI systems. Under the influence of the **Matthew effect**, a socio-economic concept highlighting how the rich become richer and the poor become poorer, a similar phenomenon occurs in collaborative model training. Participants with higher data contributions tend to dominate the training process, leaving populations with insufficient data resources unable to achieve high performance from the collaborative model, especially when their data is heterogeneous compared to that of high-contribution participants.

When AI is collaboratively trained for welfare applications, such as medical diagnosis or determining social welfare eligibility, this Matthew effect requires careful attention and to be efficiently mitigated. To this end, I proposed a novel fairness principle, Anti-Matthew fairness, which ensures that the model delivers equal performance—both in terms of accuracy and bias levels—across clients. To operationalize this principle, I formulated the training objective of *Anti-Matthew federated learning* as a multi-constrained multi-objective optimization (MCMOO) problem. The conflicts among multi-objectives make it non-trivial to obtain optimal solutions. My solution is a three-stage multi-gradient descent algorithm that enables AI model training to converge to Pareto-optimal solutions for the MCMOO problem, where the global model's performance on each objective is maximized and cannot be further improved without compromising other objectives. **My efforts provide the first efficient solution to the previously overlooked Matthew effect in collaborative model training. This work has been accepted by *European Conference on Artificial Intelligence (ECAI-24)*.**

A critical challenge in deploying fairness-oriented collaborative model training systems lies in balancing fairness and stability. Pursuing fairness often comes at the cost of model performance for data-rich participants, which may incentivize them to leave the grand coalition and form sub-coalitions better aligned with their performance goals. Despite its significance, this challenge has remained unexplored. To address this gap, I framed the question of how fairness and stability interact in such systems. My work models participants' inclination to leave based on human factors—such as altruistic behaviors and social connections—and examines their impact on the trade-off between fairness and stability. I derived tight theoretical bounds on achievable fairness levels across diverse participant scenarios, offering valuable insights for setting appropriate fairness thresholds to maintain system stability in real-world applications. **This work has been recognized and accepted at *Annual Conference on Neural Information Processing Systems (NeurIPS-24)*.**

Unique Fairness Challenges in Human-AI Collaboration

In recent work on AI-assisted decision-making, I explored a novel fairness challenge arising from heterogeneity in human cognitive capacity, driven by differences in expertise and access to knowledge shaped by diverse social backgrounds. As a result, humans exhibit varying degrees of alignment between their confidence in making positive decisions and the actual likelihood of outcomes based on observed data. Existing AI confidence reporting mechanisms typically fail to account for this heterogeneity, leading to disparities in decision-making utility—measured by accuracy—among humans assisted by AI. To address this issue, I propose a new confidence adjustment criterion, Inter-Group Alignment, which ensures both optimal and equitable utility for humans with diverse cognitive capacities. The proposed approach is model-agnostic and broadly applicable across AI systems. This work is currently in draft form and under review.

Future Research Agenda

Beyond fairness, ethical concerns in human-AI systems—such as accountability, value misalignment, and norm conflicts—require urgent attention, especially as AI technologies are increasingly integrated into real-world collaborative applications like retrieval and recommendation systems. The key obstacles to improving AI models themselves to address ethical concerns have become increasingly evident. Typical challenges include the limited representativeness of ethically-relevant datasets and the inherent trade-offs among competing performance objectives. An alternative perspective, inspired by human factors engineering, emphasizes designing systems that accommodate human limitations rather than expecting humans to overcome their inherent cognitive and behavioral constraints—an expectation that is often unrealistic. Fortunately, across many dimensions of limitation, humans and AI tend to be complementary, presenting a promising opportunity to design interaction mechanisms that mitigate each of their weaknesses. To advance human-AI complementarity in HCI, I identify the following research questions—spanning the entire HCI pipeline from human input through model processing to model output—as urgent areas for investigation:

1. Input optimization for AI processing constraint: How does the volume and nature of information provided by humans affect the AI decision-making process? Mismatches between information complexity and AI processing capabilities might negatively impact decision quality.

2. Adaptive AI responses under human heterogeneity: How can AI leverage information from human interactions to more precisely recognize and adapt to inherent human cognitive limitations or decision-making biases? Understanding individual differences in biases and cognitive constraints can enable AI systems to deliver personalized feedback, effectively mitigating these limitations.

Table 1: Difference of Humans Factors and AI Factors in HCI

Dimension	Human	AI (i.e., LLM)
Generation from Observation	Individualized subjective perception	Probabilistic sampling from learned distributions
Learning Paradigm	Interaction-based experiential learning	Data-driven statistical modeling
Emotional Expression	Genuine, diverse	Simulated emotion based on context
Creativity	Unbounded, experiential innovation	Data-bounded combinatorial output
Moral Agency	Non-enforceable, intrinsic, value-based	Enforceable, rule-driven
Information Processing	Semantic understanding, constrained by attention span and cognitive load, factually grounded	High-throughput, lacks semantic understanding, hallucination
Bias	Subject to cognitive biases, individually variable	Capable of bias detection but lacks lived context
Communication Style	Implicit, nuanced, and socially calibrated	Explicit and formal, opaque without interpretability

3. Output optimization for human perception and trust: In what ways should AI systems present their outputs or explanations to humans to enhance the perceived accuracy and fairness for humans? Effective communication strategies should not only ensure AI decisions themselves are accurate and ethically sound but also foster human confidence and clarity regarding AI-generated outcomes.

Given the outlined problems, the following technical approaches warrant further investigation.

Causal Reasoning in HCI

Enhancing the synergy between humans and AI models greatly benefits from enabling them to actively learn from their past interactions and improve their modes of interaction—a process fundamentally rooted in causal reasoning. Repeated interactions between humans and AI systems allow for deeper insights into the causal effects of various historical interaction strategies, which may potentially reflect human- and model-related factors across different contexts and scenarios. Answering counterfactual questions offers a valuable approach to exploring these causal relationships—for example: How would collaborative outcomes have changed if a different prompt or response structure had been used? Addressing such counterfactual questions provides important insights, helping to identify prompt or response constructions that consistently enhance HCI, as well as those whose effectiveness varies depending on specific contexts.

Game Theory for Explainability

Depending on the context, HCI can be modeled as either a non-cooperative or cooperative game. For instance, AI-driven job application filtering—where applicants attempt to "game" the algorithm through certain tricks—represents a non-cooperative interaction between humans and AI. In contrast, AI-assisted medical diagnosis illustrates a cooperative scenario where humans and AI systems collaborate to achieve a shared goal. Game theory provides a structured theoretical framework for modeling the HCI process, enabling analysis of the existence of Nash equilibrium, coalition formation, and the fair distribution of outcomes, etc. Moreover, considering the inherently cooperative nature of combining prompt segments—each contributing collaboratively to the AI's output—cooperative game theory concepts such as the Shapley value offer an interpretable method to quantitatively evaluate the contribution of each prompt segment to the final result. This approach may offer a viable pathway to address human limitations in crafting optimal prompts, striking a balance between the complexity (length) of input prompts and the AI's processing capabilities. It also helps reduce input-induced biases and mitigate hallucinations, which tend to worsen with longer input lengths due to the probabilistic nature of AI content generation. Research applying game theory to HCI is still in its early stages, offering significant opportunities for further exploration and development.

References

[1] Jiashi Gao, Ziwei Wang, Xiangyu Zhao, Xin Yao, and Xuetao Wei. Surviving in diverse biases: Unbiased dataset acquisition in online data market for fair model training. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):451--462, Oct. 2024.

[2] Jiashi Gao, Ziwei Wang, Xiangyu Zhao, Xin Yao, and Xuetao Wei. Does egalitarian fairness lead to instability? the fairness bounds in stable federated learning under altruistic behaviors. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[3] Jiashi Gao, Xin Yao, and Xuetao Wei. Anti-matthew fl: Bridging the performance gap in federated learning to counteract the matthew effect. In *ECAI 2024*, pages 1967--1974. IOS Press, 2024.

[4] Jiashi Gao, Ziwei Wang, Xiangyu Zhao, Xin Yao, and Xuetao Wei. Pfattack: Stealthy attack bypassing group fairness in federated learning. *arXiv preprint arXiv:2410.06509*, 2024.

[5] Jiashi Gao, Kexin Liu, Xinwei Guo, Junlei Zhou, Jiaxin Zhang, Xiangyu Zhao, Xin Yao, and Xuetao Wei. Human expertise really matters! mitigating unfair utility induced by heterogenous human expertise in ai-assisted decision-making. In *Under review*, 2024.

[6] Jiashi Gao, Ziwei Wang, and Xuetao Wei. An adaptive pricing framework for real-time ai model service exchange. *IEEE Transactions on Network Science and Engineering*, 11(5):5114--5129, 2024.

[7] Ziwei Wang, Jiashi Gao, and Xuetao Wei. Do nfts' owners really possess their assets? a first look at the nft-to-asset connection fragility. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2099–2109, New York, NY, USA, 2023. Association for Computing Machinery.