Research Overview

Modern artificial intelligence (AI) models have evolved into increasingly powerful and sophisticated systems with up to billions of parameters. However, existing commercial systems such as language models, robots, autonomous vehicles, and social media recommendation engines are still seen as immature because these AI systems do not fully act in accordance with human values, goals, and preferences. To efficiently and reliably align AI systems with human values, we must confront several key challenges:
(1) The values are taught by humans who make mistakes, harbor biases, and have complex, evolving values that are hard to completely specify. It remains unclear what is the most appropriate approach to aggregate human preferences and form an operable mathematical objective.
(2) Given the objective, the agent has to further interact with humans or a simulated/real environment to receive feedback and adapt itself. This has become increasingly costly due to the huge amount of computing and data queries.
(3) When the preferences and training algorithms are readily at hand, there are trustworthiness issues. On one hand, alignment requires a massive amount of private human data, and ensuring data security is a challenge to the data-collecting entities. On the other hand, how to utilize them robustly and safely to ensure alignment is unclear.

My research agenda revolves around these challenges and aims to develop efficient and trustworthy AI alignment approaches, that are motivated by real-world applications, yield new theoretical insights, and demonstrate tangible practical impacts. In pursuing principled ways to use data from humans for AI alignment, I have studied the three interconnected problems stated above, which can be broadly categorized into:

(1) **Preference-based Learning** to extract human preference and values.
(2) **Reinforcement Learning** for efficient AI training and alignment.
(3) **Trustworthy Machine Learning** to ensuring system privacy and robustness.

Research Accomplishment

Preference-based Learning

Learning from preference-based feedback has been one central problem across different fields such as ranking, recommendation systems, and social choice theory. Recently, reinforcement learning from human feedback (RLHF) has also shown its strong potential in utilizing weakly supervised human data (preference-based feedback) and its ability to accurately encode human values into generative models, particularly in the area of large language models. **My Ph.D. dissertation** aims to develop a comprehensive characterization of preference-based learning [1–4], focusing on the sample complexity of ranking and preference model estimation.

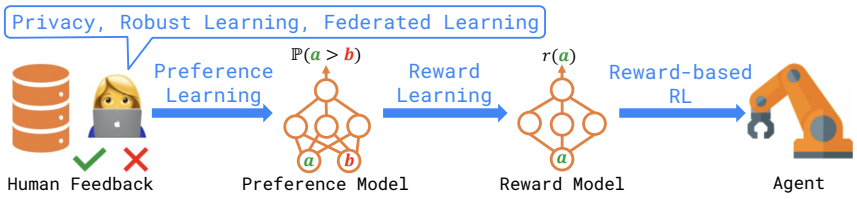

Figure 1: The overview of my research. AI Alignment (RLHF) consists of 1) preference learning, 2) reward learning, and 3) reward-based Reinforcement Learning. My research explores the possibilities of these directions and also considers the trustworthiness of AI systems due to the sensitive nature of human data.

A typical way of modeling human preference is to assign a "score" for each option, and the comparison between two options is determined by the difference of the scores [5]. This assumption can rarely hold in real applications. To this end, I explored methods for better characterization of human preferences. In [1], I studied learning to rank under the strong stochastic transitivity condition, a prevalent model without

assuming a score for each option. I proposed one of the first adaptive approaches that can effectively aggregate the feedback from different human labelers and illustrated how the relationship between the number of human queries and resulting performance depends on the properties of the human labelers. Sometimes, a close pair in the overall ranking is not necessarily harder to compare than a pair that has a large gap. One experiment on human behavior of gambling [6] shows that "people chose between adjacent gambles according to the payoff and between the more extreme gambles according to probability or expected value". In a follow-up work [2], my collaborators and I developed novel algorithms under this practical yet harder setting. Our efficient algorithm requires (up to 50%) fewer human queries compared with algorithms designed with stronger assumptions, and is provably optimal. This line of work significantly promotes our understanding of what is the least amount of human data necessary to recover the true ranking, which answers one of the central problems in recommendation systems.

In my exploration of rich and general preference models, I found constantly that human rarely makes consistent comparisons. That means humans often demonstrate contradicting preferences such as a loop within the preference relations. In [3], I considered the most general setting where the preference is determined by some rich context, and proposed algorithms that identify the Borda winner, an optimal choice even when a true underlying rank list does not exist. I showed the algorithm enjoys minimum regret, a notion that trades between exploration and exploitation. This result sheds light on the fundamental difficulty and cost of recovering human preferences under few assumptions. Working with my collaborators [4], I further our understanding of contextual preference learning by considering the variance of the comparison, and establishing the first variance-aware regret-minimization algorithms for preference-based bandits, a problem that has since received renewed interest in the theoretical machine learning community.

Efficient Reinforcement Learning Many real-world applications require grappling with the effects of dynamic feedback and systems that change over time. In these settings, there is no clear and unanimously agreed criterion to guide the agent, and the agent has to make decisions sequentially and adaptively. AI designers typically provide an objective function or reward, and Reinforcement Learning (RL) algorithms step in as a standard tool to maximize this objective function designated by humans. **My research aims to understand the core principles that govern the training process of RL algorithms, enabling them to efficiently make decisions in complex environments.** My research spans several aspects of RL: **model-based RL** and **policy-based RL**.

Model-based RL refers to those algorithms that attempt to estimate the environment by a parametric model. In [7], I for the first time studied RL with linear function approximation in the infinite-horizon average-reward setting. My proposed algorithm can fully utilize the transition structure and is efficient as I proved the regret of the algorithm enjoys optimal regret upper bound. In [8], I delved deeper into RL with general function approximation. I adopted the notion of the eluder dimension, which allows non-linear function approximation, and proposed new algorithms that are based on the multi-level partition scheme, which allocates samples with different uncertainty into different subsets. This algorithmic design allows the algorithm to directly converge to the optimal policy, with a strong guarantee that ensures both small regret and small sample complexity.

Policy-based RL refers to those algorithms that represent the policy by a parametric model and optimize directly over the parameters. Among these algorithms, the actor-critic (AC) method is one of the most widely used algorithms in practice. In the AC algorithm, the actor uses the policy gradient to improve the learning policy while the critic uses temporal difference (TD) with linear function approximation to evaluate the goodness of the policy gradient. In [9], I initiated and led the establishment of the first finite-time analysis of two time-scale actor-critic methods with linear function approximation. This significantly improves our understanding of how the actor-critic method works and justifies its great empirical success in reinforcement learning.

Trustworthy & Collaborative Machine Learning Modern AI systems do not serve individual users in a vacuum but rather must provide service simultaneously for large populations of users. Effective and robust learning approaches should have both the flexibility to model and personalize to individual users. My work has probed the boundaries of collaborative learning from large populations of users with an emphasis on the trustworthiness of AI systems, such as **privacy, robustness, and federated learning**.

For data safety, I studied federated learning and proposed algorithms [10] that enable multiple entities to build effective and personalized ML models without directly sharing data, thus ensuring data privacy and security. The algorithm for the first time explicitly handles the covariate shift (change in input distribution), provably guarantees personalization, and outperforms all state-of-the-art algorithms.

Another collaboration with practitioners leads to [11], which studies how to enable large language models to mask sensitive personal information in a principled way. I provided a principled information-theoretical analysis of the different ways of masking sensitive information and guiding the training of large language models. This work shows that LLMs can be good privacy protection learners, without the need for balancing a privacy-utility trade-off.

Some data sources are informative, others are noisy, and some are even malicious. It is important for the algorithms to adaptively query these heterogeneous data sources and maintain robust performance at the same time. In [1], I looked into robust ML in the context of preference learning, proposing algorithms that adaptively utilize multiple sources of preference feedback by filtering out noisy or adversarial sources. The number of human queries can be reduced up to 80% compared with the baselines.

Understanding Deep Learning The deep neural network constitutes the bedrock of AI advancements. During my Ph.D. study, I have also devoted my efforts to collaborations that explore the intricacies of deep neural networks. This includes a series of collaborations on deep learning theory, such as defining new notion of representational similarity [12], explaining low-frequency inductive bias [13], and understanding the mixture of experts structure [14].

Future Directions

As discussed in my research accomplishment, a central theme of my research focuses on developing interactive approaches that extract human concepts such as preferences from actively queried human feedback. Such approaches can utilize human feedback better to improve system utility for a variety of applications, ranging from helping domain experts model complex phenomena to personalizing AI applications for large populations of users.

Efficient and Rich Preference-based Reinforcement Learning The highest reward is not always synonymous with the best outcome. Identifying the most appropriate criteria to measure AI alignment and guide reinforcement learning under various settings remains an open problem.

Recent RLHF methods for large language models take a two-step approach: they first learn a reward model from the given preference dataset and then run off-the-shelf reinforcement learning algorithms on top of the learned reward model. However, acquiring an accurate reward model only from preference labels, typically provided by human labelers, poses a significant challenge as it is unclear whether the reward model can unbiasedly estimate human preferences.

Alternatively, predicting the preference itself can more accurately reflect human preference as the preference model allows inconsistent comparisons, a phenomenon observed and addressed in my previous works [2, 3]. In terms of training, learning the preference is also comparatively more straightforward since we have direct access to training labels, allowing us to leverage powerful techniques from supervised learning. Building upon this observation, I propose to design RL algorithms that bypass the need for reward function modeling

by directly learning from preference labels. The expected solution should be computationally efficient as it skips the intermediate procedure of reward learning.

Moving forward, I am interested in developing efficient reinforcement learning algorithms that provably learn optimal policies under diverse utility criteria, including preference-induced objectives and multi-task objectives.

Collaborative & Trustworthy Learning for AI Alignment AI systems are not aligned with a homogeneous group of people but rather must aggregate for large populations from heterogeneous backgrounds and with different values. Effective and fair alignment approaches should have both the flexibility to fairly represent values and preferences from different people, as well as the ability to intelligently balance the exploration/exploitation tradeoff to express overall consensus from the entire population of users. One preliminary example is my series of work on general preference models [1–3], which partially addresses these issues by establishing provably efficient convergence guarantees under various preference models.

Due to the sensitivity of personal preference data, collaborative learning naturally urges guarantees of data safety. I propose to continue my research in trustworthy ML by utilizing collaborative learning [1] and federated optimization [8] to guarantee the AI alignment is robust and private. I am also interested in developing both online and offline learning algorithms that can simultaneously leverage the shared structure among different groups of users while provably retaining the data efficiency of conventional learning methods for a single user group.

Alignment goes beyond an intellectually complicated challenge as it can be indeterminant to translate human values into mathematical objectives. I am also interested in collaborations with researchers from psychology or social science to initiate rich alignment criteria for the long-term interests that persist within populations of users.

Knowledge Incorporation with AI systems In tasks related to natural science, it is hard for human labelers to reliably rate the performance of an AI system. Even if the criterion is objective and static, human evaluation still requires a long time and expert knowledge. In my ongoing collaboration with Bytedance Research on the topics of data-driven protein design, I found that to evaluate the protein's structural plausibility, a chemical expert has to extensively read the visualized structure or even evaluate it in wet labs. I aim to develop general expert-in-the-loop frameworks to accelerate scientific discovery. More specifically, I aim to design sample-efficient methods that can align the model with only weakly supervised information from human experts. I also aim to design principled approaches to distill knowledge from the more readily available but less informative non-expert feedback, by the aggregation technique developed in my previous works [1, 3].

Aside from sample-efficient AI for scientific tasks, I am also interested in incorporating insights from natural science to develop more powerful AI models. One example is the score-based model I studied at Bytedance Research: it lifts from zeroth-order methods (normalizing flow, auto-encoders) where deep neural networks learn transformations from an object to another object, to first-order methods (diffusion models, flow matching) where networks learn the changing rate (velocity) of the object. This insight from dynamics effectively expands the flexibility and scalability of generative models. I am interested in incorporating similar physical insights that help scale up AI models.

References

[1] **Wu, Yue**, Tao Jin, Hao Lou, Pan Xu, Farzad Farnoud, and Quanquan Gu. Adaptive sampling for heterogeneous rank aggregation from noisy pairwise comparisons. In *International Conference on Artificial Intelligence and Statistics*, pages 11014–11036. PMLR, 2022.

[2] Hao Lou, Tao Jin, **Wu, Yue**, Pan Xu, Quanquan Gu, and Farzad Farnoud. Active ranking without strong stochastic transitivity. *Advances in neural information processing systems*, 35:297–309, 2022.

[3] **Wu, Yue**, Tao Jin, Hao Lou, Farzad Farnoud, and Quanquan Gu. Borda regret minimization for generalized linear dueling bandits. *arXiv preprint arXiv:2303.08816*, 2023.

[4] Qiwei Di, Tao Jin, **Wu, Yue**, Heyang Zhao, Farzad Farnoud, and Quanquan Gu. Variance-aware regret bounds for stochastic contextual dueling bandits. *arXiv preprint arXiv:2310.00968*, 2023.

[5] Ralph Allan Bradley and Milton E. Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 0006-3444. doi: 10.2307/2334029.

[6] Amos Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

[7] **Wu, Yue**, Dongruo Zhou, and Quanquan Gu. Nearly minimax optimal regret for learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3883–3913. PMLR, 2022.

[8] **Wu, Yue**, Shuaicheng Zhang, Wenchao Yu, Yanchi Liu, Quanquan Gu, Dawei Zhou, Haifeng Chen, and Wei Cheng. Personalized federated learning under mixture of distributions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37860–37879. PMLR, 23–29 Jul 2023.

[9] **Wu, Yue**, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628, 2020.

[10] **Wu, Yue**, Jiafan He, and Quanquan Gu. Uniform-PAC guarantees for model-based RL with bounded eluder dimension. In Robin J. Evans and Ilya Shpitser, editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR, 31 Jul–04 Aug 2023.

[11] Yijia Xiao, Yiqiao Jin, Yushi Bai, **Wu, Yue**, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Haifeng Chen, et al. Large language models can be good privacy protection learners. *arXiv preprint arXiv:2310.02469*, 2023.

[12] Liwei Wang, Lunjia Hu, Jiayuan Gu, Zhiqiang Hu, **Wu, Yue**, Kun He, and John Hopcroft. Towards understanding learning representations: To what extent do different neural networks learn the same representation. *Advances in neural information processing systems*, 31, 2018.

[13] Yuan Cao, Zhiying Fang, **Wu, Yue**, Ding-Xuan Zhou, and Quanquan Gu. Towards understanding the spectral bias of deep learning. In *IJCAI*, 2021.

[14] Zixiang Chen, Yihe Deng, **Wu, Yue**, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35:23049–23062, 2022.