RESEARCH STATEMENT

Xingyu Zhou

Motivation. Personalized AI/ML services are seamlessly integrated into every facet of our daily lives, from healthcare and education to commerce and large language models (e.g., ChatGPT). In this AI-driven era, individuals interact with intelligent systems that learn from their behaviors, preferences, and needs, striving to provide tailored and efficient solutions. However, this utopia of personalization brings with it substantial challenges of **trustworthiness**. Users are increasingly concerned about the *privacy* of their data, questioning how their information is used and who has access to it. Companies and service providers grapple with ensuring the *robustness* and *safety* of their systems, recognizing that even minor glitches can lead to significant consequences. Furthermore, the imperative for *fairness* cannot be overlooked, as these intelligent systems must serve diverse populations equitably, avoiding biases and ensuring equal opportunities for all.

Research Interest. Motivated by these challenges, my research centers around the following pivotal question:

*How to ensure **trustworthiness** in data-driven interactive decision-making?*

I tackle this question through a theoretical lens, striving to establish fundamental theoretical limits and uncover innovative algorithmic principles for trustworthy interactive decision-making. My exploration spans the mathematical frameworks of bandits and reinforcement learning (RL), with a goal to thoroughly understand and improve sample efficiency and algorithmic design for privacy, robustness, safety and fairness.

Diverging from the prevalent emphasis on trustworthy statistical learning (e.g., supervised learning), my research highlights that interactive decision-making presents distinct challenges and necessitates a shift in our approach to modeling, analysis, and algorithm development. This involves rethinking existing definitions and metrics, addressing unique challenges inherent to interactivity (e.g., exploitation-exploration trade-off), and examining a wider array of algorithmic solutions. In navigating these complexities, my present and forthcoming research efforts are dedicated to laying down the theoretical foundations for trustworthy interactive data-driven AI/ML systems for decision making, with a focus on ensuring data privacy, strengthening robustness, bolstering operational safety, and upholding fairness, as summarized in Fig. 1.

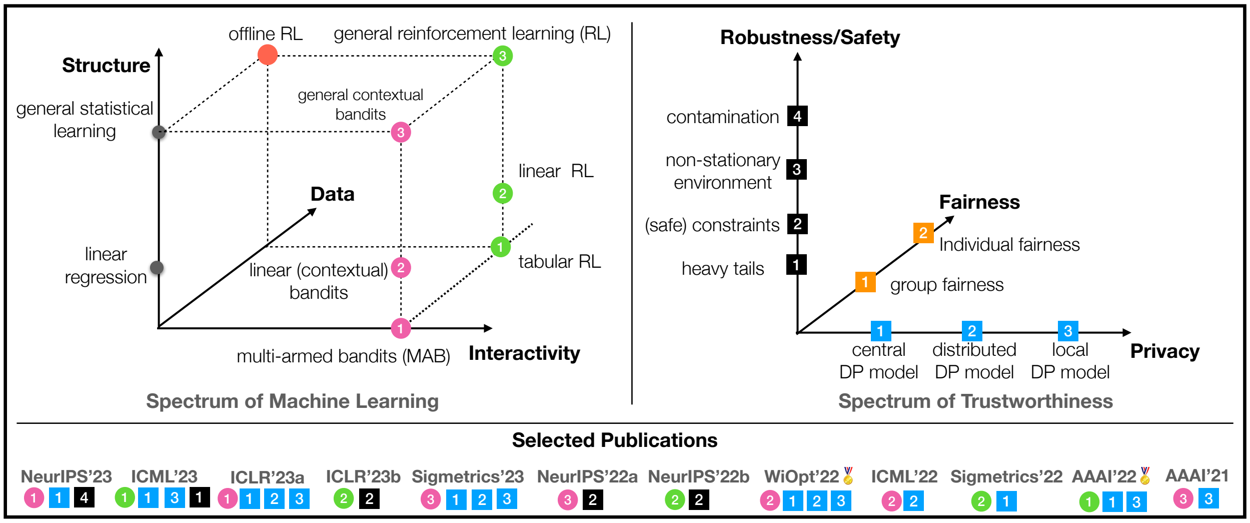

Figure 1: A landscape of my current and future research. **(left)** (i) Interactivity captures how the training data is obtained, e.g., passive vs. active; (ii) Structure captures different levels of function approximations; (iii) Data captures the complexity of observations, e.g., simple features/labels/rewards vs. full trajectory. **(right)** (i) Privacy under different differential privacy (DP) trust models; (ii) Robustness/safety from various aspects; (iii) Fairness via different notions. **(bottom)** Selected recent publications, including Best Student Paper of WiOpt'22 and Oral presentation of AAAI'22.

1 Differentially Private Bandits and RL

My research endeavors to unravel a pivotal question at the intersection of privacy and data-driven decision-making: *What is the inherent cost of privacy in terms of sample efficiency for bandit algorithms and reinforcement learning (RL)?* Specifically, I seek to quantify the additional interactions required to achieve comparable performance while ensuring rigorous privacy protections. To tackle this inquiry, I employ differential privacy (DP) [1]—a gold standard for privacy guarantees—as the foundational framework and measure sample efficiency through the lens of regret. The challenges unique to bandits and RL include establishing appropriate DP criteria and crafting private mechanisms tailored to various settings, each with its own interaction processes and sensitive data considerations. Through a systematic research strategy, I have derived a series of enlightening findings that span multiple domains—from Multi-Armed Bandits (MAB) and contextual bandits to linear RL, and across different privacy paradigms such as central DP, where the learner is a trusted entity, to local DP, where the learner is not trusted. Some representative results are listed below[1].

Contributions. My research initiates the exploration of local DP in MABs with the first optimal privacy cost characterization [2]. Following this, I advanced the field by achieving optimal regret in distributed DP, an intermediate model between central and local DP, and as a by-product provided a holistic analysis across all three DP models [3]. Beyond the realm of MABs, my work unifies the analysis within all three DP models in linear contextual bandits [4], and notably, establishes a "privacy-for-free" result for linear bandits [5]. Further, my exploration into non-linear bandits has led to the first substantial results in this complex domain [6, 7]. In collaborative multi-agent/federated scenarios, my recent work [8] identifies and rectifies foundational errors in the prevailing literature on federated linear bandits, thereby setting new standards for state-of-the-art results. In the field of RL, my contributions [9, 10] represent the first investigation into private policy optimization for both tabular and linear models, making the first step toward privatizing the most widely implemented methods in RL. These above achievements stem from novel algorithmic frameworks (e.g., distributed tree-based algorithm), and the inception of advanced technical tools (e.g., a new privacy amplification lemma).

Impacts. My research in private bandits has catalyzed subsequent studies in specialized sub-areas, including heavy-tailed [11], cascading [12], and multi-agent bandits [13], with practical implications for private dynamic pricing [14]. My established frameworks for private RL have been extended in subsequent works for both online and offline RL settings [15, 16]. This line of work has garnered recognition, including the Best Student Paper at WiOpt'22 [5] and an oral presentation at AAAI'22 [9], reflecting its scholarly impact in the field.

2 Robust Interactive Learning: Non-stationarity, Heavy-tails, and Constraints

My research aims to address a fundamental question in interactive decision-making: *Is efficient learning possible under one or more of the following robust conditions: non-stationary environments, heavy-tailed feedback, and unknown constraints?* This line of inquiry is crucial for autonomous personalized systems tasked with optimizing performance in the face of potentially time-varying and heavy-tailed feedback, all while complying with diverse constraints related to safety, fairness, and resource allocation—often without closed-form formulas. The core challenge here lies in the ability to adapt to volatile environments and heavy-tailed feedback, to learn reward structure and constraints concurrently, and to strategically balance these elements. My work provides an affirmative answer to the above question through a suite of algorithmic innovations and theoretical contributions. I have developed a collection of efficient algorithms that offer provable performance assurances, addressing both regret minimization and constraint adherence within both bandit and RL frameworks. The following summarizes the key contributions of my research.

Contributions. My research rigorously explores bandits with a general structure, specifically kernel bandits

[1] See also my blog post for summary: https://xingyuzhou.org/blog/notes/Differential-privacy-for-bandits-and-RL

(aka Gaussian process bandits or Bayesian optimization), which subsume linear bandits and MABs, as shown in Fig. 1. For environments with non-stationarity and potentially heavy-tailed rewards, I have developed a suite of algorithms that guarantee sublinear regret. These algorithms incorporate innovative exploration strategies such as restart mechanisms [17, 18], sliding windows [18], and weighted Gaussian processes for kernel bandits [19], alongside an adaptive medians-of-means technique tailored for heavy-tailed distributions [6, MoMA algorithm]. In the realm of constrained bandits, where the goal is to optimize an unknown reward function subject to unknown constraints, I have introduced a versatile primal-dual online learning framework [20]. This framework is compatible with a variety of exploration strategies, such as UCB and Thompson sampling, and it consistently achieves sublinear metrics for both regret and constraint violations. Transitioning to RL, a key contribution is the first model-free algorithm in linear Markov Decision Processes (MDPs) with large state spaces [21]. This work is further extended to infinite-horizon RL with unknown constraints [22], to non-stationary environments [23], and most recently, to constrained RL with more stringent constraint guarantees [24]. These results are underpinned by innovative analytical solutions and algorithmic techniques, such as a novel application of the covering number argument and the development of a softmax policy that balances the tradeoff between its approximation and smoothness properties.
Impacts. The algorithms I have developed for addressing non-stationarity and constraints within kernel bandits have set a new benchmark within the field, as evidenced by their adoption in subsequent research [25, 26, 27, 28]. These contributions have been recognized in a recent comprehensive survey [29]. Additionally, my co-authored work [17] has been honored with the Best Student Paper Runner-up award at WiOpt'22, underscoring the practical relevance and theoretical robustness of our approach. My research contributions to constrained RL also serve as a comparative standard across diverse RL research settings [30, 31, 32, 33].

3 Interplay between Robustness and Privacy

My research in this line aims to answer the following important question for interactive decision-making: *What is the interplay between robustness (in particular, adversary contamination) and differential privacy (DP)?* That is, instead of viewing robustness and privacy as two orthogonal aspects of trustworthiness, it is instructive to study their connection and interplay, as evidenced in recent studies on their interplay in fundamental statistical estimation [34, 35, 36]. To this end, I take the first step toward understanding the interplay between Huber contamination [37] and DP in bandits. It turns out that even for MABs, there is no simple answer to the above question and several interesting findings are revealed by my research.
Contributions. The first interesting finding is that for the central model of DP MABs, with a careful tweak of previous robust algorithms (i.e., truncation-based algorithm), one can simultaneously handle robustness (Huber contamination and heavy-tailed rewards) and privacy in an optimal way, as shown in my recent work [38]. However, when it comes to the local DP (LDP) model (where each device/user is responsible for privatizing data), there exists an interesting interplay between robustness (Huber contamination) and DP, depending on the order of the two. Specifically, my recent work [39] establishes a *separation result* between two MAB settings that differ in the order of privacy protection and corruption, i.e., LDP-then-corruption (LTC) vs. Corruption-thenLDP (CTL). That is, under LTC, corruption happens after LDP mechanism while under CTL, corruption happens before the LDP mechanism. The punchline here is that LTC is a more difficult setting that leads to worse performance in the minimax sense under both online and offline MAB settings. Moreover, this performance separation between LTC and CTL increases as stronger privacy is pursued, underscoring the necessity for meticulous design and analysis of bandit algorithms that concurrently address privacy and corruption, rather than assuming a simplistic additive relationship between these factors.
Impacts. Our theoretically optimal algorithms could serve as the basis block for practitioners when deploying real-world MAB applications. Moreover, our minimax optimal mean estimators can find wide applications.

4 Others: Ph.D. Research and Current Grants Highlights

My Ph.D. research focused on the load-balancing and queueing problems in data centers and cloud computing, from its theoretical foundations to efficient algorithm design. This line of work has been mainly published at ACM SIGMETRICS (e.g., [40, 41, 42, 43]), which includes one work that resolves an open conjecture [42]. Building upon my Ph.D. research and my recent contributions detailed in Section 2, I have been awarded two NSF grants – one CRII and one Medium Collaborative Grant – in which I am the solo PI or PI at WSU.

Research Visions and Future Directions

Looking forward, my research will delve deeper into trustworthy data-driven decision-making, with a focus on (i) fairness and its interplay with privacy and robustness, (ii) privately aligning large language models (LLM) with or without RL, and (iii) fundamental open problems in private bandits and RL.

Enabling Fairness through Constraints. My prior work in constrained bandits and RL (as detailed in Section 2) lays the foundation for my exploration into fairness within interactive systems. I propose to integrate group fairness into bandits and RL by enforcing constraints on the expected utility differences between distinct groups or subgroups. This definition aims to ensure that the disparity in utilities, such as total rewards, between different demographics (e.g., male vs. female) or qualification-based subgroups, remains within a predefined threshold. This approach could encapsulate various group fairness notions like demographic parity [44], equal opportunity and equalized odds [45], which have been extensively studied in supervised learning but are less explored in bandits and RL. A significant challenge in this endeavor is managing the potentially large number of constraints, as the regret in my previous works scales with the square root of the number of constraints. This complexity is further amplified when considering individual fairness, which can be seen as a limit of group fairness with sufficiently large group sizes [46]. Additionally, I am keen to explore the intricate relationship between fairness, privacy, and robustness, an area that has seen extensive research in supervised learning but remains underexplored in bandits and RL. This exploration will extend beyond user-centric fairness to include fairness at the level of actions and individual items [47].

Aligning LLMs with Differential Privacy. A key focus of my future research lies in the private alignment of large language models (LLMs) like ChatGPT with human values, ensuring their responsible and thrustworthy use. A common approach to this alignment involves Reinforcement Learning from Human Feedback (RLHF). This process begins with training a reward model using preference data, followed by policy optimization based on this reward model, employing RL algorithms such as Proximal Policy Optimization (PPO). My research aims to establish theoretical foundations for RLHF, particularly under Differential Privacy (DP) constraints. My recent work [48] presents a pioneering tight analysis of sample complexity bounds for reward model learning under both local and central DP models. The next phase of my research will explore the integration of private policy optimization into this framework, drawing on my prior work in the field [9, 10]. Additionally, I will explore avenues to align LLMs privately without relying on RL, leveraging cutting-edge methodologies like Direct Preference Optimization (DPO) [49] or Identity-PO (IPO) [50].

Resolving Fundamental Open Problems. Throughout my journey in private bandits and RL, I have identified several open problems critical to advancing the field. For instance, in MABs, there's an absence of regret lower bounds under approximate DP, even for stochastic cases. In adversarial scenarios, there's a notable gap between the multiplicative cost of privacy in upper bounds and the additive cost in lower bounds for pure DP. In linear contextual bandits, significant discrepancies exist between upper and lower bounds in both local and central models. Another fundamental issue in contextual bandits is the proper handling of rarely switching under DP, a factor often overlooked in the literature, leading to ungrounded theoretical guarantees. Addressing this issue could also enhance the communication efficiency in federated bandits. For RL, existing lower bound proofs have gaps due to nuances in privacy definitions, which I aim to address.

References

[1] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis". *Theory of Cryptography: Third Theory of Cryptography Conference* 2006.

[2] Wenbo Ren, **Xingyu Zhou**, Jia Liu, and Ness B Shroff. "Multi-armed bandits with local differential privacy". *arXiv preprint arXiv:2007.03121* 2020.

[3] Sayak Ray Chowdhury* and **Xingyu Zhou***. "Distributed differential privacy in multi-armed bandits". *International Conference on Learning Representations (ICLR)* 2023. *Equal Contributions.

[4] Sayak Ray Chowdhury* and **Xingyu Zhou***. "Shuffle Private Linear Contextual Bandits". *International Conference on Machine Learning (ICML)* 2022. *Equal Contributions.

[5] Fengjiao Li, **Xingyu Zhou**, and Bo Ji. "Differentially Private Linear Bandits with Partial Distributed Feedback". *International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)* 2022. **Best Student Paper**.

[6] **Xingyu Zhou** and Jian Tan. "Local Differential Privacy for Bayesian Optimization". *AAAI Conference on Artificial Intelligence (AAAI)* 2021.

[7] Fengjiao Li, **Xingyu Zhou**, and Bo Ji. "(Private) Kernelized Bandits with Distributed Biased Feedback". *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)* 2023.

[8] **Xingyu Zhou** and Sayak Ray Chowdhury. "On Differentially Private Federated Linear Contextual Bandits". *arXiv preprint arXiv:2302.13945* 2023. FL Workshop @ ICML 2023 and TDPD'23.

[9] Sayak Ray Chowdhury* and **Xingyu Zhou***. "Differentially Private Regret Minimization in Episodic Markov Decision Processes". *AAAI Conference on Artificial Intelligence (AAAI)* 2022. *Equal Contributions **Oral**.

[10] **Xingyu Zhou**. "Differentially Private Reinforcement Learning with Linear Function Approximation". *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)* 2022.

[11] Youming Tao, Yulian Wu, Peng Zhao, and Di Wang. "Optimal rates of (locally) differentially private heavy-tailed multi-armed bandits". *International Conference on Artificial Intelligence and Statistics (AISTATS)* 2022.

[12] Kun Wang, Jing Dong, Baoxiang Wang, and Shuai Li. "Cascading bandit under differential privacy". *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 2022.

[13] Alexandre Rio, Merwan Barlier, Igor Colin, and Marta Soare. "Multi-Agent Best Arm Identification with Private Communications". *International Conference on Machine Learning (ICML)* 2023.

[14] Xi Chen, David Simchi-Levi, and Yining Wang. "Privacy-preserving dynamic personalized pricing with demand learning". *Management Science* 2022.

[15] Dan Qiao and Yu-Xiang Wang. "Near-optimal differentially private reinforcement learning". *International Conference on Artificial Intelligence and Statistics (AISTATS)* 2023.

[16] Dan Qiao and Yu-Xiang Wang. "Offline Reinforcement Learning with Differential Privacy". *Conference on Neural Information Processing Systems (NeurIPS)* 2023.

[17] Yuntian Deng, **Xingyu Zhou**, Arnob Ghosh, Abhishek Gupta, and Ness Shroff. "Interference Constrained Beam Alignment for Time-Varying Channels via Kernelized Bandits". *International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)* 2022. **Best Student Paper Runner-up**.

[18] **Xingyu Zhou** and Ness Shroff. “No-regret algorithms for time-varying bayesian optimization”. *Annual Conference on Information Sciences and Systems (CISS)* 2021. **Invited**.

[19] Yuntian Deng, **Xingyu Zhou**, Baekjin Kim, Ambuj Tewari, Abhishek Gupta, and Ness Shroff. “Weighted gaussian process bandits for non-stationary environments”. *International Conference on Artificial Intelligence and Statistics (AISTATS)* 2022.

[20] **Xingyu Zhou** and Bo Ji. “On kernelized multi-armed bandits with constraints”. *Advances in Neural Information Processing Systems (NeurIPS)* 2022.

[21] Arnob Ghosh, **Xingyu Zhou**, and Ness Shroff. “Provably efficient model-free constrained rl with linear function approximation”. *Advances in Neural Information Processing Systems (NeurIPS)* 2022.

[22] Arnob Ghosh, **Xingyu Zhou**, and Ness Shroff. “Achieving Sub-linear Regret in Infinite Horizon Average Reward Constrained MDP with Linear Function Approximation”. *International Conference on Learning Representations (ICLR)* 2023.

[23] Honghao Wei, Arnob Ghosh, Ness Shroff, Lei Ying, and **Xingyu Zhou**. “Provably Efficient Model-Free Algorithms for Non-stationary CMDPs”. *International Conference on Artificial Intelligence and Statistics (AISTATS)* 2023.

[24] Arnob Ghosh, **Xingyu Zhou**, and Ness Shroff. “Towards Achieving Sub-linear Regret and Hard Constraint Violation in Model-free RL”. *submitted to International Conference on Artificial Intelligence and Statistics (AISTATS)* 2023. URL: https://xingyuzhou.org/publications/RL-hard.pdf.

[25] Kihyuk Hong, Yuhang Li, and Ambuj Tewari. “An Optimization-based Algorithm for Non-stationary Kernel Bandits without Prior Knowledge”. *International Conference on Artificial Intelligence and Statistics (AISTATS)* 2023.

[26] Paul Brunzema, Alexander von Rohr, Friedrich Solowjow, and Sebastian Trimpe. “Event-Triggered Time-Varying Bayesian Optimization”. *arXiv preprint arXiv:2208.10790* 2022.

[27] Wenjie Xu, Yuning Jiang, Bratislava Svetozarevic, and Colin Jones. “Constrained efficient global optimization of expensive black-box functions”. *International Conference on Machine Learning (ICML)* 2023.

[28] Hengquan Guo, Zhu Qi, and Xin Liu. “Rectified Pessimistic-Optimistic Learning for Stochastic Continuum-armed Bandit with Constraints”. *Learning for Dynamics and Control Conference* 2023.

[29] Xilu Wang, Yaochu Jin, Sebastian Schmitt, and Markus Olhofer. “Recent advances in Bayesian optimization”. *ACM Computing Surveys* 2023.

[30] Jianyi Yang, Pengfei Li, Tongxin Li, Adam Wierman, and Shaolei Ren. “Anytime-Competitive Reinforcement Learning with Policy Prior”. *Conference on Neural Information Processing Systems (NeurIPS)* 2023.

[31] Hoai-An Nguyen and Ching-An Cheng. “Provable Reset-free Reinforcement Learning by No-Regret Reduction”. *International Conference on Machine Learning (ICML)* 2023.

[32] Ming Shi, Yingbin Liang, and Ness Shroff. “A Near-Optimal Algorithm for Safe Reinforcement Learning Under Instantaneous Hard Constraints”. *International Conference on Machine Learning (ICML)* 2023.

[33] Xin Liu, Zixian Yang, and Lei Ying. “Online Nonstochastic Control with Adversarial and Static Constraints”. *International Conference on Machine Learning (ICML)* 2023.

[34] Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. “Robustness implies privacy in statistical estimation”. *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC)* 2023.

[35] Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. “From robustness to privacy and back”. *International Conference on Machine Learning (ICML)* 2023.

[36] Kristian Georgiev and Samuel Hopkins. “Privacy induces robustness: Information-computation gaps and sparse mean estimation”. *Advances in Neural Information Processing Systems (NeurIPS)* 2022.

[37] Peter J Huber. “Robust Estimation of a Location Parameter”. *Ann. Math. Statist.* 1964.

[38] Yulian Wu*, **Xingyu Zhou***, Youming Tao, and Di Wang. “On Private and Robust Bandits”. *Conference on Neural Information Processing Systems (NeurIPS)* 2023. *Equal Contributions.

[39] **Xingyu Zhou** and Wei Zhang. “Locally Private and Robust Multi-Armed Bandits”. *submitted to ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)* 2024. URL: https://xingyuzhou.org/publications/LDPR-MAB.pdf.

[40] **Xingyu Zhou**, Fei Wu, Jian Tan, Yin Sun, and Ness Shroff. “Designing low-complexity heavy-traffic delay-optimal load balancing schemes: Theory to algorithms”. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)* 2018.

[41] **Xingyu Zhou***, Fei Wu*, Jian Tan, Kannan Srinivasan, and Ness Shroff. “Degree of queue imbalance: Overcoming the limitation of heavy-traffic delay optimality in load balancing systems”. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)* 2018. *Equal Contributions.

[42] **Xingyu Zhou**, Jian Tan, and Ness Shroff. “Heavy-traffic delay optimality in pull-based load balancing systems: Necessary and sufficient conditions”. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)* 2018.

[43] Wentao Weng, Xingyu Zhou, and R Srikant. “Optimal load balancing with locality constraints”. *Proceedings of the ACM on Measurement and Analysis of Computing Systems (SIGMETRICS)* 2021.

[44] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. “Fairness through awareness”. *Proceedings of the 3rd innovations in theoretical computer science conference (ITCS)* 2012.

[45] Moritz Hardt, Eric Price, and Nati Srebro. “Equality of opportunity in supervised learning”. *Advances in neural information processing systems (NeurIPS)* 2016.

[46] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. “Preventing fairness gerrymandering: Auditing and learning for subgroup fairness”. *International conference on machine learning (ICML)* 2018.

[47] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. “Fairness in learning: Classic and contextual bandits”. *Advances in neural information processing systems (NeurIPS)* 2016.

[48] Sayak Ray Chowdhury*, **Xingyu Zhou***, and Nagarajan Natarajan. “Differentially Private Reward Estimation with Preference Feedback”. *arXiv preprint arXiv:2310.19733* 2023. *Equal Contributions.

[49] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. “Direct preference optimization: Your language model is secretly a reward model”. *Conference on Neural Information Processing Systems (NeurIPS)* 2023.

[50] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. “A general theoretical paradigm to understand learning from human preferences”. *arXiv preprint arXiv:2310.12036* 2023.