My research aims to address various **ethical issues** arising from artificial intelligence (AI) and **develop human-centered AI systems with positive societal impacts**. AI techniques have seen significant advances over the last decades and are playing an increasingly critical role in people's lives. While the hope is to improve societal outcomes with these techniques, they may suffer from shortcomings and behave in potentially harmful ways:

1. When AI systems are developed with people's sensitive data or deployed to make consequential decisions about people, they may violate *social norms* such as privacy, security, fairness, etc.
2. Most AI systems are built by assuming the environment is static, without accounting for people's behaviors and the *feedback loop* between people and AI systems, which may result in unintended outcomes.

By leveraging the power of both *AI systems* and *people*, I aim to develop theoretical understandings and principled approaches to tackling these problems. Drawing upon fields such as machine learning (ML), statistics, optimization, economics, etc., I have focused on addressing two critical issues: (1) *privacy* and *security*, (2) *fairness*, and meanwhile considering the human element. I am interested in answering questions such as:

- How to embed societal constraints into the design of AI systems while preserving its usefulness?
- How does the AI system interact with people? What are the long-term impacts they have on each other?
- How to design interventions to induce individual behavior that benefits people and/or AI systems?

Privacy and Security. When AI systems are developed using individuals' data, their private information is at high risk of being compromised, resulting in potentially significant harm to both data collectors and data owners. Moreover, privacy concerns have become a major source of distrust and a major obstacle to people sharing their data with data analysts, resulting in a lack of sufficient data to develop robust and accurate computational models. To address privacy and security issues, I have focused on the following research directions:

- From data analysts' perspectives, it is critical to preserve individual privacy and accomplish computations with high accuracy while building AI systems. I have ***designed novel privacy-preserving algorithms*** with rigorous guarantees in various settings (e.g., distributed optimization [1, 2, 3], sequential computations [4, 5, 6]), which *significantly improve the privacy-accuracy trade-off over the state-of-the-art methods.*
- From data owners' perspectives, investing in security (e.g., purchasing anti-virus products) can effectively avoid data breaches. However, strategic individuals are likely to under-invest and take advantage of others' security investments. I have *introduced two new methods for addressing under-investment issues* and have ***designed mechanisms that can incentivize people to voluntarily secure themselves*** [7, 8, 9, 10].
- Essentially, privacy can be regarded as a personal commodity. If privacy violations cannot be avoided, data analysts can compensate for potential losses in advance. People value their privacy but may be willing to sell their data if they are adequately compensated. I have studied the problem of ***trading private data*** [11, 12], and I *established a new transaction framework that can benefit both data analysts and data owners.*

Fairness. AI systems built with real-world data can inherit biases and exhibit discrimination against already-disadvantaged or marginalized social groups. Due to the feedback loop between AI systems and people, biases in the algorithmic decisions can be captured in the future dataset and affect future AI systems. The most commonly used approach to alleviating discrimination is enforcing certain fairness constraints when building AI systems. However, its effectiveness is mostly studied in a static framework without accounting for AI-people dynamics. To understand this interactive process, I have studied ***fairness problems in sequential decision-making contexts*** [13, 14]. My research is among the earliest works that *gives theoretical understandings of the long-term impacts of AI (fair) systems on different social groups.* I developed new theoretical frameworks to model AI-people dynamics. My work highlights the potential pitfalls of the commonly used fairness constraints and provides guidance on designing effective interventions that promote long-term social equality.

Going forward, I am excited about the opportunity to combine AI techniques and studies on human behaviors to understand their interactions better and to build more powerful hybrid systems toward good societal outcomes. In what follows, I will elaborate on each of the research topics I mentioned above, by introducing challenges, the progress I have made toward addressing them, and future directions.

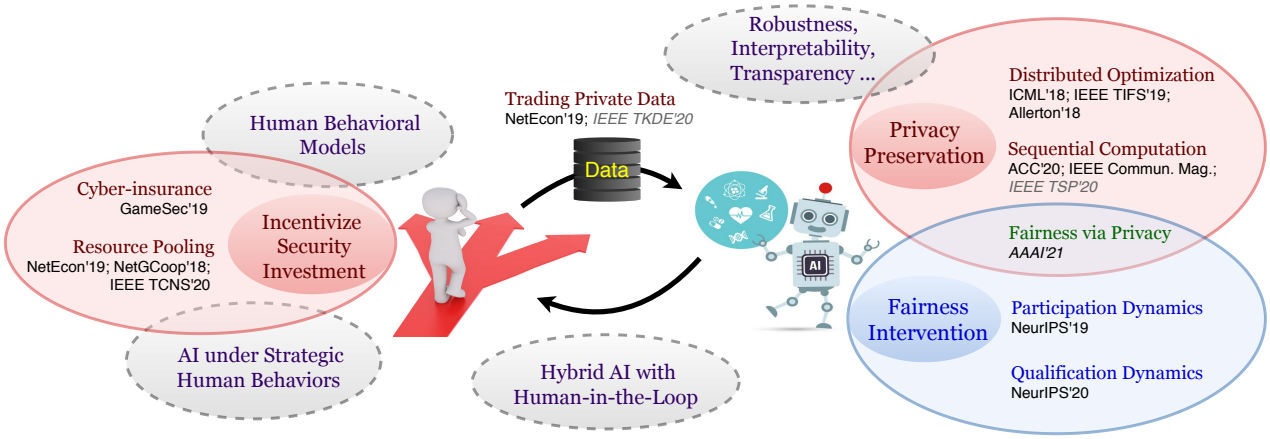

Figure 1: Overview of my completed research and future plan on *human-centered* AI.

Privacy and Security

Designing Privacy-Preserving Algorithms. One approach to leveraging people's data while preventing privacy violation, is to process sensitive data with privacy-preserving algorithms. *Differential privacy (DP)*, as a widely used notion of privacy, ensures that no one by observing the computational outcome can infer a particular individual's data with high confidence. However, DP is typically achieved by randomizing algorithms (e.g., adding noise), which inevitably leads to the trade-off between individual privacy and the outcome accuracy. This trade-off can be difficult to balance, especially for settings where the same or correlated data is repeatedly used/exposed during the computation. I have designed private algorithms that can improve the privacy-accuracy trade-off significantly over the existing algorithms for various computational tasks, including:

- *Distributed optimization.* We studied a consensus problem in a fully distributed setting where multiple entities collaboratively work toward a common optimization objective through an interactive process of local computation (over local, private data) and message passing. We focused on the Alternating Direction Method of Multiplier (ADMM)-based algorithms to solve the distributed optimization. Because attackers can infer an individual's information from all the exchanged local computations, privacy leakage accumulates substantially over time. To improve the privacy-accuracy trade-off, I have explored two ideas:
  (a) *Reuse intermediate computational results to reduce the total information leakage.*
  (b) *Improve algorithmic robustness to accommodate more randomness.*
  Intuitively, when less information is revealed, less randomization is required to achieve the same privacy guarantee, so that the accuracy can be increased; when an algorithm is more robust, it can accommodate more randomization to enhance privacy without jeopardizing too much accuracy. Based on these ideas, we designed multiple novel algorithms whose privacy-accuracy trade-off is improved significantly over conventional ADMM. Specifically, R-ADMM [2] utilizes (a) and ensures the privacy leakage only happens in *half* of the updates; M-ADMM [1] utilizes (b) which improves the algorithmic robustness; MR-ADMM [3] incorporates both ideas to improve the trade-off further.

- *Sequential computations.* Many data analytics applications rely on temporal data, generated/acquired sequentially for online analysis. How to release this type of data in a privacy-preserving manner is of great interest and more challenging than releasing one-time, static data. Due to the temporal correlation within the data sequence, total privacy leakage accumulates substantially over time. A method for alleviating this issue is to factor the correlation into the perturbation mechanism. However, existing work either focuses on offline settings or requires a priori information on the correlation in generating perturbation. In contrast, the approach we proposed in [4] can release the sequential data with DP guarantee in real-time, where the sequence correlation is not required a priori but can be learned as the sequence is generated. This method has been used to enable private vehicle-to-vehicle communication in intelligent transportation systems [5, 6].

Designing Incentive Mechanisms. Incentivizing individuals to increase their security investments (e.g., purchasing anti-virus products) voluntarily is an effective way to improve network security and mitigate cyber risks. However, in the presence of risk dependencies, strategic and selfish individuals *under-invest* in security to take advantage of others' security investments, resulting in a less secure environment. To address the under-investment issue and increase security investment, I propose two approaches:

- *Cyber-insurance.* To mitigate cyber risks, individuals can purchase *cyber-insurance* to transfer their risks to the insurer, i.e., pay premiums in exchange for coverage in the event of a loss incident (e.g., cyber attack). Due to the reduced risks, individuals (insureds) may lower their efforts, leading to a worse state of security. To address this issue, we designed a new contract using *premium discrimination*, i.e., an insured with a better security posture pays a lower premium, and showed that it could incentivize security investments [7].
- *Resource pooling.* Unlike cyber-insurance, where a social planner, i.e., cyber insurer, is required to implement the mechanism, resource pooling can address under-investment issue without any social planner. Specifically, we studied *interdependent security games* among a group of strategic and selfish individuals [8, 9, 10], where individuals are allowed to not only invest in themselves but also pool their resources to invest in others. We showed that under resource pooling, both individuals' investments and their utilities can be improved.

Trading Private Data. Another way to tackle the privacy issue is to consider private data as a personal commodity that data analysts (buyer) need to purchase from data owners (seller) before using it. We studied this problem in [11, 12] where a buyer aims to minimize the payment to sellers for a desired level of data quality, while the latter aim to obtain adequate compensation for giving up a certain amount of privacy. The transaction is facilitated by a *contract* and a *differentially private algorithm*; both are designed by a trusted, neutral third party (data broker). Specifically, the data broker collects relevant data from sellers and generates the private outcome for certain computation (requested by the buyer) using a *differentially private algorithm*; the buyer after receiving the outcome pays each individual, through the broker, an amount (determined by *contract*) commensurate with the privacy leakage the individual experiences as a result of releasing the computational outcome.

Because different people have different attitudes about their privacy, it is crucial to design the contract and private algorithm such that: (1) buyer's payment is minimized for a given accuracy level; and (2) privacy guarantee can be provided to each seller according to his/her own privacy valuation. We designed a novel differentially private algorithm and an optimal contract, under which the above two requirements are satisfied; meanwhile, the buyer's payment-accuracy trade-off can be improved significantly compared to other private algorithms.

Algorithmic Fairness

One commonly used approach to alleviating unfairness issue is to enforce fairness constraints upon the training process such that certain statistical measures (e.g., true positive rate, positive classification rate, etc.) across different social groups are (approximately) equalized. While the effectiveness of these fairness constraints has been shown in various domains, most of the studies are conducted under a *static* framework where only the immediate impact of constraints is assessed but not their long-term consequences. Because algorithmic decisions and people interact with each other over time, it is essential to study fairness problems under ML-people dynamics and examine the *long-term* impact of (fair) ML decisions on the well-being of different social groups.

I have conducted two studies under different types of dynamics, and both have shown that under certain conditions, conventional fairness constraints (e.g.,*demographic parity*, *equal opportunity*) that intend to protect disadvantaged groups may lead to unintended, pernicious long-term effects by amplifying the unfairness. These results highlight the potential pitfalls of the static fairness constraints and that long-term fairness cannot be designed in a vacuum without considering the human element. We thus emphasize the importance of performing real-time measurements and developing proper fair ML models from *dynamic* datasets.

Participation Dynamics. ML models trained on data from multiple social groups can inherit potential representation disparity in the data: the model may be less favorable to groups contributing less to the training

process; this in turn can degrade population retention in these groups over time and exacerbate representation disparity in the long-run. This problem was rigorously studied in [13]. We aimed to understand how ML models and *group representation* evolve in a sequential framework and how enforcing fairness constraints plays a role in this process. Specifically, we constructed a user *participation dynamics* model where individuals respond to perceived decisions by leaving the system uniformly at random: people who perceive mistreatment from the decisions are more likely to leave. Under such dynamics, we showed that *group representation disparity* could get exacerbated over time very easily under commonly used fair ML decisions, resulting in certain groups diminishing from the sample pool gradually. We thus developed a framework applicable to any participation dynamics, which enables us to find proper fairness constraints that can balance group representation in the long-run.

Qualification Dynamics. In high stakes applications such as lending, hiring, criminal sentencing, etc., decisions are typically made based on individuals' qualifications: assigning positive decisions to those most qualified (e.g., in lending, loans are issued to applicants that are most capable of repaying). After receiving decisions, individuals will take actions (e.g., exerting efforts, imitating others, etc.), which results in changes in their future qualifications. As a result, the qualification rate—the fraction of the qualified people—of each group changes accordingly. Understanding how the *qualification rates* of different groups evolve and examining the long-term impact of (fair) ML decisions on qualifications are important and studied in [14]. We first constructed a *qualification dynamics* model and then conducted equilibrium analyses under various fairness constraints. Our results show that imposing conventional fairness constraints may result in adverse effects and exacerbate *group qualification disparity* in the long-run, and the same fairness constraint can have opposite impacts (either exacerbate or mitigate disparity) under different problem scenarios. We thus proposed effective interventions that can improve groups' long-term qualifications and promote equality across different social groups.

Ongoing Work and Future Directions

In the future, I plan to advance my research on building human-centered AI systems with long-term societal benefits. I will broaden my research by studying issues beyond privacy & security, fairness. To make AI systems and people work along with each other better and more efficiently, I am interested in the following problems:

The Intersection Between Pillars of Trust in AI. One line of my research is to build trustworthy AI systems by embedding social norms. I have worked on several pillars of trust separately: privacy, security, and fairness. Indeed, there is a strong connection between them. It is interesting to study the impact of one on the other (e.g., whether achieving privacy helps improve fairness and vice versa). As a starting point, our work [15] studies the *compatibility* of privacy and fairness in selection problems. We identified conditions under which *perfect* fairness can be attained *for free* via differentially private algorithms. In the future, I will continue studying the relations among various pillars of trust. On the other hand, sometimes achieving one societal constraint may add difficulties to satisfy another. For instance, it becomes more difficult to develop fair ML models when protected attributes (e.g., race, gender) are *private* and *unobservable*. Building upon the relations among pillars of trust, I will also develop AI systems that simultaneously satisfy multiple social norms.

ML in the Presence of Strategic Human Behaviors. As ML models are increasingly used in making decisions about people, there is an increasing requirement for the transparency of these ML models. However, individuals are strategic: with (partial) information of ML models, they can adapt their behaviors by strategically manipulating their features or investing efforts to receive favorable decisions. For example, hiring processes that heavily depend on GPA motivate students to cheat in exams for higher scores; loan applicants may feel compelled to avoid credit card debt if a bank makes decisions based on such information. Therefore, it is crucial to develop ML models that account for such strategic behaviors. What particularly interests me is the question of how to design models that (1) *disincentivize* individuals to manipulate (e.g., cheating in exams); (2) *incentivize* individuals to invest in forms of effort that increase societal outcomes (e.g., reduce default rate in lending).

Learning Human Behavioral Models. My research has highlighted the importance of understanding human behaviors in building AI systems. When examining the interaction between strategic individuals and ML models,

most studies build on game-theoretical models, where individuals are assumed to be *fully rational*. However, this may not hold in practice. Instead, I am interested in learning interpretable human behavioral models using ML techniques via empirical studies. I plan to develop online crowdsourcing platforms or survey sites to collect *dynamic data* from people and then use ML algorithms to train a human behavioral model. Such a model is essential for building *human-centered* AI systems. It may help advance ML research towards a more interpretable domain and open up the possibility of understanding the causal relationships of human-generated data.

Hybrid AI System With Human-in-the-Loop. Both humans and machines are essential in developing AI systems. How can we integrate these two components and utilize the strengths of both? I am interested in building a collaborative environment where machines and humans can make joint decisions. Specifically, it includes designing (1) incentive mechanisms to encourage the participation of human experts, (2) the systems that aggregate and maximally utilize human contribution to improve the machine intelligence.

An Ecosystem of Trust. In addition to privacy, security, and fairness, there are many other issues that threaten the trustworthiness of AI. For example, it has been well-documented that ML models are vulnerable to adversarial attacks, and they may make hard-to-justify predictions with a lack of transparency. Therefore, a trusted AI system should also be robust, interpretable, transparent, etc. I plan to develop strong theoretical foundations and principled methods for trustworthy AI in my future research. Because of the potential strong connections between these pillars of trust, an interesting question is whether my experiences in privacy, security, and fairness can help tackle other issues. As a starting point, I will first study such connections and then, if possible, adapt the approaches I have developed in privacy, security, and fairness to broader contexts.

References

[1] **X. Zhang**, M. Khalili and M. Liu. Improving the Privacy and Accuracy of ADMM-based Distributed Algorithms. *In the 35th International Conference on Machine Learning **(ICML)**, 2018.*

[2] **X. Zhang**, M. Khalili and M. Liu. Recycled ADMM: Improve Privacy and Accuracy with Less Computation in Distributed Algorithms. *In the 56th Annual Allerton Conference on Communication, Control, and Computing **(Allerton)**, 2018.*

[3] **X. Zhang**, M. Khalili and M. Liu. Recycled ADMM: Improving the Privacy and Accuracy of Distributed Algorithms. *In IEEE Transactions on Information Forensics and Security **(TIFS)**, 2019.*

[4] **X. Zhang**, M. Khalili and M. Liu. Real-time Release of Sequential Data under Differential Privacy Constraint. *In IEEE Transactions on Signal Processing **(TSP)**, submitted, 2020.*

[5] **X. Zhang***, C. Huang*, M. Liu, A. Stefanopoulou and T. Ersal. Predictive Cruise Control with Private Vehicle-to-Vehicle Communication for Improving Fuel Consumption and Emissions. *In IEEE Communications Magazine, 2019.*

[6] C. Huang, **X. Zhang**, R. Salehi, T. Ersal and A. Stefanopoulou. A Robust Energy and Emissions Conscious Cruise Controller for Connected Vehicles with Privacy Considerations. *In 2020 American Control Conference **(ACC)**, 2020.*

[7] M. Khalili, **X. Zhang** and M. Liu. Effective Premium Discrimination for Designing Cyber Insurance Policies with Rare Losses. *In the 10th Conference on Decision and Game Theory for Security **(GameSec)**, 2019.*

[8] M. Khalili, **X. Zhang** and M. Liu. Public Good Provision Games on Networks with Resource Pooling. *In the International Conference on Network Games Control and Optimization **(NetGCoop)**, 2018.*

[9] M. Khalili, **X. Zhang** and M. Liu. Incentivizing Effort in Interdependent Security Games Using Resource Pooling. *In the 14th Workshop on the Economics of Networks, Systems and Computation **(NetEcon)**, 2019.*

[10] M. Khalili, **X. Zhang** and M. Liu. Resource Pooling for Shared Fate: Incentivizing Effort in Interdependent Security Games through Cross-investments. *In IEEE Transactions on Control of Network Systems **(TCNS)**, 2020.*

[11] M. Khalili*, **X. Zhang*** and M. Liu. Contract Design for Purchasing Private Data Using a Biased Differentially Private Algorithm. *In the 14th Workshop on the Economics of Networks, Systems and Computation **(NetEcon)**, 2019.*

[12] M. Khalili*, **X. Zhang*** and M. Liu. Designing Contracts for Trading Private and Heterogeneous Data Using a Biased Differentially Private Algorithm, *In IEEE Transactions on Knowledge and Data Engineering **(TKDE)**, submitted, 2020.*

[13] **X. Zhang***, M. Khalili*, C. Tekin and M. Liu. Group Retention when Using Machine Learning in Sequential Decision Making: the Interplay between User Dynamics and Fairness. *In the 33rd Conference on Neural Information Processing Systems **(NeurIPS)**, 2019.*

[14] **X. Zhang***, R. Tu*, Y. Liu, M. Liu, H. Kjellström, K. Zhang and C. Zhang. How Do Fair Decisions Fare in Long-term Qualification? *In the 34th Conference on Neural Information Processing Systems **(NeurIPS)**, 2020.*

[15] M. Khalili, **X. Zhang**, M. Abroshan and S. Sojoudi. Improving Fairness and Privacy in Selection Problems. *In the 35th AAAI Conference on Artificial Intelligence **(AAAI)**, 2021.*