Zhiwei Steven Wu

Contact Information

4665 Forbes Avenue
TCS Hall 424
Pittsburgh, PA 15213

Voice: (845)-475-8029
E-mail: zstevenwu@cmu.edu
http://www.zstevenwu.com

(Last updated: December 16, 2025)

Research Interests

Machine Learning, AI Safety and Alignment, Data Privacy, Algorithmic Fairness, Algorithmic Economics

Employment

Carnegie Mellon University — Pittsburgh, Pennsylvania USA

Associate Professor, School of Computer Science — July 2025 –
Assistant Professor — September 2020 – June 2025
Software and Societal Systems Department (in Societal Computing)
Machine Learning Department (affiliated)
Human-Computer Interaction Institute (affiliated)

University of Minnesota — Twin Cities, Minnesota USA

Assistant Professor of Computer Science — August 2018 – August 2020

Microsoft Research-New York City — New York City, New York USA

Postdoctoral Researcher — July 2017 – June 2018
Research Groups: Machine Learning & Algorithmic Economics

Education

University of Pennsylvania — Philadelphia, Pennsylvania USA

Ph.D., Computer Science — September 2012 – June 2017
Thesis: *Data Privacy Beyond Differential Privacy*
Advisors: Michael Kearns & Aaron Roth
Received the Morris and Dorothy Rubinoff Dissertation Award (Best Thesis)

Bard College — Annandale, NY USA

B.A., Mathematics & Computer Science — May 2012
Distinguished Scientist Scholarship (four-year full scholarship)

Budapest Semesters in Mathematics (BSM), — Budapest, Hungary

Study-abroad program in mathematics — Fall 2010

Honors and Awards

2025 - Winner of all four tracks of The Vector Institute MIDST challenge (Membership Inference over Diffusion-models-based Synthetic Tabular data) at SaTML 2025
2024 - NSF CAREER Award
2023 - FAccT Best Paper Award
2023 - First place at the U.S. Privacy-Enhancing Technologies (PETs) Prize Challenge, Pandemic Forecasting Track
2021 - Okawa Foundation Research Grant
2022, 2021, 2019 - Facebook/Meta Research Award (three times)
2019 - Google Faculty Research Award
2022, 2019 - J.P. Morgan Research Faculty Award (twice)

2017 - Morris and Dorothy Rubinoff Dissertation Award for Best Thesis
2017 - Simons-Berkeley Research Fellowship (declined)
2011 - Kenneth Bush Memorial Scholarship in Mathematics
2010 - BSM Mathematics Competition Honorable Mention
2010 - Mathematical Association of America Presentation Prize
2008–2012 - Distinguished Scientist Scholarship (four-year full scholarship)

Journal Publications

(Unless specified otherwise, authors in all papers are listed in alphabetical order. The * sign indicates equal contribution.)

- Guy Aridor, Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu Competing Bandits: The Perils of Exploration Under Competition. In *ACM Transactions on Economics and Computation,* **TEAC**, Volume 13, Issue 1, March 2025. Extended version of the EC'19 and EC'18 papers
- Satyapriya Krishna, Tessa Han, Alex Gu, Zhiwei Steven Wu, Shahin Jabbari, and Himabindu Lakkaraju The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. In *Transactions on Machine Learning Research,* **TMLR**, 2024. (Contributional order)
- Travis Dick, Cynthia Dwork, Michael Kearns, Terrance Liu, Aaron Roth, Giuseppe Vietri, Zhiwei Steven Wu. Confidence-Ranked Reconstruction of Census Microdata from Published Statistics. *Proceedings of the National Academy of Sciences* (**PNAS**), 2023
- Manish Raghavan, Aleksandrs Slivkins, Jennifer Wortman Vaughan, Zhiwei Steven Wu. Greedy Algorithm Almost Dominates in Smoothed Contextual Bandits. In *SIAM Journal on Computing* (**SICOMP**), 2023
- Shengyuan Hu, Zhiwei Steven Wu, Virginia Smith. Private Multi-Task Learning: Formulation and Applications to Federated Learning. In *Transactions of Machine Learning Research,* **TMLR**, 2023. (Contributional order)
- Ryan Steed, Terrance Liu, Zhiwei Steven Wu, and Alessandro Acquisti. Policy impacts of statistical uncertainty and privacy. In **Science**, Aug 2022. (Contributional order)
- Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Operations Research* (**OR**), 2021
- Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *IEEE Transactions on Information Theory*, 2021
- Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P Bhavnani, James Brian Byrd, Casey S. Greene and Casey S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. In *Circulation: Cardiovascular Quality and Outcomes 2019; 12* (Contributional order)
- Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. Accuracy first: Selecting a differential privacy level for accuracy-constrained ERM. In *The Journal of Privacy and Confidentiality,* **JPC**, 2019. Previously published in NeurIPS 2017
- Paul W. Goldberg, Francisco J. Marmolejo Cossío, and Zhiwei Steven Wu. Logarithmic query complexity for approximate nash computation in large games. *Theory of Computing Systems* (**TOCS**), 2019. Special issue for selected papers from SAGT 2016
- Michael Kearns, Aaron Roth, Zhiwei Steven Wu, and Grigory Yaroslavtsev. Private algorithms for the protected in social network search. *Proceedings of the National Academy of Sciences* (**PNAS**), 113(4), 2016

- Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. *SIAM Journal on Computing* (**SICOMP**), 2016. Previously published in ACM SIGACT Symposium on Theory of Computing (STOC 2014)

- Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu. Dual query: Practical private query release for high dimensional data. *Journal of Privacy and Confidentiality* (**JPC**), 2016. Previously published in ICML 2014

Conference Publications

- Terrance Liu, Eileen Xiao, Adam Smith, Pratiksha Thaker, and Zhiwei Steven Wu Generate-then-Verify: Reconstructing Data from Limited Published Statistics. In *Proceedings of the 47th IEEE Symposium on Security and Privacy,* **S&P**, 2026. (Contributional order)

- Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova Validating LLM-as-a-Judge Systems under Rating Indeterminacy. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2025. (Contributional order)

- Xiaoyu Wu, Yifei Pang, Terrance Liu, and Zhiwei Steven Wu Unlearned but Not Forgotten: Data Extraction after Exact Unlearning in LLM. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2025. (Contributional order)

- Hongyi Henry Jin, Zijun Ding, Dung Daniel Ngo, and Zhiwei Steven Wu Discretization-free Multicalibration through Loss Minimization over Tree Ensembles. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2025. (Contributional order)

- Konstantina Bairaktari, Jiayun Wu, and Zhiwei Steven Wu Kandinsky Conformal Prediction: Beyond Class- and Covariate-Conditional Coverage. In *Proceedings of the Forty-second International Conference on Machine Learning,* **ICML**, 2025

- Xiaoyu Wu, Jiaru Zhang, Zhiwei Steven Wu Leveraging Model Guidance to Extract Training Data from Personalized Diffusion Models. In *Proceedings of the Forty-second International Conference on Machine Learning,* **ICML**, 2025. (Contributional order)

- Justin Whitehouse, Christopher Jung, Vasilis Syrgkanis, Bryan Wilder, and Zhiwei Steven Wu Orthogonal Causal Calibration. In *Proceedings of the 38th Annual Conference on Learning Theory,* **COLT**, 2025. (Contributional order)

- Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas Time-Uniform Self-Normalized Concentration for Vector-Valued Processes. In *Proceedings of the 38th Annual Conference on Learning Theory,* **COLT**, 2025. (Contributional order). Full version to appear in the Annals of Applied Probability

- Kimberly Le Truong, Riccardo Fogliato, Hoda Heidari, and Zhiwei Steven Wu Persona-Augmented Benchmarking: Evaluating LLMs Across Diverse Writing Styles. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* **EMNLP**, 2025. (Contributional order)

- Kevin Ren, Santiago Cortes-Gomez, Carlos Miguel Patiño, Ananya Joshi, Ruiqi Lyu, Jingjing Tang, Alistair Turcan, Khurram Yamin, Zhiwei Steven Wu, and Bryan Wilder Predicting Language Models' Success at Zero-Shot Probabilistic Prediction. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing,* **EMNLP**, 2025. (Contributional order)

- Luke Guerdan, Devansh Saxena, Stevie Chancellor, Zhiwei Steven Wu, and Kenneth Holstein Measurement as Bricolage: Examining How Data Scientists Construct Target Variables for Predictive

Modeling Tasks. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing,* **CSCW**, 2025. (Contributional order)

- Ally Yalei Du, Dung Daniel Ngo, and Zhiwei Steven Wu Reconciling Model Multiplicity for Downstream Decision Making. In *Proceedings of the Thirteenth International Conference on Learning Representations,* **ICLR**, 2025. (Contributional order)

- Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith Unlearning or Obfuscating? Jogging the Memory of Unlearned LLMs via Benign Relearning. In *Proceedings of the Thirteenth International Conference on Learning Representations,* **ICLR**, 2025. (Contributional order)

- Santiago Cortes-Gomez, Carlos Miguel Patiño, Yewon Byun, Zhiwei Steven Wu, Eric Horvitz, and Bryan Wilder Decision-Focused Uncertainty Quantification. In *Proceedings of the Thirteenth International Conference on Learning Representations,* **ICLR**, 2025. (Contributional order)

- Xin Gu, Gautam Kamath, and Zhiwei Steven Wu Choosing Public Datasets for Private Machine Learning via Gradient Subspace Distance. In *Proceedings of the IEEE Secure and Trustworthy Machine Learning Conference,* **SaTML**, 2025. (Contributional order)

- Pratiksha Thaker, Shengyuan Hu, Neil Kale, Yash Maurya, Zhiwei Steven Wu, and Virginia Smith Position: LLM Unlearning Benchmarks are Weak Measures of Progress. In *Proceedings of the IEEE Secure and Trustworthy Machine Learning Conference,* **SaTML**, 2025. (Contributional order)

- Yuqi Pan, Zhiwei Steven Wu, Haifeng Xu, and Shuran Zheng Differentially Private Bayesian Persuasion. In *Proceedings of The Web Conference,* **WWW (Selected for a oral presentation)**, 2025

- Shuaiqi Wang, Shuran Zheng, Zinan Lin, Giulia Fanti, and Zhiwei Steven Wu Inferentially-Private Private Information. In *Proceedings of The Web Conference,* **WWW**, 2025. (Contributional order)

- Alexander Goldberg, Giulia Fanti, Nihar B. Shah, and Zhiwei Steven Wu Benchmarking Fraud Detectors on Private Graph Data. In *Proceedings of the 31st SIGKDD Conference on Knowledge Discovery and Data Mining,* **KDD**, 2025. (Contributional order)

- Adam Block, Mark Bun, Rathin Desai, Abhishek Shetty, and Zhiwei Steven Wu Oracle-Efficient Differentially Private Learning with Public Data. In *Advances in Neural Information Processing Systems,* **NeurIPS**, 2024

- Keegan Harris, Zhiwei Steven Wu, and Maria-Florina Balcan Regret Minimization in Stackelberg Games with Side Information. In *Advances in Neural Information Processing Systems,* **NeurIPS**, 2024. (Contributional order)

- Jingwu Tang, Gokul Swamy, Fei Fang, and Zhiwei Steven Wu Multi-Agent Imitation Learning: Value is Easy, Regret is Hard. In *Advances in Neural Information Processing Systems,* **NeurIPS**, 2024. (Contributional order)

- Jiayun Wu, Jiashuo Liu, Peng Cui, and Zhiwei Steven Wu Bridging Multicalibration and Out-of-distribution Generalization Beyond Covariate Shift. In *Advances in Neural Information Processing Systems,* **NeurIPS**, 2024. (Contributional order)

- Martin Andres Bertran, Shuai Tang, Michael Kearns, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu Reconstruction Attacks on Machine Unlearning: Simple Models are Vulnerable. In *Advances in Neural Information Processing Systems,* **NeurIPS**, 2024. (Contributional order)

- Pratiksha Thaker, Amrith Setlur, Zhiwei Steven Wu, and Virginia Smith On the Benefits of Public Representations for Private Transfer Learning under Distribution Shift. In *Advances in Neural Information Processing Systems,* **NeurIPS**, 2024. (Contributional order)

- Jiahao Zhang, Shuran Zheng, Renato Paes Leme, and Zhiwei Steven Wu Ex-post Individually Rational Bayesian Persuasion. In *Proceedings of the The 21st International Conference on Web and Internet Economics,* **WINE**, 2024. (Contributional order)

- Luke Guerdan, Amanda Coston, Ken Holstein, and Zhiwei Steven Wu Predictive Performance Comparison of Decision Policies Under Confounding. In *Proceedings of the 41st International Conference on Machine Learning,* **ICML**, 2024. (Contributional order)

- Gokul Swamy, Rahul Kidambi, Christoph Dann, Zhiwei Steven Wu, and Alekh Agarwal A Minimaximalist Approach to Reinforcement Learning from Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning,* **ICML**, 2024. (Contributional order)

- Juntao Ren, Gokul Swamy, Zhiwei Steven Wu, Drew Bagnell, and Sanjiban Choudhury Hybrid Inverse Reinforcement Learning. In *Proceedings of the 41st International Conference on Machine Learning,* **ICML**, 2024. (Contributional order)

- Shuai Tang, Zhiwei Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership Inference Attacks on Diffusion Models via Quantile Regression. In *Proceedings of the 41st International Conference on Machine Learning,* **ICML**, 2024. (Contributional order)

- Keegan Harris, Anish Agarwal, Chara Podimata, and Zhiwei Steven Wu. Strategyproof Decision-Making in Panel Data Settings and Beyond. In *The ACM SIGMETRICS 2024 Conference* **SIGMETRICS**, 2024 (Contributional order)

- Xinwei Zhang, Zhiqi Bu, Zhiwei Steven Wu, and Mingyi Hong. Differentially Private SGD Without Clipping Bias: An Error-Feedback Approach. In *The 12th International Conference on Learning Representations* **ICLR**, 2024 (Contributional order)

- Shuai Tang, Sergul Aydore, Michael Kearns, Saeyoung Rho, Aaron Roth, Yichen Wang, Yu-Xiang Wang, and Zhiwei Steven Wu. Improved Differentially Private Regression via Gradient Boosting. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning* **SaTML**, 2024 (Contributional order)

- Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Fair Federated Learning via Bounded Group Loss. In *2nd IEEE Conference on Secure and Trustworthy Machine Learning* **SaTML**, 2024 (Contributional order)

- Justin Whitehouse, Zhiwei Steven Wu, and Aaditya Ramdas. On the Sublinear Regret of GP-UCB. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2023 (Contributional order)

- Keegan Harris, Chara Podimata, and Zhiwei Steven Wu. Strategic Apple Tasting. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2023 (Contributional order)

- Konwoo Kim, Gokul Swamy, Zuxin Liu, Ding Zhao, Sanjiban Choudhury, and Zhiwei Steven Wu. Learning Shared Safety Constraints from Multi-task Demonstrations. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2023 (Contributional order)

- Ryan Rogers, Gennady Samorodnitsky, Zhiwei Steven Wu, and Aaditya Ramdas. Adaptive Privacy Composition for Accuracy-first Mechanisms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2023 (Contributional order)

- Martin Andres Bertran, Shuai Tang, Aaron Roth, Michael Kearns, Jamie Heather Morgenstern, and Zhiwei Steven Wu. Scalable Membership Inference Attacks via Quantile Regression. In *Advances in*

Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems, **NeurIPS**, 2023 (Contributional order)

- Anish Agarwal, Keegan Harris, Justin Whitehouse, Zhiwei Steven Wu. Adaptive Principal Component Regression with Applications to Panel Data. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2023

- Mikhail Khodak, Ilya Osadchiy, Keegan Harris, Nina Balcan, Kfir Yehuda Levy, Ron Meir, and Zhiwei Steven Wu. Meta-Learning Adversarial Bandit Algorithms. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2023

- Luke Guerdan, Amanda Coston, Kenneth Holstein, Zhiwei Steven Wu. Counterfactual Prediction Under Outcome Measurement Error. In *The Sixth ACM Conference on Fairness, Accountability, and Transparency,* **ACM FAccT**, 2023. (Contributional order). **Best Paper Award**

- Luke Guerdan, Amanda Coston, Zhiwei Steven Wu, Kenneth Holstein. Ground(less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. In *The Sixth ACM Conference on Fairness, Accountability, and Transparency,* **ACM FAccT**, 2023. (Contributional order)

- Justin Whitehouse, Aaditya Ramdas, Ryan Rogers, Zhiwei Steven Wu. Fully-Adaptive Composition in Differential Privacy. In *Proceedings of the 40th International Conference on Machine Learning,* **ICML**, 2023. (Contributional order)

- Terrance Liu, Jingwu Tang, Giuseppe Vietri, Zhiwei Steven Wu. Generating Private Synthetic Data with Genetic Algorithms. In *Proceedings of the 40th International Conference on Machine Learning,* **ICML**, 2023

- Ian Waudby-Smith, Zhiwei Steven Wu, Aaditya Ramdas. Nonparametric Extensions of Randomized Response for Private Confidence Sets. In *Proceedings of the 40th International Conference on Machine Learning,* **ICML**, 2023. (Contributional order)

- Gokul Swamy, David Wu, Sanjiban Choudhury, Drew Bagnell, Zhiwei Steven Wu. Inverse Reinforcement Learning without Reinforcement Learning. In *Proceedings of the 40th International Conference on Machine Learning,* **ICML**, 2023. (Contributional order)

- Keegan Harris, Ioannis Anagnostides, Gabriele Farina, Mikhail Khodak, Zhiwei Steven Wu, Tuomas Sandholm Meta-Learning in Games. In *The Eleventh International Conference on Learning Representations* **ICLR**, 2023 (Contributional order)

- Zhun Deng, He Sun, Zhiwei Steven Wu, Linjun Zhang, David C. Parkes. Reinforcement Learning with Stepwise Fairness Constraints. In *The 26th International Conference on Artificial Intelligence and Statistics* **AISTATS**, 2023 (Contributional order)

- Vladimir Braverman, Joel Manning, Zhiwei Steven Wu, Samson Zhou. Private Data Stream Analysis for Universal Symmetric Norm Estimation. In *The 27th International Conference on Randomization and Computation* **RANDOM**, 2023

- Gokul Swamy, Sanjiban Choudhury, J. Drew Bagnell, and Zhiwei Steven Wu. Sequence Model Imitation Learning with Unobserved Contexts. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2022 (Contributional order)

- Gokul Swamy, Nived Rajaraman, Matt Peng, Sanjiban Choudhury, J. Drew Bagnell, Zhiwei Steven Wu, Jiantao Jiao, Kannan Ramchandran. Minimax Optimal Online Imitation Learning via Replay Estimation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2022 (Contributional order)

- Keegan Harris, Valerie Chen, Joon Sik Kim, Ameet Talwalkar, Hoda Heidari, Zhiwei Steven Wu. Bayesian Persuasion for Algorithmic Recourse. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2022 (Contributional order)

- Giuseppe Vietri, Cedric Archambeau, Sergul Aydore, William Brown, Michael Kearns, Aaron Roth, Ankit Siva, Shuai Tang, Zhiwei Steven Wu. Private Synthetic Data for Multitask Learning and Marginal Queries. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2022 (Contributional order)

- Ken Liu, Shengyuan Hu, Zhiwei Steven Wu, Virginia Smith. On Privacy and Personalization in Cross-Silo Federated Learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2022 (Contributional order)

- Justin Whitehouse, Aaditya Ramdas, Zhiwei Steven Wu, Ryan Rogers. Brownian Noise Reduction: Maximizing Privacy Subject to Accuracy Constraints. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2022 (Contributional order)

- Xinyan Hu, Dung Daniel Ngo, Aleksandrs Slivkins, Zhiwei Steven Wu. Incentivizing Combinatorial Bandit Exploration. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2022

- Gokul Swamy, Sanjiban Choudhury, J. Drew Bagnell, and Zhiwei Steven Wu. Causal Imitation Learning under Temporally Correlated Noise. In *Proceedings of the 39th International Conference on Machine Learning,* **ICML**, 2022. (Contributional order)

- Keegan Harris, Daniel Ngo*, Logan Stapleton*, Hoda Heidari, and Zhiwei Steven Wu. Strategic Instrumental Variable Regression: Recovering Causal Relationships From Strategic Responses. In *Proceedings of the 39th International Conference on Machine Learning,* **ICML**, 2022. (Contributional order)

- Alberto Bietti, Chen-Yu Wei, Miro Dudik, John Langford, and Zhiwei Steven Wu. Personalization Improves Privacy-Accuracy Tradeoffs in Federated Optimization. In *Proceedings of the 39th International Conference on Machine Learning,* **ICML**, 2022. (Contributional order)

- Yahav Bechavod, Chara Podimata, and Juba Ziani, and Zhiwei Steven Wu. Information Discrepancy in Strategic Learning. In *Proceedings of the 39th International Conference on Machine Learning,* **ICML**, 2022

- Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. Understanding Clipping for Federated Learning: Convergence and Client-Level Differential Privacy In *Proceedings of the 39th International Conference on Machine Learning,* **ICML**, 2022. (Contributional order)

- Daniel Ngo, Giuseppe Vietri, and Zhiwei Steven Wu. Improved Regret for Differentially Private Exploration in Linear MDP In *Proceedings of the 39th International Conference on Machine Learning,* **ICML**, 2022

- Zuxin Liu, Zhepeng Cen, Vladislav Isenbaev, Wei Liu, Zhiwei Steven Wu, Bo Li, and Ding Zhao. Constrained Variational Policy Optimization for Safe Reinforcement Learning. In *Proceedings of the 39th International Conference on Machine Learning,* **ICML**, 2022. (Contributional order)

- Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Ken Holstein, and Haiyi Zhu. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *The Fifth ACM Conference on Fairness, Accountability, and Transparency,* **ACM FAccT**, 2022. (Contributional order)

- Logan Stapleton, Min Hun Lee, Diana Qing, Marya Wright, Alexandra Chouldechova, Zhiwei Steven Wu, Ken Holstein, and Haiyi Zhu. Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders In *The Fifth ACM Conference on Fairness, Accountability, and Transparency,* **ACM FAccT**, 2022. (Contributional order)

- Anna Kawakami*, Venkat Sivaraman*, Hao-Fei Cheng, Logan Stapleton, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Ken Holstein. 'Why Do I Care What's Similar?' Probing Challenges in AI-Assisted Child Welfare Decision-Making through Worker-AI Interface Design Concepts In *The 2022 ACM conference on Designing Interactive Systems,* **DIS**, 2022. (Contributional order)

- Hao-Fei Cheng*, Logan Stapleton*, Anna Kawakami, Venkatesh Sivaraman, Yang Cheng, Diana Qing, Adam Perer, Kenneth Holstein, Zhiwei Steven Wu, and Haiyi Zhu. How Child Welfare Workers Reduce Racial Disparities in Algorithmic Decisions. In *The 2022 ACM CHI Conference on Human Factors in Computing Systems,* **CHI**, 2022. (Contributional order)

- Anna Kawakami, Venkat Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yang Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, Kenneth Holstein. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *The 2022 ACM CHI Conference on Human Factors in Computing Systems,* **CHI**, 2022. (Contributional order). **Best Paper Honorable Mention Award (Top 5%)**

- Zheyuan Ryan Shi, Zhiwei Steven Wu, Rayid Ghani, and Fei Fang. Bandit Data-Driven Optimization for Crowdsourcing Food Rescue Platforms. In *The 36th AAAI Conference on Artificial Intelligence,* **AAAI**, 2022. (Contributional order)

- Keegan Harris, Hoda Heidari, and Zhiwei Steven Wu. Stateful Strategic Regression. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2021

- Terrance Liu, Giuseppe Vietri, and Zhiwei Steven Wu. Iterative Methods for Private Synthetic Data: Unifying Framework and New Methods. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2021

- Gokul Swamy, Sanjiban Choudhury, J. Drew Bagnell, and Zhiwei Steven Wu. Of Moments and Matching: Trade-offs and Treatments in Imitation Learning. In *Proceedings of the 38th International Conference on Machine Learning,* **ICML**, 2021. (Contributional order)

- Terrance Liu, Giuseppe Vietri, Thomas Steinke, Jonathan Ullman, and Zhiwei Steven Wu. Leveraging Public Data for Practical Private Query Release. In *Proceedings of the 38th International Conference on Machine Learning,* **ICML**, 2021. (Contributional order). **Received two best paper awards from ICLR 2021 workshops**: Distributed and Private Machine Learning and Synthetic Data Generation

- Sushant Agarwal, Shahin Jabbari, Chirag Agarwal, Sohini Upadhyay, Zhiwei Steven Wu, and Hima Lakkaraju. Towards the Unification and Robustness of Perturbation and Gradient Based Explanations. In *Proceedings of the 38th International Conference on Machine Learning,* **ICML**, 2021. (Contributional order)

- Daniel Ngo, Logan Stapleton, Vasilis Syrgkanis, and Zhiwei Steven Wu. Incentivizing Compliance with Algorithmic Instruments. In *Proceedings of the 38th International Conference on Machine Learning,* **ICML**, 2021. (Contributional order)

- Chris Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. An Algorithmic Framework for Fairness Elicitation. In *The second annual Symposium on Foundations of Responsible Computing* **FORC**, 2021

- Marcel Neunhoeffer, Zhiwei Steven Wu, and Cynthia Dwork. Private Post-GAN Boosting. In *The Ninth International Conference on Learning Representations,* **ICLR**, 2021. (Contributional order)

- Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the Ambient Dimension: Private SGD with Gradient Subspace Identification. In *The Ninth International Conference on Learning Representations,* **ICLR**, 2021. (Contributional order)

- Hao-Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Wu, and Haiyi Zhu Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *The 2021 ACM CHI Conference on Human Factors in Computing Systems,* **CHI**, 2021. (Contributional order)

- Vikas K. Garg, Katrina Ligett, Adam Kalai, and Zhiwei Steven Wu. Learn to Expect the Unexpected: Probably Approximately Correct Domain Generalization. In *The 24th International Conference on Artificial Intelligence and Statistics,* **AISTATS**, 2021

- Yahav Bechavod, Katrina Ligett, Zhiwei Steven Wu, and Juba Ziani. Gaming Helps! Learning from Strategic Interactions in Natural Dynamics. In *The 24th International Conference on Artificial Intelligence and Statistics,* **AISTATS**, 2021

- Hong Shen, Wesley Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. Value Cards: An Educational Toolkits for Teaching Social Impacts of Machine Learning through Deliberation. In *The Fourth ACM Conference on Fairness, Accountability, and Transparency* **ACM FAccT**, 2021. (Contributional order)

- Yahav Bechavod, Chris Jung, and Zhiwei Steven Wu. Metric-Free Individual Fairness in Online Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems,* **NeurIPS (Selected for a Oral Presentation: Top 1% of submissions)**, 2020

- Xiangyi Chen, Zhiwei Steven Wu, and Mingyi Hong. Understanding Gradient Clipping in Private SGD: A Geometric Perspective In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems,* **NeurIPS (Selected for a Spotlight Presentation: Top 2% of submissions)**, 2020. (Contribution order)

- Xiangyi Chen*, Tiancong Chen*, Haoran Sun, Zhiwei Steven Wu, and Mingyi Hong. Distributed Training with Heterogeneous Data: Bridging Median- and Mean-Based Algorithms. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2020. (Contributional order)

- Giuseppe Vietri, Grace Tian, Mark Bun, Thomas Steinke and Zhiwei Steven Wu. New Oracle-Efficient Algorithms for Private Synthetic Data Release. In *Proceedings of the 37th International Conference on Machine Learning,* **ICML**, 2020. (Contribution order)

- Giuseppe Vietri, Borja Balle, Akshay Krishnamurthy, and Zhiwei Steven Wu. Private Reinforcement Learning with PAC and Regret Guarantees. In *Proceedings of the 37th International Conference on Machine Learning,* **ICML**, 2020. (Contribution order)

- Seth Neel, Aaron Roth, Giuseppe Vietri, and Zhiwei Steven Wu. Oracle Efficient Private Non-Convex Optimization. In *Proceedings of the 37th International Conference on Machine Learning,* **ICML**, 2020

- Vidyashankar Sivakumar, Zhiwei Steven Wu, and Arindam Banerjee. Structured Linear Contextual Bandits: A Sharp and Geometric Smoothed Analysis. In *Proceedings of the 37th International Conference on Machine Learning,* **ICML**, 2020. (Contribution order)

- Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Zhiwei Steven Wu. Privately Learning Markov Random Fields. In *Proceedings of the 37th International Conference on Machine Learning,* **ICML**, 2020. (Contribution order)

- Raef Bassily, Albert Cheu, Shay Moran, Aleksandar Nikolov, Jonathan Ullman, and Zhiwei Steven Wu. Private Query Release Assisted by Public Data. In *Proceedings of the 37th International Conference on Machine Learning,* **ICML**, 2020

- Sivakanth Gopi, Gautam Kamath, Janardhan Kulkarni, Aleksandar Nikolov, Zhiwei Steven Wu, and Huanyu Zhang. Locally Private Hypothesis Selection. In *Proceedings of the 33rd Annual Conference on Learning Theory,* **COLT**, 2020

- Nicole Immorlica, Jieming Mao, Alex Slivkins, and Zhiwei Steven Wu. Incentivizing Exploration with Selective Disclosure In *The 21st ACM conference on Economics and Computation* **EC**, 2020

- Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi and Haiyi Zhu. Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-offs Across Multiple Objectives. In *ACM Designing Interactive Systems,* **DIS** 2020 (Contribution order)

- Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2019

- Matthew Joseph, Janardhan Kulkarni, Jieming Mao, and Zhiwei Steven Wu. Locally private Gaussian estimation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2019

- Yahav Bechavod, Katrina Ligett, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. Equal Opportunity in Online Classification with Partial Feedback. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2019

- Arindam Banerjee, Qilong Gu, Vidyashankar Sivakumar, and Zhiwei Steven Wu. Random quadratic forms with dependence: applications to restricted isometry and beyond. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems,* **NeurIPS**, 2019

- Seth Neel, Aaron Roth, and Zhiwei Steven Wu How to Use Heuristics for Differential Privacy. In *Proceedings of The 60th Annual IEEE Symposium on Foundations of Computer Science,* **FOCS**, 2019

- Alekh Agarwal, Miroslav Dudik, Zhiwei Steven Wu Fair Regression: Quantitative Definitions and Reduction-based Algorithms In *Proceedings of the 36th International Conference on Machine Learning,* **ICML**, 2019

- Aaron Schein, Zhiwei Steven Wu, Alexandra Schofield, Mingyuan Zhou, and Hanna Wallach. Locally private bayesian inference for count models. In *Proceedings of the 36th International Conference on Machine Learning,* **ICML**, 2019. (Contributional order)

- Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu Orthogonal Random Forest for Causal Inference In *Proceedings of the 36th International Conference on Machine Learning,* **ICML**, 2019

- Guy Aridor, Kevin Liu, Aleksandrs Slivkins, Zhiwei Steven Wu. The Perils of Exploration under Competition: A Computational Modeling Approach In *The 20th ACM conference on Economics and Computation* **EC**, 2019

- Nicole Immorlica, Jieming Mao, Alex Slivkins, and Zhiwei Steven Wu. Bayesian Exploration with Heterogeneous Agents In *The Web Conference 2019* **TheWebConf (Oral presentation)**, 2019

- Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the second Annual ACM Conference on Fairness, Accountability, and Transparency,* **FAccT**, 2019

- Sampath Kannan, Jamie Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems,* **NeurIPS (Selected for a Spotlight Presentation: Top 2% of submissions)**, 2018

- Manish Raghavan, Aleksandrs Slivkins, Jenn Wortman Vaughan, and Zhiwei Steven Wu. The unfair externalities of exploration and how data diversity helps exploitation. In *The 31st Annual Conference on Learning Theory,* **COLT**, 2018

- Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning,* **ICML**, 2018

- Akshay Krishnamurthy, Zhiwei Steven Wu, and Vasilis Syrgkanis. Semiparametric Contextual Bandits. In *Proceedings of the 35th International Conference on Machine Learning,* **ICML**, 2018. (Contributional order)

- Jinshuo Dong, Aaron Roth, Zachary Schutzman, Bo Waggoner, and Zhiwei Steven Wu. Strategic classification from revealed preferences. In *The 19th ACM conference on Economics and Computation* **EC**, 2018

- Yishay Mansour, Aleksandrs Slivkins, and Zhiwei Steven Wu. Competing bandits: Learning in competition. In *Proceedings of the 2018 ACM Conference on Innovations in Theoretical Computer Science,* **ITCS**, 2018

- Katrina Ligett, Seth Neel, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. Accuracy first: Selecting a differential privacy level for accuracy-constrained ERM. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems,* **NIPS**, 2017

- Sampath Kannan, Michael Kearns, Jamie Morgenstern, Mallesh M. Pai, Aaron Roth, Rakesh V. Vohra, and Zhiwei Steven Wu. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation,* **EC**, 2017

- Aaron Roth, Aleksandrs Slivkins, Jonathan Ullman, and Zhiwei Steven Wu. Multidimensional dynamic pricing for welfare maximization. In *Proceedings of the 2017 ACM Conference on Economics and Computation,* **EC**, 2017. Invited to the special issue of ACM Transactions on Economics and Computation for EC'17

- Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Meritocratic fairness for cross-population selection. In *Proceedings of the 34th International Conference on Machine Learning,* **ICML**, 2017

- Michael Kearns and Zhiwei Steven Wu. Predicting with distributions. In *Proceedings of the 30th Conference on Learning Theory,* **COLT**, 2017

- Shahin Jabbari, Ryan Rogers, Aaron Roth, and Zhiwei Steven Wu. Learning from rational behavior: Predicting solutions to unknown linear programs. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems,* **NIPS**, 2016

- Yishay Mansour, Aleksandrs Slivkins, Vasilis Syrgkanis, and Zhiwei Steven Wu. Bayesian exploration: Incentivizing exploration in bayesian games. In *Proceedings of the 2016 ACM Conference on Economics and Computation,* **EC**, 2016

- Aaron Roth, Jonathan Ullman, and Zhiwei Steven Wu. Watch and learn: optimizing from revealed preferences feedback. In *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing,* **STOC**, 2016

- Rachel Cummings, Katrina Ligett, Kobbi Nissim, Aaron Roth, and Zhiwei Steven Wu. Adaptive learning with robust generalization guarantees. In *Proceedings of the 29th Conference on Learning Theory,* **COLT**, 2016

- Paul W. Goldberg, Francisco J. Marmolejo Cossío, and Zhiwei Steven Wu. Logarithmic query complexity for approximate nash computation in large games. In *Proceedings of the 9th International Symposium on Algorithmic Game Theory,* **SAGT**, 2016. Invited to the special issue of Theory of Computing Systems for SAGT'16

- Justin Hsu, Zhiyi Huang, Aaron Roth, and Zhiwei Steven Wu. Jointly private convex programming. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms,* **SODA**, 2016

- Rachel Cummings, Katrina Ligett, Jaikumar Radhakrishnan, Aaron Roth, and Zhiwei Steven Wu. Coordination complexity: Small information coordinating large populations. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science,* **ITCS**, 2016

- Rachel Cummings, Michael Kearns, Aaron Roth, and Zhiwei Steven Wu. Privacy and truthful equilibrium selection for aggregative games. In *Proceedings of the 11th International Conference on Web and Internet Economics,* **WINE**, 2015

- Ryan Rogers, Aaron Roth, Jonathan Ullman, and Zhiwei Steven Wu. Inducing approximately optimal flow using truthful mediators. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation,* **EC**, 2015

- Rachel Cummings, Katrina Ligett, Aaron Roth, Zhiwei Steven Wu, and Juba Ziani. Accuracy for sale: Aggregating data with a variance constraint. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science,* **ITCS**, 2015

- Sampath Kannan, Jamie Morgenstern, Aaron Roth, and Zhiwei Steven Wu. Approximately stable, school optimal, and student-truthful many-to-one matchings (via differential privacy). In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms,* **SODA**, 2015

- Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu. Dual query: Practical private query release for high dimensional data. In *Proceedings of the 31th International Conference on Machine Learning,* **ICML**, 2014

- Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. In *Proceedings of the 46th ACM Symposium on Theory of Computing,* **STOC**, 2014. Invited to the special issue of ACM Transactions on Economics and Computation for STOC'14 (declined)

SURVEYS/ NEWSLETTERS

- Aaron Roth, Jonathan Ullman, and Zhiwei Steven Wu. Watch and learn: optimizing from revealed preferences feedback. *SIGecom Exchanges*, 2015

ADVISING

Ph.D. Students

- Gokul Swamy (co-advised by Drew Bagnell). 2020–present

- Luke Guerdan (co-advised by Ken Holstein). 2021–present

- Terrance Liu. 2022–present

- Jingwu Tang (co-advised with Fei Fang). 2023–present
- Anusha Sinha (co-advised with Hoda Heidari). 2023–present
- Jiahao Zhang. 2024–present

- Keegan Harris. 2020-2025. Now a post-doc at Berkeley.
- Justin Whitehouse. 2021–2024. Now a post-doc at Stanford.
- Logan Stapleton 2019-2024. Now on the Vassar College CS Faculty.
- Dung Ngo. 2019–2024. Now at JP Morgan AI Research.
- Hao-Fei Cheng. 2019–2021. Now at Amazon.
- Giuseppe Vietri. 2018-2023. Now at Amazon.

Post-docs

- Pratiksha Thaker (co-advised by Virginia Smith). 2022–present
- Shuran Zheng. 2023. Now on the IIIS faculty of Tsinghua University

REU Students
Xinyan Vicky Hu, Allen Marquez, Manish Nagireddy, Harry Tian, Grace Tian, Diana Qing, Jocelyn Chou.

Thesis Committee Member
Vidyashankar Sivakumar (UMN), Qilong Gu (UMN), Anthony Zhenhuan Zhang (UMN), Haoran Sun (UMN), Gautam Goel (Caltech), Yingxue Zhou (UMN), Zinan Lin (CMU), Chris Jung (Penn)

Teaching

Carnegie Mellon University — Pittsburgh, PA
Instructor
- 17740: Algorithmic Foundations of Interactive Learning — Spring 2025
- 17757: Modern Techniques in Uncertainty Quantification — Spring 2024
- 18734 / 17731: Foundations of Privacy — Fall 2021, 2022
- 05-318 / 05-618: Human AI Interaction — Fall 2020, 2021
- 17880: Algorithms For Private Data Analysis — Spring 2021

University of Minnesota — Twin-Cities, MN
Instructor
- CSCI 5525: Machine Learning — Fall 2019, Spring 2020
- CSCI 8980: The Algorithmic Foundations of Data Privacy — Fall 2018

Bard Prison Initiative, — Eastern Correctional Facility, NY
Math tutor: gave math tutorials to inmates — Spring 2011

Research Funding

- Role: Sole-PI
Funding Agency: NSF
(Awarded) "CAREER: New Frontiers of Private Learning and Synthetic Data"
3/1/2024–
Award amount: $680,000

- Role: Sole-PI
  Funding Agency: Department of Air Force Office of Small Business
  (Awarded) "Mission Assurance for Autonomy and Reinforcement Learning" (partnered with Edge Case Research)
  10/1/2024–1/31/2026
  Award amount to CMU: $720,000

- Role: Co-PI
  (PI: Lei Li, co-PI: Daphne Ippolito)
  Funding Agency: CMU CyLab
  (Awarded) "Security Attacks and Robust Defense against LLMs"
  Award amount: $50,000

- Role: Co-PI
  (PI: Gauri Joshi)
  Funding Agency: CMU CyLab
  (Awarded) "Synergies and Trade-offs in Private Federated Learning"
  Award amount: $50,000

- Role: Co-PI
  (PI: Jonathan Ullman (Northeastern), co-PIs: Adam Smith (BU))
  Funding Agency: NSF
  (Awarded) "SaTC: CORE: Medium: Private Model Personalization"
  4/15/2023–
  Award amount: $300,000 for CMU

- Role: Sole-PI
  Funding Agency: FutureEnterprise@CyLab
  (Awarded) "Evaluating Large Language Models' Privacy Risks with Privacy Attacks"
  Total award amount: $60,000

- Role: Co-PI
  Funding Agency: Adobe
  (Awarded) "A New Approach to Build Individual Level Model with Walled Gardens Data"
  Award Amount: $25,000 for CMU

- Role: Sole-PI
  Funding Agency: Google
  (Awarded) "Label Inferential Privacy"
  Total award amount: $50,000

- Role: Sole-PI
  Funding Agency: Google Collabs Award
  (Awarded) "Private Synthetic Data Sharing via Generative Transfer Learning"
  Total award amount: $80,000

- Role: Sole-PI
  Funding Agency: FutureEnterprise@CyLab
  (Awarded) "Differentially Private Synthetic Data Generation"
  Total award amount: $60,000

- Role: Sole-PI
  Funding Agency: JP Morgan Faculty Research award
  (Awarded) "Advancing Privacy-Preserving Data Sharing with Synthetic Data Generation"
  Total award amount: $110,000

- Role: Co-PI
  (PI: Haiyi Zhu, Co-PI: Ken Holstein)
  Funding Agency: CMU Block Center
  (Awarded) "Supporting Effective AI-Augmented Decision-Making in Content Moderation"
  Total award amount: $80,000

- Role: Co-PI
  (PI: Ken Holstein, Co-PI: Haiyi Zhu)
  Funding Agency: Center for Advancing Safety of Machine Intelligence (CASMI), Northwestern University
  (Awarded) "Supporting Effective AI-Augmented Decision-Making in Social Contexts"
  Total award amount: $275,000

- Role: Co-PI
  (PI: Virginia Smith)
  Funding Agency: Apple Research Award
  (Awarded) "Pushing the Privacy-Utility Frontier with MTL"
  Total award amount: $99,072

- Role: Co-PI
  (PI: Virginia Smith)
  Funding Agency: Meta Research Award
  (Awarded) "Private Multi-Task Learning"
  Total award amount: $100,000

- Role: Collaborator (PI: Anusha Sinha, Team Members: Nathan VanHoudnos, Sumanyu Gupta, CMU Collaborators: Matt Frederickson, Hoda Heidari)
  Funding Agency: Software Engineering Institute
  (Awarded) "Leveraging Adversarial Machine Learning Techniques to Perform Query- Access Fairness Evaluations"
  Award amount for CMU team: $200,000

- Role: Sole-PI
  Funding Agency: Cisco
  (Awarded) "Foundations for Private Synthetic Data Generation"
  Total award amount: $150,000

- Role: Sole-PI
  Funding Agency: The Okawa Foundation
  (Awarded) "Enabling the Next Generation of Privacy-Preserving Machine Learning"
  Total award amount: $10,000

- Role: PI
  (co-PI: Virginia Smith)
  Funding Agency: CMU ATLAS Moonshot Award
  (Awarded) "Private and Fair Federated Learning with Applications to Energy Control"
  Total award amount: $87,077

- Role: Co-PI
  (PI: Hoda Heidari, co-PIs: Haiyi Zhu)
  Funding Agency: Meta
  (Awarded) "A Tool to Study the Efficacy of Fairness Algorithms on Specific Bias Types"
  Total award amount: $100,000

- Role: Co-PI
  (PI: Jonathan Ullman (Northeastern), co-PIs: Roxana Geambasu(Columbia), Alina Oprea (North-

eastern), Adam Smith (BU))
Funding Agency: NSF
(Awarded) "SaTC: CORE: Small: Foundations for the Next Generation of Private Learning Systems"
10/1/2021–9/30/2022
Total award amount: $549,786 total, $100,000 for CMU

- Role: Co-PI
(PI: Ken Holstein, co-PI: Alexandra Chouldechova, Emily Putnam-Hornstein (UNC), Haiyi Zhu)
Funding Agency: CMU Block Center
(Awarded) "Supporting Responsible Use of Algorithmic Decision Support in Child Welfare"
Total award amount: $40,000

- Role: Co-PI
(PI: Haiyi Zhu (CMU), co-PIs: Gord Burtch (BU), Yanhua Li (WPI), Min Kyung Lee (UT Austin))
Funding Agency: NSF
(Awarded) "SCC-IRG Track 1: Empowering and Enhancing Workers Through Building A Community-Centered Gig Economy"
10/1/2020–9/30/2023
Total award amount: $1,997,764 total

- Role: PI
(Co-PIs: Alexandra Chouldechova (CMU), Min Kyung Lee (UT Austin), Haiyi Zhu (CMU))
Funding Agency: NSF and Amazon
(Awarded) "FAI: Advancing Fairness in AI with Human-Algorithm Collaborations"
1/1/2020–12/31/2022
Total award amount: $1,037,000 total, $338,286 for UMN

- Role: UMN PI
(PI: Haiyi Zhu (CMU), co-PIs: Mark Snyder, Loren Terveen)
Funding Agency: NSF
(Awarded) "EAGER: AI-DCL: Capture, Explain and Negotiate the Inherent Trade-offs in Machine Learning Algorithms"
10/01/2019–9/30/2021
Total award amount: $295,713, $193,267 for UMN

- Role: Co-PI
(PI: Haiyi Zhu (CMU), co-PIs: Mark Snyder, Loren Terveen)
Funding Agency: NSF
(Awarded) "CHS: Small: Incorporating and Balancing Stakeholder Values in Algorithm Design"
8/1/2019–7/31/2022
Total award amount: $500,000, $243,941 for UMN

- Role: PI
(Co-PI: Yuvraj Agarwal (CMU))
Funding Agency: CMU CyLab
(Awarded) "Enabling Privacy-Preserving IoT Apps and Data Analytics"
Awarded in 2021
Total award amount: $50,000

- Role: Co-PI
(PI: Ding Zhao (CMU))
Funding Agency: Mobility21 University Transportation Center
(Awarded) "Towards a Smart, Safe, and Sustainable Sidewalk: A Quantitative Analysis on How Sidewalk Infrastructure Affect Personal Delivery Devices"

Awarded in 2021
Total award amount: $99,997

- Role: PI
(Co-PI: Haiyi Zhu (CMU))
Funding Agency: Facebook
(Awarded) "Promoting Diversity in Peer Production through Mechanism Design"
Awarded in 2019
Total award amount: $50,000, $50,000 for UMN

- Role: Sole PI
Funding Agency: J.P. Morgan
(Awarded) "Preventing Unfair Discrimination in Interactive Learning"
3/4/2019–3/3/2021
Total award amount: $155,034

- Role: Sole PI
Funding Agency: Google
(Awarded) "Incentive-Aware Learning via Algorithmic Stability"
Awarded in 2019
Total award amount: $50,000

- Role: Sole PI
Funding Agency: Mozilla
(Awarded) "DP-Fathom: Private, Accurate, and Communication-Efficient"
Awarded in 2019
Total award amount: $25,000

Service and Outreach

Organizer of Recent Developments in Research on Fairness. The Simons Institute for the Theory of Computing, Berkeley, CA. July 8-10, 2019.

Program Committee: ALT 2021, COLT 2023 (SPC), AISTATS 2021 (Area Chair), NeurIPS 2020, 2021 (Area Chair), ICML 2022, 2020 (Area Chair), ICLR 2020, 2021, 2022 (Area Chair), SODA 2022, ITCS 2022, WWW 2020, EC 2020, TPDP 2019, EC 2019, FAccT 2019, 2021, 2023 (Area Chair), AAAI 2019, EC 2018, WWW 2018, ICML 2018, ICML 2017.

Conference Reviewer: STOC 2019, SODA 2018, ITCS 2018, NIPS 2017, ALT 2017, FOCS 2017, EC 2017, ICALP 2017, SODA 2017, COLT 2016, ESA 2016, TEAC, WINE 2015, ISAAC 2015, NIPS 2015, FOCS 2015, STOC 2015, FOCS 2014, WINE 2014, WINE 2013

Journal Reviewer: Proceedings of the National Academy of Sciences (PNAS), Machine Learning, Journal of Machine Learning Research, Operations Reserach, Journal of Privacy and Confidentiality, Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Information Theory.

Selected Talks

What can differential privacy do for advertising?
- Google fourth annual Ads Future of Technology Conference (FACT), Aug 2022

Choosing Epsilon for Differential Privacy, Adaptively
- Google Federated Learning Seminar, Aug 2022

Policy impacts of statistical uncertainty and privacy
- Fields Institute Workshop on Differential Privacy and Statistical Data Analysis, July 2022

Leveraging Public Data for Private Synthetic Data Generation
- 2022 Institute of Mathematical Statistics (IMS) Annual Meeting, June 2022

Of Moments and Matching: Trade-offs and Treatments in Imitation Learning
- Simons Institute Workshop on Adversarial Approaches in Machine Learning, Feb 2022
- Baidu Research AI Colloquium, June 2022

Panel Discussion: Differential privacy and its disparate impacts
- The Third AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI-22), Feb 2022

Leveraging Strategic Interactions for Causal Discovery
- StratML workshop at NeurIPS'21, Dec 2021

Recent Advances in Private Synthetic Data Generation
- The 14th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2021), December 2021
- The 2021 FCSM Research and Policy Conference, Nov 2021
- MIT AIPF Health Workshop, Feb 2022

Private Multi-Task Learning
- Google's Federated Learning and Analytics Workshop, Nov 2021

A Geometric View on Private Gradient-Based Optimization
- Federated Learning One World Seminar (FLOW), March 2021
- Google TechTalks, March 2021

Involving Stakeholders in Building Fair ML Systems
- Foundations of Algorithmic Fairness Workshop, March 2021
- IDEAL Quarterly Theory Workshop: Algorithms and their Social Impact, March 2021
- Trustworthy ML Initiative (TrustML) Seminar, Feb 2021

Leveraging Heuristics in Private Synthetic Data Generation
- CMU Crypto/Applied crypto seminar, March 2021
- PPAI workshop 2021, Feb 2021
- Boston-area Data Privacy Seminar, Feb 2021

Differential Privacy Techniques Beyond Differential Privacy
- FOCS 2019 Workshop "A TCS Quiver", November 2019

Between Individual and Group Fairness
- DIMACS 30th Birthday Conference "Three Decades of DIMACS: The Journey Continues"

Locally Private Bayesian Inference for Count Models
- Simons Workshop on Privacy and the Science of Data Analysis, April 2019

How to Use Heuristics for Differential Privacy
- Simons Institute Seminar, Feb 2019
- IMA Workshop: Recent Themes in Resource Tradeoffs: Privacy, Fairness, and Robustness, June 2019

Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness
- Google Research Seminar, April 2018
- CalTech Theory Seminar, March 2018

Privacy-Preserving GANs Support Clinical Data Sharing
- Microsoft Research-NYC tea talk, March 2019
- Banff workshop on "Mathematical Foundations of Data Privacy", May 2018

A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem
- Rutgers/DIMACS Theory of Computing Seminar, Oct 2017
- UMass Machine Learning and Friends Lunch (MLFL), Nov 2017

Differential Privacy: A Rigorous Notion for Data Privacy
- Muhlenberg College Math/CS Colloquium, May 2017
- Carleton College CS Tea Talk, Oct 2019

Leveraging No-Regret Algorithms in Private Data Analysis
- Princeton CS theory lunch, Feb 2017

Social Norms for Data-Driven Algorithms: Privacy, Incentive-Compatibility and Fairness
- SIGAI CNC, Boston, MA, Oct 2016
- NY Area Theory Day, New York, NY, Dec 2016

Adaptive Data Analysis and Differential Privacy
- Guest Lecture in the course Computational Learning Theory at UPenn

Adaptive Learning with Robust Generalization Guarantees
- COLT, New York City, June 2016

Coordination Complexity: Small Information Coordinating Large Populations
- Northeastern University Theory Seminar, January 2016
- UPenn Theory Lunch, September 2015
- University of Hong Kong, Theory Seminar, December 2015

Bayesian Exploration: Incentivizing Exploration in Bayesian Games
- Harvard EconCS Seminar, September 2016
- EC, Maastricht, July 2016
- Microsoft Research NYC Tea Talk, July 2015

Watch and Learn: Optimizing from Revealed Preferences Feedback
- STOC, Cambridge, June 2016
- Caltech Theory Lunch, April 2015
- The First Workshop on Algorithmic Game Theory and Data Science, Portland, June 2015

Inducing Approximately Optimal Flow Using Truthful Mediators
- EC, Portland, June 2015

Privacy for the Protected (Only)
- Columbia CS Seminar, Dec. 2016
- Cornell Theory Seminar, Nov. 2016
- Workshop on The Theory of Bringing Privacy into Practice, Pasadena, April 2015

Privacy and Truthful Equilibrium Selection in Aggregative Games
- UPenn Theory Lunch, September 2014
- WINE, December 2015

Dual Query: Practical Private Query Release for High Dimensional Data
- ICML, Beijing, June 2014

Private Matchings and Allocations

- STOC, New York, June 2014
- UPenn Theory Lunch, May 2014