Yuanshun (Kevin) Yao

Email: kevinyaowork@gmail.com
[Website] [Google Scholar]

Education

09/2017-08/2020 **The University of Chicago** Chicago, IL
Ph.D. in Computer Science; Area: AI safety
Thesis: Practical Backdoor Attacks and Defenses in Deep Learning Systems
Advisor: Prof. Ben Y. Zhao and Prof. Heather Zheng

09/2015-06/2017 **University of California, Santa Barbara** Santa Barbara, CA
Ph.D. in Computer Science (transferred to UChicago)

09/2011-05/2015 **University of Minnesota – Twin Cities** Minneapolis, MN
B.S. in Computer Science, Mathematics, Statistics (triple major)
Area: Spatio-temporal data mining; Advisor: Prof. Vipin Kumar

Work

09/2020-present **ByteDance Research** San Jose, CA
Research Scientist
- Research in trustworthy large language models (LLMs), i.e. LLM safety, privacy, general alignment (e.g. unlearning, red teaming, watermark, hallucination, etc.)
- Research in trustworthy AI, e.g. AI fairness, AI explainability, AI privacy
- Implement responsible AI principles in Tiktok production systems

06/2018-09/2018 **Google** Sunnyvale, CA
Internship in Security & Privacy (Safe Browsing) team
- Trained machine learning models to detect mobile malware on a global scale
- Improved model interpretability for malware manual analysts
- Diagnosed feature engineering and improved model training pipeline

06/2017-09/2017 **Google** Mountain View, CA
Internship in Google Shopping team
- Trained deep learning models to recognize and localize commodities in images
- Improved model performance with online hard example mining
- Implemented a prototype of deep learning based image retrieval system

06/2013-05/2014 **IBM** Princeton, NJ
Software Engineering Intern
- Worked on IBM InfoSphere Optim Test Data Management
- Prototyped a machine learning system to predict customer behaviors

Preprint

[1] **Yuanshun Yao**, Xiaojun Xu, Yang Liu. "Large Language Model Unlearning." https://arxiv.org/abs/2403.10553

[2] Sijia Liu, **Yuanshun Yao**, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun Xu, Yuguang Yao, Hang Li, Kush R Varshney, Mohit Bansal, Sanmi Koyejo, Yang Liu. "Rethinking Machine Unlearning for Large Language Models." https://arxiv.org/abs/2402.08787

[3] Yang Liu*, **Yuanshun Yao***, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, Hang Li. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." https://arxiv.org/abs/2308.05374

[4] **Yuanshun Yao**, Yang Liu. "On the Cause of Unfairness: A Training Sample Perspective." https://arxiv.org/abs/2306.17828

[5] Xiaojun Xu, **Yuanshun Yao**, Yang Liu. "Learning to Watermark LLM-generated Text via Reinforcement Learning." https://arxiv.org/abs/2403.10553

[6] Jiaheng Wei, **Yuanshun Yao**, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, Yang Liu. "Measuring and Reducing LLM Hallucination without Gold-Standard Answers via Expertise-Weighting." https://arxiv.org/abs/2402.10412

[7] Hongyi Guo, **Yuanshun Yao**, Wei Shen, Jiaheng Wei, Xiaoying Zhang, Zhaoran Wang, Yang Liu. "Human-Instruction-Free LLM Self-Alignment with Limited Samples." https://arxiv.org/abs/2401.06785

[8] Wei Shen, Xiaoying Zhang, **Yuanshun Yao**, Rui Zheng, Hongyi Guo, Yang Liu. "Improving Reinforcement Learning from Human Feedback Using Contrastive Rewards." https://arxiv.org/abs/2403.07708

[9] **Yuanshun Yao**, Yang Liu, Chong Wang, Hang Li. "Measuring Training Influence in Recommender Systems via Surrogate Function"

[10] **Yuanshun Yao**, Chong Wang, Hang Li. "Learning to Counterfactually Explain Recommendations."

[11] Jiankai Sun, Xin Yang, **Yuanshun Yao**, Junyuan Xie, Di Wu, Chong Wang. "Differentially Private AUC Computation in Vertical Federated Learning."

[12] Xin Yang, Jiankai Sun, **Yuanshun Yao**, Junyuan Xie, Chong Wang. "Differentially Private Label Protection in Split Learning."

[13] Jiankai Sun, Xin Yang, **Yuanshun Yao**, Chong Wang. "Label Leakage and Protection from Forward Embedding in Vertical Federated Learning."

[14] Jiankai Sun, Xin Yang, **Yuanshun Yao**, Aonan Zhang, Weihao Gao, Junyuan Xie, Chong Wang. "Vertical Federated Learning without Revealing Intersection Membership."

Publication

[1] Tongxin Yin, Jean-François Ton, Ruocheng Guo, **Yuanshun Yao**, Mingyan Liu, and Yang Liu. "Fair Classifiers that Abstain without Harm." Proceedings of *International Conference on Learning Representations* (ICLR), May 2024.

[2] **Yuanshun Yao**, Xiaojun Xu, Yang Liu. "Large Language Model Unlearning." Workshop of Socially Responsible Language Modelling Research (SoLaR) at NeurIPS 2023.

[3] Yang Liu*, **Yuanshun Yao***, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, Hang Li. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." Workshop of Socially Responsible Language Modelling Research (SoLaR) at NeurIPS 2023.

[4] **Yuanshun Yao**, Yang Liu. "Understanding Unfairness via Training Concept Influence." Workshop of Data-centric Machine Learning Research (DMLR) at ICML 2023.

[5] Zhaowei Zhu*, **Yuanshun Yao***, Jiankai Sun, Hang Li, Yang Liu. "Weak Proxies are Sufficient and Preferable for Fairness with Missing Sensitive Attributes." Proceedings of *International Conference on Machine Learning* (ICML), July 2023.

[6] Zhujun Xiao, Jenna Cryan, **Yuanshun Yao**, Yi Hong Gordon Cheo, Yuanchao Shu, Stefan Saroiu, Ben Y. Zhao, Haitao Zheng. ""My face, my rules": Enabling Personalized Protection against Unacceptable Face Editing." Proceedings of *Privacy Enhancing Technologies Symposium* (PETS), July 2023.

[7] Jiankai Sun, Xin Yang, **Yuanshun Yao**, Junyuan Xie, Di Wu, Chong Wang "DPAUC: Differentially Private AUC Computation in Federated Learning." Proceedings of *AAAI Conference on Artificial Intelligence* (AAAI), February 2023.

[8] Shangyu Xie, Xin Yang, **Yuanshun Yao**, Tianyi Liu, Taiqing Wang, Jiankai Sun. "Label Inference Attack against Regression Model under Split Learning." Proceedings of *AAAI workshop on Privacy Preserving Artificial Intelligence*, February 2023.

[9] Mimee Xu, Jiankai Sun, Xin Yang, **Yuanshun Yao**, Chong Wang. "Netflix and Forget: Fast Severance From Memorizing Training Data in Recommendations." Proceedings of NeurIPS ML Safety Workshop, November 2022.

[10] Ruihan Wu, Xin Yang, **Yuanshun Yao**, Jiankai Sun, Tianyi Liu, Kilian Q Weinberger, Chong Wang. "Differentially Private Multi-Party Data Release for Linear Regression." Proceedings of *Conference on Uncertainty in Artificial Intelligence* (UAI), August 2022.

[11] Jiankai Sun, **Yuanshun Yao**, Weihao Gao, Junyuan Xie, Chong Wang. "Defending against Reconstruction Attack in Vertical Federated Learning." Proceedings of *International Workshop on Federated Learning for User Privacy and Data Confidentiality* (Conjunction with ICML), July 2021.

[12] Emily Wenger, Josephine Passanati, Arjun Nitin Bhagoji, **Yuanshun Yao**, Haitao Zheng, Ben Y. Zhao. "Backdoor Attacks Against Deep Learning Systems in the Physical World." Proceedings of *IEEE Computer Vision and Pattern Recognition* (CVPR), June 2021.

[13] **Yuanshun Yao**, Huiying Li, Haitao Zheng and Ben Y. Zhao. "Latent Backdoor Attacks on Deep Neural Networks." Proceedings of *ACM Conference on Computer and Communications Security* (CCS), London, UK, November 2019.

[14] Bolun Wang, **Yuanshun Yao**, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng and Ben Y. Zhao. "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks." Proceedings of *IEEE Symposium on Security and Privacy* (IEEE S&P), San Francisco, CA, May 2019.

[15] Bolun Wang, **Yuanshun Yao**, Bimal Viswanath, Haitao Zheng and Ben Y. Zhao. "With Great Training Comes Great Vulnerability: Practical Attacks against Transfer Learning." Proceedings of *USENIX Security Symposium* (USENIX Security), Baltimore, MD, August 2018.

[16] **Yuanshun Yao**, Zhujun Xiao, Bolun Wang, Bimal Viswanath, Haitao Zheng and Ben Y. Zhao. "Complexity vs. Performance: Empirical Analysis of Machine Learning as a Service." Proceedings of *ACM SIGCOMM Internet Measurement Conference* (IMC), London, UK, November 2017.

[17] **Yuanshun Yao**, Bimal Viswanath, Jenna Cryan, Haitao Zheng and Ben Y. Zhao. "Automated Crowdturfing Attacks and Defenses in Online Review Systems." Proceedings of *ACM Conference on Computer and Communications Security* (CCS), Dallas, TX, October 2017.

[18] Yanzi Zhu, **Yuanshun Yao**, Ben Y. Zhao and Haitao Zheng. "Object Recognition and Navigation using a Single Networking Device." Proceedings of *International Conference on Mobile Systems, Applications, and Services* (MobiSys), Niagara Falls, NY, June 2017.

[19] Zhijing Li, Ana Nika, Xinyi Zhang, Yanzi Zhu, **Yuanshun Yao**, Ben Y. Zhao and Haitao Zheng. "Identifying Value in Crowdsourced Wireless Signal Measurements." Proceedings of *World Wide Web Conference* (WWW), Perth, Australia, April 2017.

[20] Xi C. Chen, **Yuanshun Yao**, Sichao Shi, Snigdhansu Chatterjee, Vipin Kumar and James H. Faghmous. "A General Framework to Increase the Robustness of Model-based Change Point Detection Algorithms to Outliers and Noise." Proceedings of *SIAM International Conference on Data Mining* (SDM), Miami, FL, May 2016.

[21] James H. Faghmous, Ivy Frenger, **Yuanshun Yao**, Robert Warmka, Aron Lindel and Vipin Kumar. "A Daily Global Mesoscale Ocean Eddy Dataset From Satellite Altimetry." *Scientific Data 2, Nature Publishing Group*, June 2015.

Academic Service

Conference Reviewer	ICML'20, CVPR'21, CSCW'21, ICML'22, NeurIPS'22, AISTATS'22, AAAI'22, ICML'23, UAI'23, KDD'23, NeurIPS'23, ICLR'24, ICML'24
Journal Reviewer	Transactions on Pattern Analysis and Machine Intelligence, Transactions on Dependable and Secure Computing, Neural Computing and Applications, Special Issue On Pre-Trained Large Language Models - IEEE Transactions on Big Data, Data-centric Machine Learning Research
Program Committee	IJCAI'23, FAccT'23, DMLR'23, SoLaR'23, TheWebConf Industry Track'24

Award

2020	Siebel Scholarship
2018	UU Fellowship, University of Chicago
2011-2015	Deans list, University of Minnesota
2011-2015	Maroon Global Excellence Scholarship, University of Minnesota
2014	Undergraduate Research Opportunity Program Grant, University of Minnesota
2014	NSF Student Travel Grant

Teaching

Spring 2016	CS 16 Problem Solving with Computers I, University of California, Santa Barbara
Winter 2016	CS 16 Problem Solving with Computers I, University of California, Santa Barbara
Fall 2015	CS 8 Introduction to Computer Science, University of California, Santa Barbara
Spring 2015	Csci 2033 Elementary Computational Linear Algebra, University of Minnesota
Spring 2013	Math 5651 Basic Theory of Probability and Statistics, University of Minnesota

Media Coverage

12/16/2017	Artificial intelligence is killing the uncanny valley and our grasp on reality. **Wired**.
10/16/2017	Could AI be the future of fake news and product reviews? **Scientific American**.
09/05/2017	Many people can't tell the difference between Yelp reviews written by an AI and a human. Can you? **Forbes**.
09/01/2017	AI writes Yelp reviews that pass for the real thing. **Engadget**.
08/31/2017	AI trained on Yelp data writes fake restaurant reviews "indistinguishable" from real deal. **The Verge**.
08/31/2017	Robots learned how to write fake Yelp reviews like a human. **New York Post**.
08/30/2017	AI writes believable fake Yelp reviews. **Nvidia Developer**.
08/30/2017	Restaurant reviews could be generated by AI without you noticing.**Fortune**.
08/29/2017	Researchers taught AI to write totally believable fake reviews, and the implications are terrifying. **Business Insider**.