Chaowei Xiao
Arizona State University

Dear Members and Search Committee,

I am writing to express my strong interest in the Assistant Professor position at the School of Cybersecurity and Privacy at the Georgia Institute of Technology. I am confident that my expertise, specifically in the area of the intersection between cyber security and machine learning (ML), will make me an excellent fit for this role. I am currently a new Assistant Professor at the Center for Cybersecurity and Trusted Foundations at Arizona State University starting from August 2022, but I am seeking a new position due to an emerging family issue.

My academic training has prepared me to be a strong researcher. My research starts with studying security in ML models and later expands to Socially Responsible Machine Learning (SRML) with broader considerations, including robustness, privacy, social bias, explainability, and their interconnection. My goal is to establish principled approaches to build SRML models and explore their applications in cyber security, vision, robotics, IoT/CPS (e.g., autonomous driving), and healthcare. Within this context, I tackle the security threats of current model learning models by (1) broadening the scope of security threats beyond traditional L_p regime by discovering new non-L_p threats, (2) demonstrating severe physical consequences of adversarial threats in real-world robotic and cyber-physical applications such as autonomous driving systems (e.g., Baidu Apollo), voice controllable systems (e.g., Alexa), robotic and IoT systems and (3) proposing principled purification-based solutions to defend against various unseen threats from both empirical and theoretical perspectives. For SRML, I take a two-pronged approach to co-design ML security and SRML. On one hand, I apply ML security principles for the SRML domain to improve model robustness against distribution shift, discover new privacy attacks, use explainability (i.e., sharply value) for detecting security threats, and reduce large language models' social bias. On the other hand, I work on both theory and practice to propose principled solutions for SRML from the perspectives of security, robustness, privacy, explainability, and social bias, respectively. In addition to SRML, I also work on data-driven cyber-security and robust malware detection.

My research has brought significant impacts to industry, academia, and society. My research has been implemented by companies (NVIDIA, IBM, etc.) to protect users against adversarial attacks, deployed in the second-largest airport in the world for baggage surveillance, and triggered 20+ autonomous driving companies to investigate newly-discovered vulnerabilities. My research results in 24 papers in top-tier machine learning conferences with several spotlights and oral presentations and 8 papers in top-tier security and system conferences. I won Best Paper Awards in MobiCom 2015, ESWN 2021, and ICLR 2022 SRML workshop, and ACM Gordon Bell Special Prize for HPC-Based COVID-19 Research. My work has led to a wide range of media coverage, as well as an exhibition at the London Science Museum.

In the future, I will continue to design SRML models and study SRML for computer security and social good. My research area is new and emerging: within four months at ASU, I have submitted 5 grants and received a $200,000 single PI grant. I believe my research could contribute to the strengths of your department and lead to successful collaborations.

I was also fortunate to serve as an instructor, guest lecturer, and research advisor. These professional experiences provided me abilities to teach and mentor students. I look forward to both teaching fundamental courses and designing new ones related to my research (e.g. SRML). Additionally, I would like to continue my commitment to strengthen diversity, equity, and inclusion in the future.

I am excited about the opportunity to contribute to the research and teaching mission of your school. Thank you for considering my application. I look forward to discussing this opportunity further.

Thank you in advance!

Chaowei Xiao

600 W Grove Pkwy, Apt 2039, Tempe, AZ, 85283
734-2392-561
xiaocw@asu.edu
http://xiaocw11.github.io/
Google Scholar: citations 5000+

Chaowei Xiao

Education

2015.8- **University of Michigan, Ann Arbor**.
2020.8 Ph.D. in Computer Science, EECS. Advisor: Prof. Mingyan Liu.
Dissertation: "Secure Learning in Adversarial Environments".

2011.8- **Tsinghua University**.
2015.7 B.S. in Computer Software.
B.S. in Economics, School of Economic and Management.

Employments

2022.8 **Assistant Professor**, *School of Computing and Augmented Intelligence*, Arizona State University.

2020.9 **Research Scientist**, *AI Algorithm Group*, NVIDIA Corporation.

Honors & Awards

2022 **ACM Gordon Bell Special Prize** for HPC-Based COVID-19 Research
2022 **Best Paper Award** in ICML SRML workshop
2021 **Best Paper Award** in ESWN
2019 Exhibition of "Physical Stop Sign" in London Science Museum.
2014 **Best Paper Award** in MobiCom
2014 First Prize in the 32nd Tsinghua Great Challenge Cup
2014 Intel Chinese Outstanding Student Scholarship
2013-2014 National Innovation and Entrepreneurship Training Program
2013 Tencent Chinese Outstanding Student Scholarship
2012-2014 First Class Scholarship for Overall Excellence

Three Representative Publications (* indicates equal contributions)

[3] **Chaowei Xiao***, Zhongzhu Chen*, Kun Jin*, Jiongxiao Wang*, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, Dawn Song. *DensePure: Understanding Diffusion Models towards Adversarial Robustness.* ICLR 2023. **TL;DR**: Based on our previous work [31], which, for the first time, discovered the effective diffusion model to defend against unseen adversarial examples without adversarial training, our work theoretically explained why and how diffusion models could enhance adversarial robustness.

[2] Yulong Cao*, Ningfei Wang*, **Chaowei Xiao***, Dawei Yang*, Jin Fang, Ruigang Yang, Alfred Chen, Mingyan Liu, Bo Li. *Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks*. IEEE Symposium on Security and Privacy 2021. **TL;DR**: We were the first to reveal vulnerabilities of the multi-sensor fusion perception framework of real-world autonomous driving systems (e.g., Baidu Apollo) by physically generating adversarial objects; Our research has triggered over 20 autonomous companies to start investigating these newly-discovered vulnerabilities.

[1] **Chaowei Xiao***, Jun-Yan Zhu*, Bo Li, Warren He, Mingyan Liu, Dawn Song. *Spatially Transformed Adversarial Examples*. ICLR 2018. **TL;DR**: Our work broadened the traditional Lp-based adversarial examples by discovering non-Lp-bounded adversarial examples; It opened up a new domain on non-Lp-bounded adversarial examples.

Publications (* indicates equal contributions)

Summary Total Citations: 5858, H-Index: 23 (Google Scholar [link (click)], as of Feb 1, 2023)

27 papers in commonly-recognized top-tier machine learning conferences (NeurIPS, ICML, ICLR, CVPR, ICCV, ECCV, CORL, ICDM, AAMAS, IJCAI)

8 papers in commonly-recognized top-tier security and system conferences (IEEE Security & Privacy, USENIX Security, ACM CCS, ACM MobiCom, IEEE INFOCOM)

1 paper in Top Survey journal (ACM Computing Survey) and 2 papers in Top journals (TMC, TDSC)

Students mentored by me are underlined

[39] **Chaowei Xiao***, Zhongzhu Chen*, Kun Jin*, Jiongxiao Wang*, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, Dawn Song. *DensePure: Understanding Diffusion Models towards Adversarial Robustness*. ICLR 2023.

[38] Shutong Wu, Jiongxiao Wang, Wei Ping, Weili Nie, **Chaowei Xiao**. *Defending against Adversarial Audio via Diffusion Model*. ICLR 2023

[37] Zichao Wang, Weili Nie, Zhuoran Qiao, **Chaowei Xiao** , Richard Baraniuk, Anima Anandkumar . *Retrieval-based Controllable Molecule Generation* ICLR 2023 (spotlight)

[36] Zhiyuan Yu, Yuanhaur Chang , Ning Zhang, **Chaowei Xiao**. *SMACK: Semantically Meaningful Adversarial Audio Attack* USENIX Security 2023

[35] Maxim Zvyagin*, Alexander Brace*, Kyle Hippe*, Yuntian Deng*, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, **Chaowei Xiao**, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, Arvind Ramanathan, GenSLMs: Genome-scale language models reveal SARS-CoV-2 evolutionary dynamics. (ACM Gordon Bell Special Prize)

[34] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, **Chaowei Xiao**. *Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models*. NeurIPS 2022

[33] Boxin Wang, Wei Ping,**Chaowei Xiao**, Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Bo Li, Anima Anandkumar, Bryan Catanzaro. *Exploring the Limits of Domain-Adaptive Training for Detoxifying Large-Scale Language Models* NeurIPS 2022

[32] Yulong Cao, Danfei Xu, Xinshuo Weng, Z. Morley Mao, Anima Anandkuma, **Chaowei Xiao**, Marco Pavone. *Robust Trajectory Prediction against Adversarial Attacks.* CORL 2022 (Oral presentation)

[31] Weili Nie, Brandon Guo, Yujia Huang, **Chaowei Xiao**, Arash Vahdat, Anima Anandkumar. *Diffusion Models for Adversarial Purification.* ICML 2022.

[30] Daquan Zhou, Zhiding Yu, Enze Xie, **Chaowei Xiao**, Anima Anandkumar, Jiashi Feng, Jose M Alvarez. *Understanding the robustness in vision transformers.* ICML 2022

[29] Yulong Cao, **Chaowei Xiao**, Anima Anandkumar, Danfei Xu, Marco Pavone. *AdvDO: Realistic Adversarial Attacks for Trajectory Prediction.* ECCV 2022

[28] Zhuowen Yuan, Fan Wu, Yunhui Long, **Chaowei Xiao**, and Bo Li. *SecretGen: Privacy Recovery on Pre-trained Models*. ECCV 2022

[27] Sina Mohseni, Zhiding Yu, **Chaowei Xiao**, and Jay Yadawa, Haotao Wang and Zhangyang Wang. *Taxonomy of Machine Learning Safety: A Survey and Primer.* ACM Computing Survey 2022. **TL; DR**: a comprehensive survey to discuss the problem and opportunities in machine learning safety.

[26] Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, **Chaowei Xiao**, Yuke Zhu, Song-Chun Zhu, Anima Anandkumar. *RelViT: Concept-guided vision transformer for visual relational reasoning.* ICLR 2022

[25] Jianwie Liu, **Chaowei Xiao**, Kaiyan Cui, Jinsong Han, Xian Xu, Kui Ren. *Behavior Privacy Perserving in RF Sensing.* IEEE Transcations on Dependable and Secure Computing, 2022

[24] Jianwei Liu, Yinghui He, **Chaowei Xiao**, Jinsong Han, Le Cheng, Kui Ren, Physical-World Attack towards WiFi-based Behavior Recognition, IEEE International Conference on Computer Communications (INFOCOM) 2022

[23] Xinlei Pan*, **Chaowei Xiao***, Warren He, Jian Peng, Mingjie Sun, Jinfeng Yi, Mingyan Liu, Bo Li, Dawn Song . *Characterizing Attacks on Deep Reinforcement Learning.* AAMAS 2022

[22] Jiachen Sun, Yulong Cao, Christopher Choy, Zhiding Yu, Anima Anandkumar, Z. Morley Mao, and **Chaowei Xiao**. *Adversarially Robust 3D Point Cloud Recognition Using Self-Supervisions.* NeurIPS 2021

[21] Haotao Wang, **Chaowei Xiao**, Jean Kossaifi, Zhiding Yu, Animashree Anandkumar, and Zhangyang Wang. *AugMax: Adversarial Composition of RandomAugmentations for Robust Training.* NeurIPS 2021

[20] Chen Zhu, Wei Ping, **Chaowei Xiao**, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, Bryan Catanzaro. *Efficient Transformers for Language and Vision.* NeurIPS 2021

[19] Mingjie Sun*,**Chaowei Xiao***, Zichao Li*, Haonan Qiu, Mingyan Liu, Bo Li*Can Shape Structure Features Improve Model Robustness under Diverse Adversarial Settings?*. ICCV 2021

[18] Aria Rezaei, **Chaowei Xiao**, Bo Li, Jie Gao. *Application-driven Privacy-preserving Data Publishing with Correlated Attributes*. EWSN 2021 (Best Paper Award)

[17] Yulong Cao*, Ningfei Wang*, **Chaowei Xiao***, Dawei Yang*, Jin Fang, Ruigang Yang, Alfred Chen, Mingyan Liu, Bo Li. *Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks*. IEEE Symposium on Security and Privacy 2021.

[16] Huan Zhang, Hongge Chen, **Chaowei Xiao**, Bo Li, Mingyan Liu, Duane Boning, Cho-Jui Hsieh. *Robust Deep Reinforcement Learning against Adversarial Perturbations on State Observations*. NeurIPS 2020 (Spotlight)

[15] Haonan Qiu*, **Chaowei Xiao***, Lei Yang*, Xinchen Yan, Honglak Lee, Bo Li. *SemanticAdv: Generating Adversarial Examples via Attribute-conditional Image Editing*. ECCV 2020

[14] Huan Zhang, Hongge Chen, **Chaowei Xiao**, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, Cho-Jui Hsieh. *Towards Stable and Efficient Training of Verifiably Robust Neural Networks*. ICLR 2020

[13] **Chaowei Xiao***, Dawei Yang*, Bo Li, Jia Deng, Mingyan Liu. *Realistic Adversarial Examples in 3D Meshes*. CVPR 2019 (Oral Presentation)

[12] **Chaowei Xiao**, Ruizhi Deng, Bo Li, Taesung Lee, Benjamin Edwards, Jinfeng Yi, Dawn Song, Mingyan Liu, Ian Molloy. *AdvIT: Characterizing Adversarial Frames in Videos Based on Temporal Information*. ICCV 2019

[11] Yulong Cao, **Chaowei Xiao**, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, Z. Morley Mao. *Adversarial Sensor Attack on LIDAR-based Perception in Autonomous Driving*. CCS 2019

[10] Liang Tong, Bo Li, Chen Hajaj, **Chaowei Xiao**, Ning Zhang, Yevgeniy Vorobeychik. *Improving Robustness of ML Classifiers against Realizable Evasion Attacks Using Conserved Features*. USENIX Security 2019

[9] Kin Sum Liu, **Chaowei Xiao**, Bo Li, Jie Gao. *Performing Co-Membership Attacks Against Deep Generative Models*. ICDM 2019

[8] **Chaowei Xiao**, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, Dawn Song. *Characterize Adversarial Examples Based on Spatial Consistency Information for Semantic Segmentation*. ECCV 2018

[7] **Chaowei Xiao***, Jun-Yan Zhu*, Bo Li, Warren He, Mingyan Liu, Dawn Song. *Spatially Transformed Adversarial Examples*. ICLR 2018.

[6] **Chaowei Xiao**, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, Dawn Song. *Generating Adversarial Examples with Adversarial Networks*. IJCAI 2018

[5] **Chaowei Xiao**, Armin Sarabi, Yang Liu, Bo Li, Tudor Dumitra, Mingyan Liu. *From Patching Delays to Infection Symptoms: Using Risk Profiles for an Early Discovery of Vulnerabilities Exploited in the Wild*. Usenix Security 2018

[4] Kevin Eykholt*, Ivan Evtimov*, Earlence Fernandes, Bo Li, Amir Rahmati, **Chaowei Xiao**, Atul Prakash, Tadayoshi Kohno, Dawn Song. *Robust Physical-World Attacks on Deep Learning Visual Classification*. CVPR 2018

[3] Chenshu Wu, Zheng Yang, **Chaowei Xiao**. *Automatic Radio Map Adaptation for Indoor Localization using Smartphones*. TMC 2017

[2] Chenshu Wu, Zheng Yang, **Chaowei Xiao**, Chaofan Yang, Yunhao Liu, Mingyan Liu. *Static Power of Mobile Devices: Self-updating Radio Maps for Wireless Indoor Localization*. INFOCOM 2015

[1] Lei Yang, Yekui Chen, Xiangyang Li, **Chaowei Xiao**, Mo Li and Yunhao Liu. *Tagoram: Real-time Tracking of Mobile RFID Tags to High Precision Using COTS Devices*. MobiCom 2014 (Best Paper Award)

Workshop Papers

[4] **Chaowei Xiao***, Zhongzhu Chen*, Kun Jin*, Jiongxiao Wang*, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, Dawn Song. *DensePure: Understanding Diffusion Models towards Adversarial Robustness* . NeurIPS SRML 2022 (Contributed Talk)

[3] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, **Chaowei Xiao**, Z Morley Mao. *Benchmarking robustness of 3d point cloud recognition against common corruptions*. ICML SRML 2021 (Best Paper Award)

[2] Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Anima Anandkumar, **Chaowei Xiao**. *MoleculeCLIP: Learning Transferable Molecule Multi-Modality Models via Natural Language*. NeurIPS AI4Science

[1] Boyi Li, Zhiding Yu, De-An Huang, Weili Nie, Linxi Fan, **Chaowei Xiao**,Serge Belongie, Kilian Q. Weinberger, Anima Anandkumar. *Weakly-Supervised Referring Image Segmentation with Multimodal Transformers*. NeurIPS InterNLP 2022

Funding (Total: My share $481,660)

Granted Exploring and mitigating unrestricted adversarial examples. Open Philanthropy. Role: Single PI. My share: $200,000.

Granted A Federated Query Optimizer for Privacy-Preserving Analytics and Machine Learning. Department of Homeland Security 2022 CAOE Research Grants for privacy. Total: $874,998 Role: Co-PI. My share: $281,660

Submitted DARPA ECOLE TA1&TA2. DOD-DARPA: Information Innovation Office (IIO). Total:$5,420,547 Role: Co-PI. My share: $600,951

Submitted DARPA Castle TA1. DOD-DARPA: Information Innovation Office (IIO). Total:$10,835,576 Role: Co-PI. My share: $530,425

Submitted DARPA Castle TA3. DOD-DARPA: Information Innovation Office (IIO) Total:$2,136,688 Role: Sub-contractor. My share: $480,352

Submitted Interdisciplinary Systems-based Training in Precision Nutrition (INTERAICT). NIH grant of advanced training in AI for precision nutrition science research (RFA-OD-22-027).

Selected Media Press

2022 FIERCE Electronics.Nvidia, others work to use LLMs to predict Covid variants

2022 HPCWire. Gordon Bell Special Prize Goes to LLM-Based Covid Variant Prediction

2022 Insightsfy.Speaking the Language of the Genome: Gordon Bell Finalist Applies Large Language Models to Predict New COVID Variants

2022 Scientific Computing World. ACM awards researchers for HPC-Based COVID-19 Research

2022 Tencent News. Speaking the Language of the Genome: Gordon Bell Finalist Applies Large Language Models to Predict New COVID Variants

2019 Analytics. Elon Musk Might Be Right. New Research Exposes Vulnerabilities In LiDAR-based Autonomous Vehicle.

2019 Synced. Researchers Fool LiDAR with 3D-Printed Adversarial Objects.

2019 Popular Science. Self-driving cars still have major perception problems

2019 GCN and Conversation. Autonomous vehicles can be fooled to see nonexistent obstacles

2017 Wired. Security News This Week: A Whole New Way to Confuse Self-Driving Cars.

2017 Fortune. Researchers Show How Simple Stickers Could Trick Self-Driving Cars

2017 Nature. Why deep-learning AIs are so easy to fool.

2017 IEEE SPECTRUM. Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms.

2017 Yahoo News. Researchers demonstrate the limits of driverless car technology.

2017 Telegraph. Graffiti on stop signs could trick driverless cars into driving dangerously.

Selected Industry Discussions & Responses

Triggered 20+ Autonomous Driving (AD) companies such as Tesla, GM, Daimler, Baidu, TuSimple, Aptiv, Hyundai, Volkswagen, Bosch, Lyft, Nuro, Toyota, Hyundai, Kia, and Volvo to start investigating our newly-discovered security vulnerabilities in AD localization and/or perception algorithms

Talks

Tiny Step towards Socially Responsible Machine Learning.

2022/8 Virtual Seminar series on Challenges and Opportunities for Security & Privacy in Machine Learning

2022/4 ICLR SRML opening remarks

Machine Learning in Adversarial Environment and Beyond.

2022/2 AAAI 2022 1st International Workshop on Practical Deep Learning in the Wild

2022/2 AAAI 2022 workshop on Adversarial Machine Learning and Beyond

2021/10 Hong Kong Baptist University

2021/9 University of Science and Technology of China

2021/8 Tsinghua University

2021/8 Zhejiang University

2021/7 ICML SRML opening remarks

Machine Learning in Adversarial Environment.

2020/11 Waterloo ML + Security + Verification Workshop

2020/3 Google Brain

2020/3 Facebook AI Research

2020/3 Nvidia Research

2020/3 Uber ATG Research

2020/3 Amazon AWS

2020/2 Visa Research

2020/2 Ant Finance

2020/2 ByteDance

Machine Learning: the Good, the Bad, and the Ugly.

2019/9 Microsoft Research

2019/3 Amazon Graduate Research Symposium

2019/2 University of Michigan, Ann Arbor

2018/6 Baiduxlab

Adversarial Objects for Lidar-Based Autonomous Driving System.

2019/8 Microsoft Security Workshop

2019/6 CVPR workshop on Adversarial Machine Learning in Real-World Computer Vision Systems

Characterizing Adversarial Frames in Videos Based on Temporal Information.

2018/8 IBM Watson Research Lab

Research Experience

2019 Microsoft Research, Redmond, USA. Research Intern at Deep Learning Group

2018 IBM Watson Research Lab, New York, USA. Research Intern at IBM Research AI group

Teaching & Mentoring Experience

2022-2023 Instructor, CSE 598 Machine Learning Security, Privacy and Fairness, ASU, Fall 2022. Design a new course on current topics in machine learning security, privacy and fairness. Course Evaluation: 4.7/5.0

2021-now Jiongxiao Wang (ASU Ph.D.): machine learning security

2021-now Yijin Yang (ASU Ph.D.): machine learning privacy

2018-now Jiachen Sun (UMich Ph.D.): adversarial examples in 3D vision systems

2018-now Yulong Cao (UMich Ph.D.): adversarial attacks in autonomous driving systems

2021 Manli Shu (UMaryland Ph.D.): zere-shot robustness in foundation models

2021 Boxin Wang (UIUC Ph.D.): large language model detoxification

2021 Chen Zhu (UMaryland Ph.D.): efficient and robust transformer for vision and language

2016-2018 Ruizhi Deng (UMich B.S, SFU M.S and now SFU Ph.D.): Adversarial attacks in semantic segmentation and audios

Academic Services

Organizer Workshop on Socially Responsible Machine Learning (Founder) in NeurIPS 2022, ICLR 2022, ICML 2021

Workshop on Neural Architectures: Past, Present and Future in ICCV 2021

Workshop on 1st International Workshop on Adversarial Learning for Multimedia in ACM-MM 2021

Workshop on Adversarial Machine Learning in Real-World Computer Vision System in CVPR 2019

PC member CVPR, ICCV, ECCV, ICLR, ICML, NeurIPS, CCS

Area Chair AAAI, CVPR

Panelist NSF 2023, NSERC 2021

600 W Grove Pkwy, Apt 2039, Tempe, AZ, 85283
734-2392-561
xiaocw@asu.edu
http://xiaocw11.github.io/
Google Scholar: citations 5000+

Chaowei Xiao

References

Mingyan Liu (mingyan@umich.edu), Peter and Evelyn Fuss Chair of Electrical and Computer Engineering, University of Michigan, Ann Arbor.

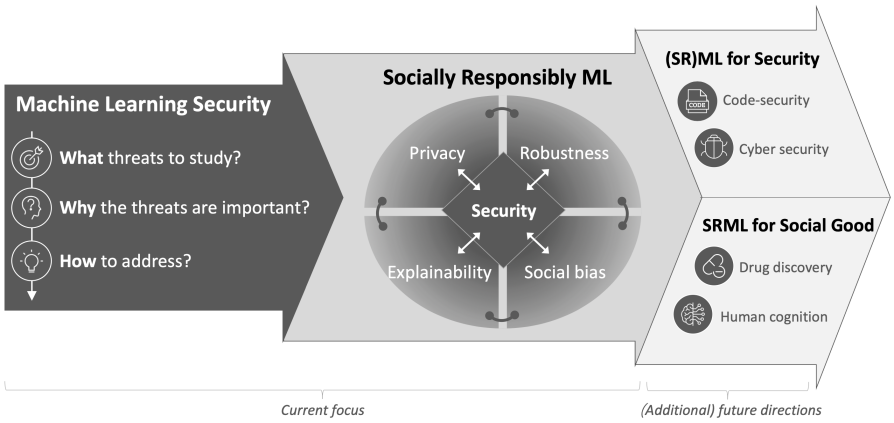
Anima Anandkumar (animakumar@gmail.com), Bren Professor, California Institute of Technology.

Marco Pavone (pavone@stanford.edu), Associate Professor, Stanford University

Research Statement

Machine Learning (ML), especially Deep Neural Nets (DNNs), has achieved great success in a variety of applications. However, various socially responsible concerns have also emerged, threatening the trustworthiness and adoption of ML techniques in real-world applications. Unlike classical algorithms that could be formally analyzed, there is less understanding of DNNs. This lack of understanding by either formal methods or empirical observations leaves potential issues in security, robustness, privacy, explainability, and social bias.

My research aims to establish principled approaches to build socially responsible machine learning (SRML) models and to explore its applications in vision, cyber security, language, health care, and IoT/CPS applications.



The above Figure highlights some of my research projects to achieve the above goal. My research goes beyond traditional disciplinary barriers and innovates at the crossroads of security and SRML. My research starts with studying security in ML models and later expands to SRML with broader considerations, including robustness, privacy, social bias, explainability, and their interconnection.

To study security in ML models, my research tackles the security threats at the inference stage to answer the following three core questions:

- **What threats to study:** my research broadens the scope of security threats (i.e., adversarial examples) beyond the traditional L_p regime by discovering new non-L_p threats. My contributions to this line have led to a single PI grant of $200,000.
- **Why these threats are important:** I am one of the pioneers in demonstrating severe physical consequences of adversarial examples by generating physical adversarial examples in real-world safety-critical applications such as industry-level autonomous driving systems (e.g., Baidu Apollo), voice controllable systems (e.g., Alexa), robotic and IoT systems. My research was shown in the London Science Museum to raise the general audiences' awareness of the risk.
- **How to address:** I propose principled solutions to defend against various unseen threats from both empirical and theoretical perspectives. I open a new direction to purify the adversarial perturbation by using the off-the-shelf diffusion model for free and show the STOA results in defending various unseen adversarial examples in both empirical robustness and (probabilistic) certified robustness settings.

Furthermore, I take a two-pronged approach to co-design ML security and SRML with a border consideration of *robustness, privacy, explainability,* and *social bias* On one hand, I apply ML security principles for the SRML domain to improve model robustness against distribution shift, discover new privacy attacks, use explainability (i.e., sharply value) for detecting security threats, and reduce large language models'(LLM) social bias. On the other hand, I work on both theory and practice to propose principled solutions for SRML from the perspectives of security, robustness, privacy, explainability, and social bias respectively. I also work on the best practice for building SRML models in the industry by proposing a commercial product, ModelCard++. Looking forward, I'm passionate about SRML for computer security and social good.

My research has brought significant impact to industry, academia and society. My research has been implemented by companies (IBM, NVIDIA, Baidu, etc.) to protect billions of users against adversarial attacks,

deployed in the second largest airport in the world (Beijing Capital International Airport) for baggage surveillance and triggered > 20 Autonomous Driving companies (e.g., Tesla, GM, Daimler, Baidu, TuSimple, Hyundai, Volkswagen, Lyft, Nuro, Volvo, etc.,) to start investigating newly-discovered vulnerabilities. My research has appeared in top-tier machine learning conferences, including NeurIPS, ICML, ICLR, CVPR, ECCV, ICCV, CORL with several spotlight and oral presentations, top-tier security and system conferences, including IEEE S&P, USENIX Security, CCS, and MOBICOM. My research won **Best Paper Awards** in MobiCom 2015 and ESWN 2021, **ACM Gordon Bell Special Prize** for HPC-Based COVID-19 Research, **workshop Best Paper Award** at ICLR 2022 SRML workshop and top positions in the public challenge. My work has also led to a wide range of media coverage - e.g., Wired, Nature, Fortune, Forbes, Telegraph, IEEE SPECTRUM, and Yahoo News. In addition, results from my work were shown in the London Science Museum to raise the public's awareness about the risks of adversarial examples.

1 Current Research

1.1 Machine Learning Security

I take the unique principle to study ML security by considering practical (security) impact, involving answering the following three core questions:

What threats to study. As a field, the first step in studying the security problem in ML is to discover security threats (i.e., adversarial examples). In the literature, adversarial examples are proposed under the L_p-norm regime: adversarial examples are the data points added by imperceptible perturbations bounded L_p-norm regime to mislead models.

My research sought to broaden the machine learning security domain beyond L_p-norm regime. I identified the limitation of the current L_p regime and argued that we should study more border threats beyond the L_p regime, which opened a new direction of non-L_p threats. I was the first to show the security threat beyond L_p regime by designing a new attack and laid the foundations. I proposed `stAdv`, which generated realistic non-Lp adversarial examples to mislead ML models without misleading humans by changing pixel positions instead of directly manipulating existing pixel values [12]. Following it, my research proposed several new attacks by considering practical security impacts including the semantically meaningful adversarial images[6] and audios [23], which manipulate the inherent semantic attributes of input (e.g., hair color for human image and prosody for audio), 3D adversarial examples, which manipulates the 3D shape and texture [17], adversarial WIFI [2] , and adversarial malware [16].

My new attacks bring new opportunities and challenges to the community. Since they use a completely different principle, existing adversarial training-based defenses, trained on L_p adversarial examples, are less effective. Furthermore, since my attacks do not require imperceptible criteria, they pave the way for adversarial examples in the physical world.

Why these threats are important. At the earlier stage, threats were designed in the digital domain. An opposite view on "No need to worry about adversarial examples" was brought up. Thus, understanding the security impact of adversarial examples in the physical/3D world and real ML systems was an important step in answering why we need to study adversarial examples.

I was one of the pioneers in performing physical attacks. My research focused on the unique angle to attack industry-level ML-based cyber-physical systems such as Autonomous driving systems (e.g., Baidu) and voice-controllable systems (e.g., Alexa). These systems operate in noisy physical environments, that can destroy L_p-based invisible perturbations created using current digital algorithms. Such systems usually consist of complicated machine learning models or sophisticated pre-and post-processing, challenging the standard differentiable end-to-end back-propagation when generating adversarial examples and

I designed several physical attacks to study the security impacts of these systems in a holistic and systematic way. Taking autonomous driving systems, as an example, my research successfully generated diverse robustly physical adversarial examples in AD systems ranging from traffic sign recognition to complicated Multi-Sensor Fusion (consisting of Lidar and Camera) object detection, from preception modules to prediction and planning modules and demonstrated the severe consequences (e.g., collision) in all of these components. In addition, my research also generated adversarial audio by perturbing prosody to physically attack real-world commercial products (Alexa and Google Home) [23], generated adversarial WIFI signals to attack WiFI-based behavior recognition [2], and generated physical adversarial environments to mislead reinforcement learning-based robots [15].

These explorations towards the physical world contributed to a systematic understanding of the susceptibility of ML-based cyber-physical systems against physical adversarial examples. These works provided cases for severe consequences of adversarial examples on deep learning models that interacted with the physical world and enlightened communities to raise awareness of risks and to find solutions by building robust models.

How to address. Many defense strategies have been proposed to find solutions, yet only to be broken thereafter. Adversarial training has become a standard defense form, due to its effectiveness. However, adversarial training has its intrinsic limitations: (1) ineffective against unseen threats, (2) higher training complexity, (3) un-scalable to large-scale datasets and complicated tasks (e.g., object detection) and so on, challenging its real-world applications.

To bridge this gap, I worked on both theory and practice to provide principled solutions. I proposed novel purification-based solutions [5, 13], that successfully purified adversarial examples by using diffusion models and theoretically explained why and how the diffusion model improves the adversarial robustness. Compared with the adversarial training methods, my methods could defend against unseen threats in a plug-and-play manner without re-training the classifiers and establish the new state-of-the-art (STOA) robustness against diverse unseen threats including various L_p-based attacks and `stAdv`. It exhibited a significant improvement, up to +36% in robust accuracy against unseen threats, and also offered a (probabilistic) *provable guarantee* of model robustness (i.e., certified robustness), outperforming existing methods by a large margin(over 12% on average compared with non-diffusion model-based methods, and over 37% on average compared with the off-the-shelf-based method). In addition, I successfully extended to 3D [10] and audio domain [23] by considering the domain properties, achieving the STOA adversarial robustness as well.

1.2 Socially Responsible Machine Learning (SRML)

Besides ML security, I am passionate about building SRML models with a broader consideration of robustness, privacy, explainability, and social bias. My research takes unique angles of security-SRML co-design, followed by building scalable and parameter-efficient SRML algorithms.

Security-SRML Co-design. My research took strong co-design between ML security and SRML with the consideration of *robustness, privacy,* and *explainability*.

From the *robustness* side, my research [20] utilized ML security techniques, adversarial training, to improve model robustness against unseen data distribution shifts. From the privacy side, my research [3, 24] discovered privacy threats of the public-available pre-trained models by designing membership [3] and model inversion attacks [24]. From the explainability side, my research leveraged a Shapley value (SV) based interpreter on detecting training-time security threats (backdoor attacks) [9].

Scalable and Parameter-efficient SRML. Beyond the interconnection, my research focuses on building scalable and parameter-efficient SRML models by considering robustness, privacy, and social bias respectively.

To improve *robustness*, my research went over the whole processing of the ML pipeline. At the training stage, my research proved that the optimization step of the information bottleneck (IB), which increases model robustness, could be rewritten into a self-attention form. Based on it, we designed scalable and efficient attention-based architectures to improve robustness including (1) designing an efficient self-attention [26] with linear complexity w.r.t input length, and (2) designing a family of fully attentional networks, FAN [25], with a novel attention-based channel processing design. As a result, my approach won ImageNet-C Domain Generalization benchmark compared with methods in the same training setting [1]. At the test-time stage side, I proposed a parameter-efficient method, test-time prompt tuning to improve robustness [8].

For *privacy*, my research worked on striking a balance between theory and practice, with a focus on scalable and parameter-efficient training with a privacy guarantee. I built privacy-enhanced algorithms from three different perspectives: (1) scalable and parameter-efficient differential private (DP) training [4] ; (2) scalable privacy-preserving data publishing [7] and (3) parameter-efficient personalized federate learning [22]. My results in this direction won the ESWN Best Paper Awards.

For *social bias*, I worked on the toxicity and biased issues of the large language model (LLM) and proposed to use the generative power of LMs itself to detoxify the LLM in a parameter-efficient way [19]. My method was the first practical method that was deployed at 530 billion Large language model at NVIDIA and reduced the toxicity probability by over 30%.

For *explainability*, I focused on training-time explanation and proposed GKNN-Shapley by taking advantage of KNN-Shapley to approximate the Shapley value of nodes at training time [9].

SRML in Practice. Besides algorithm designs, I proposed and designed ModelCard++ to enhance SRML in

practice [14] at NVIDIA. ModelCard++ provided information about a model with diverse trustworthy properties including safety, robustness, privacy, bias, and explainability. Since it explicitly listed various trustworthy properties of the model, it could be helpful in understanding the capabilities and limitations of a model and transparently showed the steps that the developer/researcher was taking to enhance trustworthiness. NVIDIA plans to start rolling out Model Card++ by the end of the year, with all commercial models using it by the end of 2023.

2 Future Research Directions

I will continually work on strengthening the interconnection between security and machine learning, and building socially responsible machine learning models. In particular, I would like to extend my investigation in the following directions.

2.1 Security for Machine Learning

SRML and Diffusion Model co-design. My theoretical and empirical studies of diffusion model-based purification methods have demonstrated success in enhancing ML model security. Yet there are still many open research questions and opportunities for these methods. For instance, how to increase the inference stage efficiency; how to secure visible adversarial attacks (e.g., patch-based adversarial examples); how to increase the robustness against domain distribution shift; how to enhance privacy and handle security, robustness, and privacy simultaneously with a unified framework for SRML. Looking forward, I plan to use my previous achievements as the foundation to further answer such questions and build SRML models.

SRML in the Era of Foundation Models. We are witnessing a paradigm shift in AI. Looking back last decade, AI systems have been developed for different problems in a task-specific way, each of which is trained on a specific large, well-labeled dataset. However, there was a sea change to replace task-specific models. Foundation models (e.g., Stable-Diffusion, Codex, GPT-3, etc.), trained on a broad range of (unlabeled) data and can be adapted to various tasks, have been proposed to bring new opportunities for building SRML. For instance, as shown by previous research, we used the foundation model (CLIP) to improve the model robustness against data distribution shift [8]. However, new socially responsible risks are also being raised. For instance, from the *security perspective*, since these models are usually trained from the web-crawl data, training-time backdoor attacks might be "unintentionally" injected into these models. From the *privacy perspective*, GPT-3 can leak sensitive information (e.g., SSN). From the *ethical perspective*, code generation models (e.g., Codex, Copilot) can generate code with copyright issues. From the *social bias perspective*, it is more challenging to detoxify the large language models (e.g., GPT-3, Megatron-LM) as demonstrated by my previous work [19]. Looking forward, I am excited to keep exploring socially responsible problems in the foundation models in a holistic way, developing foundations to understand them, and building socially responsible foundation models.

2.2 SRML for Security.

My first two Ph.D. projects applied machine learning to cyber-security with a focus on discovering early vulnerabilities [11] and building robust PDF malware detection [16, 18]. For instance, I have developed an ML-based method for the early discovery of vulnerability exploited by studying the correlation between the symptomatic botnet data (in the form of a set of spam blacklists) and the risk behavior of end-host collected from end-host software patching behavior [11] and building robust PDF malware detection using conserved features [18]. With my research foundations in both security and machine learning, I am passionate about the following two directions.

Code-Security Co-design. The code generation models, such as Codex and Copilot, offer new opportunities (e.g., assisting humans in automating time-consuming coding tasks). Since these models are trained on untrustworthy open-source code repo without any security specifications, they introduce security concerns about generating insecure codes. Although researchers in the machine learning community have proposed advanced ways (e.g., using reinforcement learning) to improve the code generation model, the main focus is still on improving the pass rate on competitive programming, leaving security analysis largely unexplored. I will leverage my experience in both security and machine learning domains to perform code generation and security co-design. Taking advantage of my expertise in toxicity studies for large language models [19], I hope to build secure code testing frameworks for code generation models and enhance secure code generation in return.

SRML for Cyber Security. Currently, fruitful security-related knowledge has been aggressively collected via natural language, enabling the community to gain visibility into the fast-evolving threat landscape, tactics, and techniques. For instance, Open-Source Cyber Threat Intelligence (OSCTI) collects the knowledge of cyber threats; MITRE offers a globally-accessible knowledge platform for adversary tactics and techniques based on real-world observations; security conference, journal, and medium articles provide underlying principles about diverse threats. Such knowledge provides rich opportunities to build interpretable cyber security. Based on my previous experience in natural language models [19] and cyber security [11], I hope to build security-specific generative models (e.g., Security-GPT/T5) for structuring the knowledge, building a knowledge graph, explaining the threat and enabling a series of security applications.

2.3 SRML for Social Good

I dream of building SRML models for social good to ultimately improve health outcomes and wellbeing. I have taken my initial steps in the healthcare domain by (1) building a retrieval-based controllable molecules editing framework for drug discovery (e.g., treating SARS-CoV-2 virus) [21] and (2) building genome-scale language models (GenSLMs) to learn the evolutionary landscape of SARS-CoV-2 genomes [27]. My research won ACM Gordon Bell Special Prize for HPC-Based COVID-19 Research. In the future, I plan to work in the following two directions.

SRML for Drug Discovery. AI methods have been used to augment and accelerate current computational pipelines. Existing methods mainly focus on modeling the molecules by solely using the chemical structure information with limited and expensive annotations. It challenges the way to build AI for drug foundation models, which require training on huge amounts of data. Compared with expensive annotations, there are vast amounts of domain-specific natural language data. Such natural language data contains professional and trustworthy molecule descriptions and knowledge. For instance, PubChem contains information on millions of chemical compounds, providing a vast amount of information on their structures, physical properties, and biological activities. PubMed contains millions of citations and abstracts for articles, books, and other documents in the biomedical and life sciences fields. Thus, I plan to use the natural language as a bridge to design multi-modality unified molecule foundation models that work on diverse AI-for-molecule tasks, promising to be transformative for drug discovery.

SRML via human cognition. Although current foundation models (e.g., CLIP, GPT-3, Dalle-2) demonstrate promising results, their designs are still different from our human cognition systems. As an example, the current GPT-3 model is designed to learn from huge datasets with large model sizes. Although it shows a certain extrapolation ability, it is still be viewed as memorizing training data with limited applicability in complicated tasks (e.g., numerical reasoning).In addition, due to its learning mechanism as well as the unreliable training data source, it faces severe factuality and trustworthy issues. Since it requires training on a very large dataset, the current situation becomes a competition among big giant companies, introducing accessibility issues to academic researchers with limited resources.

I have a dream to build a future accessible SRML inspired by the human cognition perspective. Compared with current ML models, humans have a retrieval mechanism and working memory, supporting learning in a continuous, symbolic, and efficient way. Besides memorizing the data, humans learn the underlying principles (or concepts) and learn how to retrieve the relevant information for extrapolation in a symbolic way. These lead to a gap between human cognition learning and current ML. In the future, I am interested in bridging this gap. I plan to build future SRML models in a more holistic and human-like way via the following steps. First, I hope to focus on trustworthy data sources, e.g., scientific articles, which contain the underlying knowledge and principles of science and can be used for many social good applications (e.g., healthcare, education). Second, I hope to add a retrieval-based mechanism into the model to suggest retrieval from a large set as a complementary path to scaling models based on my previous experience [21]. Third, I hope to embed the executable program (i.e., code) as the human in-working memory in a symbolic way.

Interdisciplinary research and collaborations. To practically solve big and important problems, I consistently collaborate with experts from both academia in different domains such as machine learning, computer vision, cyber security, cyber-physical system/IoT, natural language processing and health care. In the future, I am always opening to other potential collaborations with different domains.

References

[1] Paperswithcode domain generation on imagenet-c, 2022. https://paperswithcode.com/sota/domain-generalization-on-imagenet-c.

[2] J. Liu, Y. He, **C. Xiao**, J. Han, L. Cheng, and K. Ren. Physical-world attack towards wifi-based behavior recognition. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 400–409. IEEE, 2022.

[3] K. S. Liu, **C.Xiao**, B. Li, and J. Gao. Performing co-membership attacks against deep generative models. *ICDM*, 2019.

[4] Z. Luo, Y. Zou, Y. Yang, Z. Durante, D.-A. Huang, Z. Yu, **C. Xiao**, F.-F. Li, and A. Anandkumar. Differentially private video classification with multi-clip dp-sgd and parameter-efficient transfer learning. 2022.

[5] W. Nie, B. Guo, Y. Huang, **C. Xiao**, A. Vahdat, and A. Anandkumar. Diffusion models for adversarial purification. *ICML*, 2022.

[6] H. Qiu, **C. Xiao**, L. Yang, X. Yan, H. Lee, and B. Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision*, pages 19–37. Springer, 2020.

[7] A. Rezaei, **C. Xiao**, J. Gao, B. Li, and S. Munir. Application-driven privacy-preserving data publishing with correlated attributes. *EWSN (Best Paper Award)*, 2018.

[8] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and **C. Xiao**. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022.

[9] A. Sun, H. Li, X. Xu, **C. Xiao**, C. Zhang, C. A. Gunter, and B. Li. Gknn-shapley: Training-time explanation for gnns. 2022.

[10] J. Sun, W. Nie, Z. Yu, Z. M. Mao, and **C. Xiao**. Pointdp: Diffusion-driven purification against adversarial attacks on 3d point cloud recognition. *arXiv preprint arXiv:2208.09801*, 2022.

[11] **C. Xiao**, A. Sarabi, Y. Liu, B. Li, M. Liu, and T. Dumitras. From patching delays to infection symptoms: using risk profiles for an early discovery of vulnerabilities exploited in the wild. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 903–918, 2018.

[12] **C. Xiao**, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018.

[13] **C. Xiao**, Z. Chen, K. Jin, J. Wang, W. Nie, M. Liu, A. Anandkumar, B. Li, and D. Song. Densepure: Understanding diffusion models towards adversarial robustness. *arXiv preprint arXiv:2211.00322*, 2022.

[14] **C. Xiao**, B. Michael, P. Nikki, and A. Anandkumar. Enhancing ai transparency and ethical considerations with model card++. https://developer.nvidia.com/blog/enhancing-ai-transparency-and-ethical-considerations-with-model-card/, May 2022.

[15] **C. Xiao**, X. Pan, W. He, J. Peng, M. Sun, J. Yi, M. Liu, B. Li, and D. Song. Characterizing attacks on deep reinforcement learning. *AAMAS*, 2022.

[16] **C. Xiao**, H. Qiu, W. Guo, G. Wang, X. Xing, M. Liu, and B. Li. Paintmal: Inpainting network based malware evasion generation. 2019.

[17] **C. Xiao**, D. Yang, B. Li, J. Deng, and M. Liu. Meshadv: Adversarial meshes for visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6898–6907, 2019.

[18] L. Tong, B. Li, C. Hajaj, **C. Xiao**, N. Zhang, and Y. Vorobeychik. Improving robustness of {ML} classifiers against realizable evasion attacks using conserved features. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 285–302, 2019.

[19] B. Wang, W. Ping, **C. Xiao**, P. Xu, M. Patwary, M. Shoeybi, B. Li, A. Anandkumar, and B. Catanzaro. Exploring the limits of domain-adaptive training for detoxifying large-scale language models. *NeurIPS*, 2022.

[20] H. Wang, **C. Xiao**, J. Kossaifi, Z. Yu, A. Anandkumar, and Z. Wang. Augmax: Adversarial composition of random augmentations for robust training. *NeurIPS*, 34:237–250, 2021.

[21] Z. Wang, W. Nie, Z. Qiao, **C. Xiao**, R. Baraniuk, and A. Anandkumar. Retrieval-based controllable molecule generation. *arXiv preprint arXiv:2208.11126*, 2022.

[22] C. Xie, D.-A. Huang, B. Li, A. Anandkumar, and **C. Xiao**. Parameter-efficent federated learning. 2022.

[23] Z. Yu, Y. Chang, N. Zhang, and **C. Xiao**. Semantically meaningful adversarial audio attack. *USENIX Security*, 2023.

[24] Z. Yuan, F. Wu, Y. Long, **C. Xiao**, and B. Li. Secretgen: Privacy recovery on pre-trained models via distribution discrimination. In *European Conference on Computer Vision*, pages 139–155. Springer, 2022.

[25] D. Zhou, Z. Yu, E. Xie, **C. Xiao**, A. Anandkumar, J. Feng, and J. M. Alvarez. Understanding the robustness in vision transformers. In *ICML*, pages 27378–27394. PMLR, 2022.

[26] C. Zhu, W. Ping, **C. Xiao**, M. Shoeybi, T. Goldstein, A. Anandkumar, and B. Catanzaro. Long-short transformer: Efficient transformers for language and vision. *NeurIPS*, 34:17723–17736, 2021.

[27] M. T. Zvyagin, A. Brace, ..., **C. Xiao**, et al. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *ACM Gorden Bell special prize for HPC-based COVID-19 Research*, 2022.

Teaching Statement

Chaowei Xiao

In my past experience in educational experience ranging from classroom teaching, student mentoring, and community outreach, I feel that I deeply enjoy the sense of achievement from teaching and mentoring. It is also one of the most important reasons for me to become a faculty.

Philosophy. I strongly believe that the central goal of teaching is to help students develop strong study skills and nurture creativity and subjective initiative to apply the learned knowledge to real-world problems. Following this goal, my teaching confused on (1) igniting students' curiosity and interest, (2) encouraging teamwork and creativity through open-ended team projects and (3) fostering subjective initiative via carefully designed step-by-step questions to inspire. I hope students can become active learners, team players, and innovators by this design.

Teaching Interests. In my past teaching and research experience, I have covered a wide range of topics in computer science at both the undergraduate and graduate levels. As a result, I am able to teach a variety of computer science-related undergraduate courses, including programming, data structure, cyber security, algorithms, artificial intelligence, and computer networks. Additionally, I am qualified to teach graduate-level courses such as machine learning and security, and I am also capable of developing and teaching new courses, such as trustworthy machine learning

My past teaching and research experience have covered a wide range of topics in electrical engineering and computer sciences (EECS) at both undergraduate and graduate-level courses. Therefore, I can teach any EECS-related undergraduate courses such as programming, data structure, cyber security, algorithms, artificial intelligence, computer network, and so on. I can also teach graduate courses such as machine learning, security. I am also capable of developing and teaching new courses such as socially responsible machine learning.

Teaching experience. I recently taught a new course, CSE 598: Machine Learning Security, Privacy, and Fairness, at Arizona State University in 2022 Fall and received very positive feedback from my students. Some comments included: "This course was excellent and amazing"; "Even though I had no prior experience in this field, I now feel confident in my ability to do research in this domain thanks to this amazing course." During my teaching, I used three methods to keep the class attractive and creative. First, *live demos*. During the class, I demonstrated various examples to the students to help illustrate concepts and make the material more engaging. For instance, I showed them demos of Baidu Apllo's autonomous driving car and discussed adversarial examples. I found that this approach was effective in capturing the students' attention and motivating them to apply their learned knowledge to real-world problems. Second, *open-end group projects and detailed guidance*. I designed open-end course projects. I encourage students to think creatively and explore the trustworthiness issues of machine learning models. I held office hours to discuss their projects and provide guidance, and I found that this approach was very effective in inspiring the students' creativity. Many of them proposed interesting ideas, and some of the projects were able to submit to top-tier conferences. The students continue to work with me to prepare for conference submissions. Third, *step-by-step questions*. During my lectures, I did not simply provide complete solutions to problems. Instead, I designed step-by-step questions to guide the students through the process of solving the problems on their own. This approach was effective in helping the students develop strong study skills and build their confidence.

Additionally, I enhanced diversity, equity, and inclusion by creating support, consideration, encouragement, and inclusive environments during my teaching. I shared my course information and materials to other departments (e.g., the school of Electrical, Computer and Energy Engineering) to provide them the opportunity to take my course. During the class, I noticed some students struggling to understand the material and participating less actively in discussions. In response, I offered them additional office hours and encouraged them to ask questions during class to improve their sense of participation and achievement. In the future, I will continue to work towards addressing the challenges of teaching diverse students and creating an inclusive environment for all students, and I will encourage students from other departments to take my courses as well.

Mentoring experience. In addition to teaching, I have been fortunate to mentor many talented students. Fortunately, most of them have at least one co-author paper with me. I am grateful that all of the master and undergraduate students decided to apply for the Ph.D. program and kept working with me. During my mentoring, I encouraged students to actively think about the question before I provide answer as I found junior students tend to enjoy doing detailed tasks with a lack of thinking in depth. In discussions, I start by participating in high-level research direction discussions and encourage the students to describe their ideas. Then, I inspire them to think about the strengths and weaknesses of their ideas before providing my own solutions. This approach has been very effective, as students often come up with more reasonable methods and experiments than I had expected.

Community outreach. I am always enthusiastic about sharing my knowledge and research with the broader public. For example, the results of my paper "Adversarial Stop Signs" were displayed at the London Science Museum to educate the general public about the security issues of current machine learning models. I have also organized workshops at top-tier conferences such as the Socially Responsible Machine Learning workshop at ICLR, ICML, and NeurIPS, and the Neural Architectures: Past, Present and Future workshop at ICCV. In addition, I have participated in engineering camps for kids at the University of Michigan, Ann Arbor and shared my research experiences and insights on social media with high school students in my hometown. I have given lectures on my latest research at universities, seminars, and companies, and have enjoyed chatting with students about both research and life in general. I plan to continue this type of outreach in the future.

Diversity Statement

Chaowei Xiao

My commitment to promoting diversity, equity and inclusion is grounded in my personal life, leadership, teaching and mentoring experience.

As an ethnic minority of Hui (0.79% of the Chinese population) growing up in a remote western part of China, where educational resources are scarce, I have firsthand experience of the value of working and interacting with diverse groups and the importance of equitable circumstances. Before I went to university, I had never studied programming due to a lack of professional instructors, and there were no minorities from my hometown choosing computer science as their major in my undergraduate university due to a lack of programming background. This education gap was a huge obstacle for me in my freshman year. To overcome it, I studied hard and asked questions of students with diverse educational backgrounds. That diversity and effort helped me achieve good grades and showed that minorities from my hometown can succeed and have a career in computer science. During my sophomore year at Tsinghua, I co-founded a non-profit organization to foster communication between minorities from my hometown and students at Tsinghua. We recruited volunteers with different backgrounds to go to high schools in my hometown to teach both students‘ courses as well as sharing their study methodologies and university's life. I was encouraged to hear that some students felt that these events gave them the opportunity to connect with people from outside their hometown and strengthened their ambition to study hard to develop their hometowns.

During my teaching, I have made an effort to enhance diversity, equity, and inclusion by creating supportive, considerate, and inclusive environments. I had shared the information outside the CSE department about my course information and encouraged all students to take my courses. In my courses, I have noticed that some students struggle with course projects, team grouping, and participating in discussions. To address these issues, I offer additional office hours and encourage students to ask questions during class to improve their sense of participation and achievement. Additionally, I believe that diverse teams make academia stronger. For the final project, I encouraged groups to be composed of students with different genders, countries, and cultural backgrounds to give every student the opportunity to work with a diverse group. In the final anonymous survey, all students were satisfied with the course design.

During my mentorship, I strive to recognize and eliminate implicit biases and create a welcoming environment for all. I strongly believe that all people, regardless of race, ethnicity, gender, sexual orientation, disability, culture, and age, should be provided with equal opportunities and that everyone has the potential and ability to achieve successful careers. Therefore, I have mentored students from different countries, with different genders and ages. Half of my Ph.D. students are female, and over 50% of my interns at NVIDIA are in minority groups. I use different strategies for different students and encourage them to work together. For example, I worked with Max Wolff from Viewpoint School, a high school student, on a project with more empirical experiments, and I worked with Kaizhao, an undergraduate student with a strong theoretical background, on a project related to certified defense. I am encouraged that all of the students have continued working with me and plan to apply to Ph.D. programs

I emphasize diversity in my research as well. For example, I have worked on reducing the bias of machine learning models. In addition, I consistently collaborated with people with different research backgrounds, including machine learning, cybersecurity, the Internet of Things, and so on. These collaborations with diverse researchers have broadened my horizons and resulted in many papers.

To be a faculty in the Engineering field, I could realize that the gap of diversity still exists (e.g., women only earn 18% of computer science bachelor's degrees in the US). By discussing with students from underrepresented groups, the gap could also come from unconscious bias from parents or teachers. For instance, parents will tell female students that girls usually are not good at engineering unconsciously. This bias could make female students lose their interest in Engineering from an early age. To enhance the diversity in terms of demographics in Engineering related departments, I will work on avoiding this unconscious bias and conduct outreach to raise awareness of the parents' and educators' unconscious bias. Looking forward, my commitment to promoting diversity, equity, and inclusion is solid. I believe that diversity is the source of innovative ideas and creative accomplishments. It is also the core values that every faculty member should actively contribute to advancing. I would like to play an active role in strengthening diversity, equity, and inclusion in the future. As a faculty member, I would like to conduct outreach to hire a diverse group of students. In teaching, I will strive to address the challenges of teaching diverse students, create an inclusive environment for students and encourage students from other departments to take my courses. Additionally, I also hope to be a mentor for the under-representative PhDs in EECS with the help of other faculties if possible. Besides these, I hope to conduct outreach to open a summer school to invite the local under-representative high-school students and their parents, which aims to promote diversity at high school. Overall, I hope to play an active role in strengthening diversity, equity, and inclusion. In the end, let me use one of my favorite sentences to finish my statement.

“Diversity is not just something to be achieved; it is the source of life.”