Department of Computer Science
香港城市大學
City University of Hong Kong

COMPUTER SCIENCE COLLOQUIUM

Some Personal Perspectives and Experiences on Trustworthy AI

SPEAKER **Dr. Leo Zhang**

Senior Lecturer
School of Information and Communication Technology, Griffith University, Australia

DATE 19 Apr, 2024 (Fri)
TIME 5:00 PM - 6:00 PM
VENUE CS Seminar Room, Y6405, 6th Floor, Yellow Zone, Yeung Kin Man Academic Building, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon Tong

ABSTRACT

In today's world, Artificial Intelligence (AI) permeates every facet of our lives, underscoring the critical importance of trustworthy AI. Trustworthy AI encompasses various dimensions, including safety & robustness, privacy, generalizability, fairness, and explainability. The core of many challenges in achieving trustworthy AI lies in distribution shifts within data. For instance, adversarial/evasion attacks stem directly from distribution shifts between training and test datasets. This talk delves into the intersection of safety & robustness, and generalizability in AI through the lens of our recent research endeavors. Focusing primarily on safety and robustness, we explore the susceptibility of AI models to poisoning attacks, which can stealthily introduce trojans or backdoors, compromising their integrity. Furthermore, we investigate the underlying reasons behind the notable generalizability of adversarial attacks across diverse data samples and neural network architectures. By unraveling these complexities, we aim to shed light on crucial aspects of building trustworthy AI systems in an era where their reliability is paramount.

BIOGRAPHY

Dr. Leo Zhang is currently a Senior Lecturer with the School of Information and Communication Technology, Griffith University, Australia. Prior to this, he was a faculty member with the School of Information Technology, Deakin University, from 2018 to 2023. He received the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2016. Leo's research interests focus on trustworthy AI (adversarial/poisoning/privacy attacks & defenses) and applied cryptography (privacy-preserving computation, authentication in emerging areas). He has published over 100 conference and journal articles in these fields (with 3100+ citations and h-index 29), many of them appear in top venues like IEEE S&P, Esorics, AsiaCCS, NeurIPS, CVPR, ICCV, AAAI and IJCAI. His research is supported by NSFC (China), Cyber CRC (Australia), Bosch, and Nvidia. He was the recipient of the 2021 Australian Information Security Association Researchers of the Year Award, and he is an Associate Editor for IEEE Transactions on Dependable and Secure Computing.

All are welcome!

In case of questions, please contact Prof. WANG Cong at congwang@cityu.edu.hk, or visit the CS Departmental Seminar Web at https://www.cs.cityu.edu.hk/events/cs-seminars/recent-cs-colloquiums.