MS02: Trustworthy ML: Theory and Applications

Organizers: Shuren Qi, Fenglei Fan, Youzhi Zhang, and Xiaoge Zhang

July 7, 2025 (Monday, Morning)	**Speaker**
10:20-10:30	Opening
10:30-11:00	Xiaochun Cao
11:00-11:30	Tianwei Zhang
11:30-12:00	Youzhi Zhang
12:00-12:30	Zitong Yu

July 8, 2025 (Tuesday, Morning)	**Speaker**
10:20-10:30	Opening
10:30-11:00	Bo Han
11:00-11:30	Xiaoge Zhang
11:30-12:00	Zhiyong Yang
12:00-12:30	Shuren Qi

Xiaochun Cao
Sun Yat-sen University
caoxiachun@mail.sysu.edu.cn

Attribution-based Safety in Agent Decision-making

Ensuring the safety of agent decision-making is a critical challenge in artificial intelligence, with significant implications for improving decision reliability, safety monitoring, and risk prevention. Attribution techniques, as a core component of explainability, face the key question of whether they can effectively enhance decision safety and reliability. Prior research shows that attribution analysis not only clarifies model decision processes but also plays a vital role in model quality monitoring and risk control. High-quality explanations are crucial, as models with better performance often exhibit more reasonable attribution distributions, which help identify decision anomalies. Certain training strategies can also improve the reasonableness of attributions. This talk systematically explores attribution-based safety in agent decision-making across three areas: interpretable attribution techniques, attribution-guided training, and attribution-based anomaly monitoring and correction. We first review attribution methods at the input, parameter, and data levels to identify factors influencing decisions and provide theoretical support. Next, we analyze existing attribution-guided training methods and their effects on attribution reasonableness and model performance. Finally, we investigate whether attribution distributions correlate with decision reliability in deployment, explore how monitoring attribution anomalies can enhance safety, and assess low-cost online correction methods for decision failures. Our goal is to offer theoretical foundations and technical pathways for safer agent decision-making, supporting future research and implementation. We also evaluate attribution techniques in embodied and non-embodied agent scenarios, assess their maturity, and outline future trends.

Bio: Xiaochun Cao is the Dean of School of Cyber Science and Technology, Sun Yat-sen University. His research interests include: artificial intelligence especially computer vision, and content analysis in cyber space, etc. He received the B.E. and M.E. degrees both in computer science from Beihang University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. Before joining SYSU, he was a professor at Institute of Information Engineering, Chinese Academy of Sciences. He has authored and coauthored over 300 journal and conference papers. In 2004 and 2010, he was the recipients of the Piero Zamperoni best student paper award at the International Conference on Pattern Recognition. Dr. Cao was the recipients of Outstanding Young Scientists Fund and Excellent Young Scientists Fund of National Natural Science Foundation of China, in 2020 and 2014, respectively. He is on the editorial boards of IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, and Acta Electronica Sinica. He was on the editorial board of IEEE Transactions on Circuits and Systems for Video Technology.

Tianwei Zhang
Nanyang Technological University
tianwei.zhang@ntu.edu.sg

Safety Benchmarking and Testing of Multimodal Large Language Models

Multimodal Large Language Models (MLLMs), epitomized by ChatGPT, Stable Diffusion and Heygen, have made remarkable strides in very recent years. They significantly simplify the tasks of creating high-quality content in various modalities such as images, videos, and coherent text based on users' demands. However, the widespread adoption of generative models has also raised ethical and societal concerns. Issues related to data privacy, bias in AI algorithms, and the potential for model misuse have become subjects of intense debate. In this talk, I will introduce some works towards the safety benchmarking and testing of MLLMs. These studies accentuate the pressing challenges and opportunities in securing large model ecosystems.

Bio: Tianwei Zhang is currently an associate professor at College of Computing and Data Science, Nanyang Technological University, Singapore. He is the deputy director of cyber security research centre @ NTU, and associate director of NTU Centre Computational Technologies for Finance. He received his Bachelor's degree at Peking University in 2011, and Ph.D degree at Princeton University in 2017. He has been involved in the organization committee of numerous technical conferences. He serves on the editorial board of IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) since 2021, and receives the best editor award in 2023. His research focuses on building efficient and trustworthy computer systems. He has published more than 150 papers in top-tier security, AI, and system conferences and journals. He has received several research awards, including Distinguished Paper Award @ ASPLOS'23, Distinguished Paper Award @ ACL'24, Distinguished Artifact Award @ Usenix Security'24, Distinguished Artifact Award @ CCS'24.

Youzhi Zhang
Hong Kong Institute of Science & Innovation
youzhi.zhang.lgc@gmail.com

A Game-theoretic Approach for Trustworthy AI

Progress in AI has often been measured by the mastery of games, and game-theoretical AI systems have been applied to solve many real-world problems as well. Game theory is an important tool for trustworthiness, especially security. In this talk, I will first present the game-theoretic approach for public security, especially about a real-world problem of urban network security games, where law enforcement officers must respond quickly to apprehend a criminal who is choosing an escape route while the criminal strategically selects a path to evade capture. Based on that, I will show how game theory can be used for trustworthy AI.

Bio: Youzhi Zhang is an associate professor at the Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science & Innovation, Chinese Academy of Sciences. Before joining CAIR, he was a postdoctoral researcher at the Northwestern University, and the Dartmouth College, USA. He received the Ph.D degree in Computer Science from the Nanyang Technological University, Singapore. His research interests include artificial intelligence, game theory and security, AI for science. He has published over 30 referred conference/journal papers, including top conference papers at AAMAS, IJCAI, AAAI, NeurIPS, ICLR, and ICML. For contributions to algorithms for equilibrium computation in multiplayer games and their applications, he won the Ph.D Best Thesis Award from the School of Computer Science and Engineering, Nanyang Technological University, in 2021. He won the Best Paper Award at the IEEE DASC 2022. For outstanding contributions to top AI conferences, he won the ICML 2022 Outstanding Reviewer Award and the IJCAI 2023 Distinguished Program Committee Member Award.

Zitong Yu
Great Bay University
yuzitong@gbu.edu.cn

Subtle Visual Computing

Subtle visual signals, though often imperceptible to the human eye, contain subtle yet crucial information that can reveal hidden patterns within visual data. By applying advanced computer vision and representation learning techniques, we can unlock the potential of these signals to better understand and interpret complex environments. This ability to detect and analyze subtle signals has profound implications across various fields, e.g., 1) from medicine, where early identification of minute anomalies in medical imaging can lead to life-saving interventions, 2) from industry, where spotting micro-defects in production lines can prevent costly failures, 3) from affective computing, where understanding micro-expression and micro-gesture under human interaction scenarios can benefit the deception detection. In this talk, we will introduce foundation models and methods to detect and decode these 'subtle visual signals' on representative downstream applications.

Bio: Zitong Yu received the Ph.D. degree in Computer Science and Engineering from the University of Oulu, Finland, in 2022. Currently, he is an Assistant Professor at Great Bay University, China. He was a Postdoctoral researcher at ROSE Lab, Nanyang Technological University. He was a visiting scholar at TVG, University of Oxford, from July to November 2021. He is a Senior Member of IEEE. His research focus on subtle visual computing. He has published more than 40 works on top-tier journals and conferences such as TPAMI/IJCV/CVPR/ICCV/ECCV, and received 6200 google citations. He was AC/SPC of ACM MM'25, ICME'23, BMVC'24/25, IJCB'24, and IJCAI'25. He organized the 1st Workshop & Challenge on Subtle Visual Computing on ACM MM'25 and Special Issue Subtle Visual Computing on Machine Intelligence Research. He won the 1st Place in the ChaLearn Multi-Modal Face Anti-spoofing Attack Detection Challenge with CVPR'20. He was a recipient of IAPR Best Student Paper Award, IEEE Finland Section Best Student Conference Paper Awar, second prize of the IEEE Finland Jt. Chapter SP/CAS Best Paper Award, Best Paper Candidate of ICME'24, Best Paper Honorable Mention Award of CCBR'24, and World's Top 2% Scientists 2023/2024 by Stanford.

Bo Han
Hong Kong Baptist University
bhanml@comp.hkbu.edu.hk

Exploring Trustworthy Foundation Models: Benchmarking, Finetuning and Reasoning

In the current landscape of machine learning, where foundation models must navigate imperfect real-world conditions such as noisy data and unexpected inputs, ensuring their trustworthiness through rigorous benchmarking, safety-focused finetuning, and robust reasoning is more critical than ever. In this talk, I will focus on three recent research advancements that collectively advance these dimensions, offering a comprehensive approach to building trustworthy foundation models. For benchmarking, I will introduce CounterAnimal, a dataset designed to systematically evaluate CLIP's vulnerability to realistic spurious correlations, revealing that scaling models or data quality can mitigate these biases, yet scaling data alone does not effectively address them. Transitioning to finetuning, we delve deep into the process of unlearning undesirable model behaviors. We propose a general framework to examine and understand the limitations of current unlearning methods and suggest enhanced revisions for more effective unlearning. Furthermore, addressing reasoning, we investigate the reasoning robustness under noisy rationales by constructing the NoRa dataset and propose contrastive denoising with noisy chain-of-thought, a method that markedly improves denoising-reasoning capabilities by contrasting noisy inputs with minimal clean supervision.

Bio: Bo Han is currently an Associate Professor in Machine Learning and a Director of Trustworthy Machine Learning and Reasoning Group at Hong Kong Baptist University, and a BAIHO Visiting Scientist of Imperfect Information Learning Team at RIKEN Center for Advanced Intelligence Project (RIKEN AIP), where his research focuses on machine learning, deep learning, foundation models, and their applications. He was a Visiting Research Scholar at MBZUAI MLD (2024), a Visiting Faculty Researcher at Microsoft Research (2022) and Alibaba DAMO Academy (2021), and a Postdoc Fellow at RIKEN AIP (2019-2020). He received his Ph.D. degree in Computer Science from University of Technology Sydney (2015-2019). He has co-authored three machine learning monographs, including Machine Learning with Noisy Labels (MIT Press), Trustworthy Machine Learning under Imperfect Data (Springer Nature), and Trustworthy Machine Learning from Data to Models (Foundations and Trends). He has served as Senior Area Chair of NeurIPS, and Area Chairs of NeurIPS, ICML and ICLR. He has also served as Associate Editors of IEEE TPAMI, MLJ and JAIR, and Editorial Board Members of JMLR and MLJ. He received paper awards, including Outstanding Paper Award at NeurIPS, Most Influential Paper at NeurIPS, and Outstanding Student Paper Award at NeurIPS Workshop, and service awards, including Notable Area Chair at NeurIPS, Outstanding Area Chair at ICLR, and Outstanding Associate Editor at IEEE TNNLS. He received the RGC Early CAREER Scheme, IEEE AI's 10 to Watch Award, IJCAI Early Career Spotlight, INNS Aharon Katzir Young Investigator Award, RIKEN BAIHO Award, Dean's Award for Outstanding Achievement, Microsoft Research StarTrack Scholars Program, and Faculty Research Awards from ByteDance, Baidu, Alibaba and Tencent.

Xiaoge Zhang
The Hong Kong Polytechnic University
xiaoge.zhang@polyu.edu.hk

Implementing Trust in Non-Small Cell Lung Cancer Diagnosis with a

Conformalized Uncertainty-Aware AI Framework in Whole-Slide Images

Ensuring trustworthiness is fundamental to the development of artificial intelligence (AI) that is considered societally responsible, particularly in cancer diagnostics, where a misdiagnosis can have dire consequences. Current digital pathology AI models lack systematic solutions to address trustworthiness concerns arising from model limitations and data discrepancies between model deployment and development environments. To address this issue, we developed TRUECAM, a framework designed to ensure both data and model trustworthiness in non-small cell lung cancer subtyping with whole-slide images. TRUECAM integrates 1) a spectral-normalized neural Gaussian process for identifying out-of-scope inputs and 2) an ambiguity-guided elimination of tiles to filter out highly ambiguous regions, addressing data trustworthiness, as well as 3) conformal prediction to ensure controlled error rates. We systematically evaluated the framework across multiple large-scale cancer datasets, leveraging both task-specific and foundation models, illustrate that an AI model wrapped with TRUECAM significantly outperforms models that lack such guidance, in terms of classification accuracy, robustness, interpretability, and data efficiency, while also achieving improvements in fairness. These findings highlight TRUECAM as a versatile wrapper framework for digital pathology AI models with diverse architectural designs, promoting their responsible and effective applications in real-world settings.

Bio: Xiaoge Zhang is an Assistant Professor in the Department of Industrial and Systems Engineering (ISE) at The Hong Kong Polytechnic University. His research interests center on risk management, reliability engineering, and trustworthiness assurance of AI/ML-powered intelligent systems using uncertainty quantification, knowledge-enabled AI, and fail-safe measures. He received his Ph.D. in Systems Engineering and Operations Research at Vanderbilt University, Nashville, Tennessee, United States in 2019. He has won multiple awards, including Peter G. Hoadley Best Paper Award, Chinese Government Award for Outstanding Self-Financed Students Studying Abroad, Bravo Zulu Award, Pao Chung Chen Fellowship, among others. He has published more than 80 papers in leading academic journals, such as Nature Communications, IEEE Transactions on Artificial Intelligence, IEEE Transactions on Information Forensics and Security, IEEE Transactions on Reliability, IEEE Transactions on Cybernetics, IEEE Transactions on Industrial Informatics, IEEE Transactions on Automation Science and Engineering, Reliability Engineering & Systems Safety, Risk Analysis, Decision

Support Systems, and Annals of Operations Research, among others. He is on the editorial board of Journal of Organizational Computing and Electronic Commerce. He is a member of INFORMS, IEEE and IISE.

Zhiyong Yang
University of Chinese Academy of Sciences
yangzhiyong21@ucas.ac.cn

Pursuing a Proper Allocation of the Probability Mass in Knowledge Distillation

Knowledge Distillation (KD) transfers knowledge from a large teacher model to a smaller student model by minimizing the divergence between their output distributions, typically using forward Kullback-Leibler divergence (FKLD) or reverse KLD (RKLD). It has become an effective training paradigm due to the broader supervision information provided by the teacher distribution compared to one-hot labels. In this talk, we identify that the core challenge in KD lies in balancing two mode-concentration effects: the Hardness-Concentration effect, which refers to focusing on modes with large errors, and the Confidence-Concentration effect, which refers to focusing on modes with high student confidence.

Through an analysis of how probabilities are reassigned during gradient updates, we observe that these two effects are entangled in FKLD and RKLD, but in extreme forms. Specifically, both are too weak in FKLD, causing the student to fail to concentrate on the target class. In contrast, both are too strong in RKLD, causing the student to overly emphasize the target class while ignoring the broader distributional information from the teacher.

To address this imbalance, we propose ABKD, a generic framework with α-β-divergence. Our theoretical results show that ABKD offers a smooth interpolation between FKLD and RKLD, achieving a better trade-off between these effects. Extensive experiments on 17 language/vision datasets with 12 teacher-student settings confirm its efficacy.

Bio: Zhiyong Yang is an Associate Professor and Ph.D. advisor at the University of Chinese Academy of Sciences. His research focuses on decision-invariant machine learning methods and theories. He has contributed to the development of the XCurve framework (https://xcurveopt.github.io/). He has published over 30+ in TPAMI/ICML/NeurIPS He serves as an Area Chair for ICML, NeurIPS, and ICLR. He has received multiple honors, including the CCF Doctoral Dissertation Award (formerly known as the CCF

Excellent Doctoral Dissertation Encouragement Program), the championship of the 2024 NeurIPS Large Model Safety Challenge, the 2025 CVPR Fine-Grained Video Understanding Challenge, and the 2025 CVPR Compositional 3D Vision Challenge, as well as the Asian Trustworthy Machine Learning (ATML) Fellowship.

Shuren Qi
The Chinese University of Hong Kong
shurenqi@cuhk.edu.hk

Rethink Deep Learning with Invariance in Data Representation

Integrating invariance into data representations is a principled design in intelligent systems. Representations play a fundamental role, where systems and applications are both built on meaningful representations of digital inputs (rather than the raw data). In fact, the proper design/learning of such representations relies on priors w.r.t. the task of interest. Here, the concept of symmetry from the Erlangen Program may be the most fruitful prior — informally, a symmetry of a system is a transformation that leaves a certain property of the system invariant. Symmetry priors are ubiquitous, e.g., translation as a symmetry of the object classification, where object category is invariant under translation.

The quest for invariance is as old as pattern recognition itself. Invariant design has been the cornerstone of various representations in the era before deep learning, such as the SIFT. As we enter the early era of deep learning, the invariance principle is largely ignored and replaced by a data-driven paradigm, such as the CNN. However, this neglect did not last long before they encountered bottlenecks regarding robustness, interpretability, efficiency, and so on. The invariance principle has returned in the era of rethinking deep learning, forming a new field known as Geometric Deep Learning (GDL).

In this talk, I will give a historical perspective of the invariance in data representations. More importantly, I will introduce research dilemmas, promising works, future directions, and our contributions.

Bio: Shuren Qi is currently a Postdoctoral Fellow with Department of Mathematics, The Chinese University of Hong Kong. His research focuses on Geometric Deep Learning, with applications in Trustworthy AI and Science AI. He has authored 12 papers in top-tier journals and conferences, such as

IEEE TPAMI and USENIX Security. His works offer some new designs of invariant representations — from global to local and hierarchical assumptions. More information is available at https://shurenqi.github.io/.