Statement of Purpose

Driven to improve AI's interpretability and safety, I have developed a keen interest in trustworthy AI and causality. My solid background in this field fuels my enthusiasm to further my research within a PhD program in Computer Science, under the guidance of esteemed mentors.

Before my current research focus, I delved into heuristic studies in computer vision (CV) and natural language processing (NLP). Under Prof. Weishi Zheng at SYSU and before my MPhil at HKUST, I pioneered a project in mixture generation, blending various image concepts to create new domains unseen during training. This solo-driven work resulted in a first-author paper at IJCAI. My first MPhil semester in 2020 included an internship at Noah's Ark Lab, where I worked on integrating lexicons into BERT to improve sequence labeling, contributing to a second-author paper in the EMNLP 2021 Long Paper. Next, I will briefly introduce my previous and future work, as depicted in Fig 1.

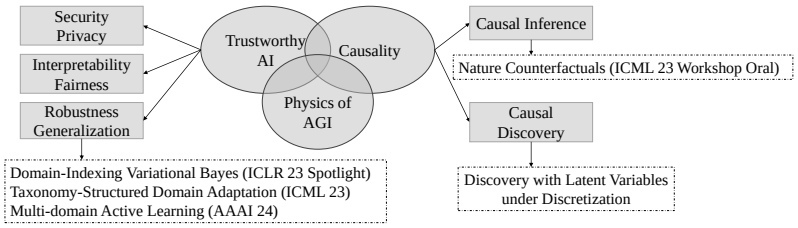

Figure 1: The main focus of previous and future work.

Research in Trustworthy AI. Publishing two papers and completing my MPhil coursework across various AI disciplines, including ML, CV, NLP, and data mining, laid a robust groundwork in AI and programming. These experiences sharpened my focus on developing general machine learning methods with theoretical guarantees and interpretability. Since 2021, under Prof. Hao Wang (Rutgers) 's advisement, I have delved into trustworthy AI, emphasizing out-of-distribution generalization for AI system robustness, leading to three notable projects.

In domain adaptation research, our team discovered that using domain indices—real-value vectors embedding domain semantics (like "age" in medical datasets)—enhances domain adaptation more effectively than traditional, semantically void domain identities (such as one-hot labels). To tackle the issue of often unavailable semantic domain indices, we proposed inferring them as latent variables from multi-domain data. We introduced VDI, an adversarial variational Bayesian deep-learning model, to formalize domain index inference. This model delineates the conditional dependencies among data inputs, labels, encodings, and domain indices. My independent theoretical work, grounded in information theory, demonstrated that maximizing VDI's evidence lower bound optimally infers domain indices. Our experiments showcased VDI's capability to infer meaningful domain indices, achieving state-of-the-art (SOTA) performance and resulting in a first-author Spotlight paper at ICLR 2023.

The second project leveraged existing semantic domain indices and specifically addressed generalization challenges in taxonomy-structured domains, which exhibit nested, hierarchical similarities (e.g., animal species, product catalogs). We extended the classic adversarial framework by introducing a "taxonomist" model to preserve taxonomy information. I developed a core theory to identify conditions under which data features retain taxonomy information, leading to a publication in ICML 2023.

Building on these experiences, my recent research aimed to address the limitations of traditional active learning (AL) models, which primarily focus on single domains and struggle with out-of-distribution scenarios in multi-domain settings, such as medical data from various hospitals. Introducing Composite Active Learning (CAL), a novel method tailored for multi-domain AL, we approached domain-level budget allocation based on domain importance while managing distribution shifts. I developed a theoretical framework demonstrating that CAL achieves a lower generalization error bound than existing AL methods, validated through empirical testing across diverse datasets. This research has resulted in a first-author paper currently under review for AAAI 2024, and the positive feedback received includes one "Accept" and three "Weak Accept" evaluations.

Each of my projects emphasizes interpretability, as demonstrated through detailed visualization techniques in their respective papers.

Research in Causality. Causality plays a pivotal role in enhancing AI's interpretability and performance. As a research assistant mentored by Prof. Kun Zhang (CMU) and Prof. Jiji Zhang (CUHK), I

have been actively engaged in the realm of causal inference and discovery since 2022.

My initial project addressed a critical issue in causality: the challenge of implausible interventions in non-backtracking counterfactuals, a concept developed by Judea Pearl. Consider a scenario where Tom, on a suddenly braking bus, falls and injures Jerry. Traditional non-backtracking counterfactuals suggests that if Tom had not fallen, Jerry would not have been injured. However, this ignores the implausibility of Tom not falling due to the laws of physics. To address this, we introduced "natural counterfactuals," which consider interventions plausible within the observed data. For instance, a more realistic intervention might involve the bus slowing down earlier, necessitating some degree of backtracking. Our approach minimizes backtracking, ensuring changes to causally preceding variables are limited and the counterfactual world alteration is minimal. Independently, I developed an optimization framework for generating natural counterfactuals. Empirical results using both simulated and real data showcased the method's superiority over traditional approaches, leading to an oral presentation at the ICML 2023 workshop, selected as one of five among 38 accepted papers, and a modified version submitted to ICLR 2023.

My ongoing project explores causal discovery with latent variables in the context of data discretization. While previous research has tackled causal discovery with latent variables under linear Gaussian assumptions and continuous data, categorical data—common in fields like biology, psychology, and economics—pose new challenges. These data types often result from discretizing continuous data, rendering standard hypothesis testing methods inapplicable. I am developing a method using multi-polychoric correlations and adapting original testing methods for discretized data. The preliminary experiments are promising, and the next step is to theoretically validate this approach.

Research Plan Before PhD. In addition to my current causality research, I am collaborating with Prof. Yuanzhi Li and Prof. Kun Zhang at CMU on a project exploring the physics of Large Language Models (LLMs). Our objective is to delve into LLMs' capacity to comprehend concepts and elucidate their behaviors, potentially from a causal perspective, to improve their trustworthiness.

Future Research Direction. My overarching mission is to advance the interpretability, robustness, fairness, and security of AI systems.

During my PhD, I am eager to delve more deeply into the realm of Safe and Trustworthy AI. In an era marked by the growing influence of AI technologies and their potential for misuse, it is imperative to engineer secure and transparent AI solutions. My research will maintain its focus on cross-domain generalization, a crucial element in enabling AI to learn and adapt effectively to a multitude of environments. For example, it is evident that even GPT-4's performance is compromised when faced with previously unseen data. Beyond my potential exploration of interpretability, fairness, and security, an exciting direction is studying the physics of Artificial General Intelligence (AGI), particularly understanding how intelligence emerges in LLMs. Developing corresponding theories may pave the way for creating more robust and secure AGI systems.

Furthermore, I intend to focus on causality, unlocking scientific discoveries and providing clarity to large AI systems, including LLMs. Key aspects of my work will involve causal discovery to map the graph of latent variables responsible for generating generate natural languages, images, videos, and other modalities. This effort aims to construct fundamentally interpretable causal models, enabling easier detection and understanding of issues that may arise.