Tractable Probabilistic Reasoning for Trustworthy AI

Automated decision-making systems are increasingly being deployed in areas with personal and societal impacts, leading to growing interest and need for trustworthy AI and ML systems—that is, models that are robust, explainable, fair, and so on. It is important to note that these guarantees only hold with respect to a certain model of the world, with inherent uncertainties. In this talk, I will present how probabilistic modeling and reasoning, by incorporating a distribution, offer a principled way to handle different kinds of uncertainties when learning and deploying trustworthy AI systems. For example, when learning classifiers, the labels in the training data may be biased; I will show that probabilistic circuits, a class of tractable probabilistic models, can be effective in enforcing and auditing fairness properties by explicitly modeling a latent unbiased label. In addition, I will also discuss recent breakthroughs in tractable inference of more complex queries such as information-theoretic quantities, to demonstrate the potential of probabilistic reasoning for trustworthy AI. Finally, I will conclude with my future work towards a framework to flexibly reason about and enforce trustworthy AI/ML system behaviors.

Bio: YooJung Choi is a PhD student in the Computer Science department at UCLA, advised by Guy Van den Broeck. She is part of the Statistical and Relational Artificial Intelligence (StarAI) lab. Her research interests are in tractable probabilistic models, graphical models, knowledge compilation, and trustworthy AI/ML (robustness, fairness, explainability, and more).