Probabilistic Reasoning and Learning for Trustworthy AI*

YooJung Choi

School of Computing and Augmented Intelligence
Arizona State University
yj.choi@asu.edu

As automated decision-making systems are increasingly deployed in areas with personal and societal impacts, there is a growing demand for artificial intelligence and machine learning systems that are fair, robust, interpretable, and generally trustworthy. Ideally we would wish to answer questions regarding these properties and provide guarantees about any automated system to be deployed in the real world. This raises the need for a unified language and framework under which we can reason about and develop trustworthy AI systems. This talk will discuss how tractable probabilistic reasoning and learning provides such framework.

It is important to note that guarantees regarding fairness, robustness, etc., hold with respect to the distribution of the world in which the decision-making system operates. For example, to see whether automated loan decisions are biased against certain gender, one may compare the average decision for each gender; this requires knowledge of how the features used in the decision are distributed for each gender. Moreover, there are inherent uncertainties in modeling this distribution, in addition to the uncertainties when deploying a system in the real world, such as missing or noisy information. We can handle such uncertainties in a principled way through probabilistic reasoning. Taking fairness-aware learning as an example, we can deal with biased labels in the training data by explicitly modeling the observed labels as being generated from some probabilistic process that injects bias/noise to hidden, fair labels, particularly in a way that best explains the observed data.

A key challenge that still needs to be addressed is that: we need models that can closely fit complex real-world distributions—i.e. *expressive*—while also being amenable to exact and efficient inference of probabilistic queries—i.e. *tractable*. I will show that probabilistic circuits, a family of tractable probabilistic models, offer both such benefits. In order to ultimately develop a common framework to study various areas of trustworthy AI (e.g., privacy, fairness, explanations, etc.), we need models that can flexibly answer different questions, even the ones it did not foresee. This talk will thus survey the efforts to expand the horizon of complex reasoning capabilities of probabilistic circuits, especially highlighted by a modular approach that answers various queries via a pipeline of a handful of simple tractable operations.

*Presented in the New Faculty Highlights Program