AFRL SUNY Trusted AI Challenge Series

Topic #4: Trustworthy AI Certification

Sponsors: IBM and SUNY

Objective: This topic seeks to create novel methodologies and tools for eliciting and fusing accuracy, fairness, robustness, and explainability metrics and beliefs to certify trustworthiness of AI systems.

Description: AI systems that are worthy of trust in mission-critical and high-stakes domains have high aptitude not only in basic accuracy, but also in distributional robustness, adversarial robustness, algorithmic fairness, and explainability. In recent years, there has been significant progress in developing metrics for these considerations beyond accuracy and in toolkits for computing them. However, the metrics for the different pillars of trust are operationalized in different ways depending on the precise characteristics of the application along with its regulations and policies, and also depending on the normative stances and politics of the consumers of the metrics. It is also not clear how the different elements of trust interact with each other and with accuracy: are they tradeoffs or non-tradeoffs that can be achieved simultaneously? Thus, certification of AI systems for trustworthiness is not as simple as only computing fairness, robustness, and explainability metrics.

Guidance: Using expertise from human factors engineering, machine learning, software engineering, game theory, cognitive psychology, policy, law, and other disciplines, we challenge researchers to develop novel methods for:

- Eliciting feasible trust-related policies from committees of multiple stakeholders, including people from traditionally marginalized groups,
- Operationalizing these trust policies as means for certifying AI models, including by fusing beliefs towards different elements of trust in summaries that can be regulated, and
- Creating end-to-end tools for developers and third parties to govern AI models.

Summary: Certifying AI systems for trustworthiness is not straightforward because trustworthiness involves several dimensions whose interrelationships are not clear and that have different politics. Develop components of a system that measures trust metrics and elicits acceptable ranges for these trust metrics from stakeholders.