GSAW 2021 Tutorial M:

Trusted Artificial Intelligence

Overview:
Artificial Intelligence systems – enabled by advancement in sensor and control technologies, artificial intelligence, data science, and machine learning – promise to deliver new and exciting applications to a broad range of industries. However, a fundamental trust in their application and execution must be established in order for them to succeed. People, by and large, do not trust a new entity or system in their environment without some evidence of trustworthiness. To trust an artificial intelligence system, we need to know which factors affect system behaviors, how those factors can be assessed and effectively applied for a given mission, and the risks assumed by trusting.

This course aims to provide a foundation for building trust in artificial intelligence systems. Elements of artificial intelligence systems are defined, and in that context, the perception of trust is explored. A framework for evaluating trust highlights several key aspects: Objective/Data Specification, Reproducibility, Confidence & Uncertainty, Adversarial Robustness, Interpretability, Fairness, Monitoring and Control. An overview of the state-of-the art in research, methods, and technologies for achieving trusted artificial intelligence systems is reviewed along with current applications. The course concludes by identifying important open issues and outlining a roadmap for Trusted Artificial Intelligence Systems.

Outline

- AI systems – needs, environments, and challenges
    - Operational context (real-time vs non-real time / critical vs non-critical)
    - Differentiating training and deployment
    - Facets for Evaluating Trust in AI Systems
- Challenges for trust
    - Human perception of trust
    - Drivers for defining levels of trust
    - Examples of trust challenges
- Overview of Trusted AI Framework
    - Objective/Data Specification
    - Reproducibility
    - Confidence & Uncertainty
    - Adversarial Robustness
    - Interpretability
    - Fairness
    - Monitoring
    - Control
- Topic Focus: Adversarial Robustness
    - AI model attacks
    - AI model defenses
    - Securing AI models

Instructors: Andrew Brethorst, The Aerospace Corporation and Erik Linstead, Chapman University

Biographies:
Andrew Brethorst is the Associate Department Director for the Data Science and AI Department at The Aerospace Corporation. Mr. Brethorst completed his undergraduate degree in cybernetics from UCLA, and later completed his master's degree in computer science with a concentration in machine learning from UCI. Much of his work involves applying machine learning techniques to image

exploitation, telemetry anomaly detection, intelligent artificial agents using reinforcement learning, as well as collaborative projects within the research labs.

Dr. Erik Linstead is a professor in AI at Chapman University. Dr. Linstead completed his undergraduate degree in computer science from Stanford University, and later went on to complete his PhD in Artificial Intelligence and machine learning from UC Irvine. He currently operates a research lab where he focuses on using AI technology for enhancing learning as well as studying new treatment affects for autism.

Description of Intended Students and Prerequisites:
This course is intended for decision makers, program managers, chief engineers, systems architects, analysts, AI scientists and practitioners from defense-related businesses interested in the application and ramifications of trusted artificial intelligence systems. Having prior knowledge or training in AI, while encouraged, is not required.

What can Attendees Expect to Learn:
Learning Objectives

1. Need for artificial intelligence systems and the environments in which they may operate.
2. Elements of artificial intelligence systems and challenges for trust – sensor data interpretation, rapid stimulus-response, learning, high-level cognitive models of planning and decision-making.
3. Discussion regarding the need for a framework for evaluating trust in artificial intelligence systems – data, algorithms, and cyber considerations
4. State-of-the-art and examples in evaluation of trust in artificial intelligence systems