Trustworthy AI: New Properties for New Complex Systems

Submitted by Jeannette Wing on Fri, 08/14/2020 - 1:19pm. Contributor: Jeannette Wing

Recent years have seen an astounding growth in deployment of AI systems in critical domains such as autonomous vehicles, criminal justice, healthcare, hiring, housing, human resource management, law enforcement, and public safety, where decisions taken by AI agents directly impact human lives. Consequently, there is an increasing concern if these decisions can be trusted to be correct, reliable, fair, and safe, especially under adversarial attacks. How then can we deliver on the promise of the benefits of AI but address these scenarios that have life-critical consequences for people and society? In short, how can we achieve trustworthy AI?

Under the umbrella of trustworthy computing, there is a long-established framework employing formal methods and verification techniques for ensuring trust properties like reliability, security, and privacy of traditional software and hardware systems. Just as for trustworthy computing, formal verification could be an effective approach for building trust in AI-based systems. However, the set of properties needs to be extended beyond reliability, security, and privacy to include fairness, robustness, probabilistic accuracy under uncertainty, and other properties yet to be identified and defined. Further, there is a need for new property specifications and verification techniques to handle new kinds of artifacts, e.g., data distributions, probabilistic programs, and machine learning based models that may learn and adapt automatically over time. This talk will pose a new research agenda, from a formal methods perspective, for us to increase trust in AI systems.
Jeannette Wing
License: Creative Commons 2.5