Developing Trustworthy, Causal, and Aligned Artificial Intelligence

Arjan Durresi
Indiana University Indianapolis, USA
adurresi@iupui.edu

This talk will discuss the engineering of trust among agents, humans, and algorithms to develop practical solutions. The problems we will address include Trustworthy and Causal AI in various applications, system resilience, and creating a framework for Artificial Conscience to control AI. The trustworthiness of AI solutions is now considered essential for optimal AI usage. We have developed metrics using our trust system to evaluate AI solutions' acceptance, explainability, and fairness. Additionally, we emphasize the importance of causality as a crucial component of trustworthy AI. An essential application of Causal AI is the Manufacturing production line Root Cause Analysis. Lastly, we present our framework for Artificial Conscience, in which AI algorithms are controlled by agents who negotiate using our trust engine to produce a solution that maximizes "Artificial Feeling." This framework can be easily implemented in any AI system, incorporating metrics such as trustworthiness and causality, ultimately aligning AI with human values.