JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Electrical and Computer Engineering

Electrical & Computer Engineering Seminar

Building Safe and Reliable AI: Perspectives from robustness, uncertainty estimation and privacy.

Krishnamurthy Dvijotham

Oct. 17, 2023

Krieger 170

3 PM – 4 PM

Abstract

As we make ever-accelerating progress on AI systems with powerful capabilities, the challenge of building AI systems that are safe and reliable becomes increasingly pressing. I will discuss recent work on three specific challenges: 1) How do we measure, certify and enhance robustness of deep learning systems to test-time adversarial attacks? 2) How do we train AI systems to be aware of their uncertainty and defer back to a user when necessary? 3) How do we address these challenges while preserving the privacy of users the AI was trained on?

Bio

Krishnamurthy (Dj) Dvijotham is a staff research scientist in the Brain team at Google Research. His research focuses on building safe, reliable and trustworthy AI, including concerns like privacy, uncertainty estimation, fairness and formal verifiability.