Ensuring Safety and Trust in AI-Based Autonomous Systems

RAVI SHANKAR

Synthetic Wisdom LLC.

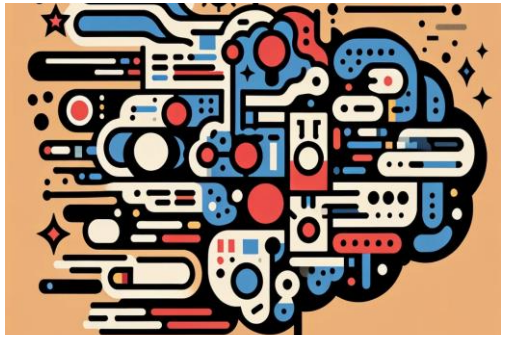
rshankar@synth-wisdom.com

RAMESH BHARADWAJ

US Naval Research Laboratory

ramesh.bharadwaj.civ@us.navy.mil

Increased Risk in AI based Systems



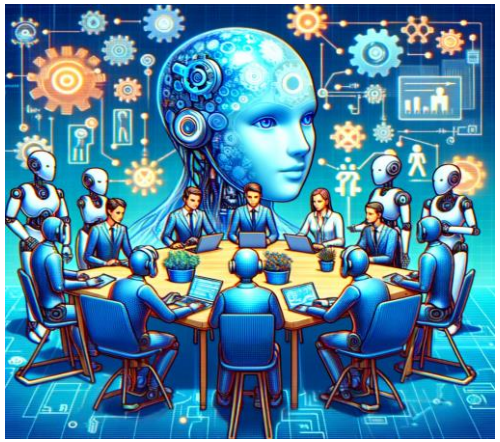
Defects And Vulnerabilities In AI Generated Code.

Hallucination

Lack Of Supervision And Transparency While Building The System

Gen-AI based Framework to build Confidence

Leverage Existing Approaches For Safety Critical Systems.



Introduce AI-Based Domain Experts and Collaborate

Strengthen With Critique Agents with Fact Grounding.

Hybrid Framework with Checks and Balances

Adopt and enhance existing safety assurance process by:

Adding Gen-AI based Domain experts (finetuned, RAG, specially pretrained)

Automating the process with multi-agent collaboration using mandatory human intervention

Bolstering confidence with multiple critique-loops and fact based grounding

Automated test-case generation to probe faults and derive qualification suites

Requalification after maintenance or modifications to the system

Transparent, auditable, explainable process

Let AI be your trusted colleague!