AI-enabled capabilities are increasingly being incorporated into mission planning, perception, analytics, and decision-support workflows across defense environments. As these systems move closer to operational use, programs need practical ways to evaluate whether model behavior remains reliable and consistent under representative operating conditions. Existing software assurance processes and documentation reviews often provide limited visibility into how AI models behave under stress, how failures can be reproduced, or how weaknesses may affect downstream mission workflows.

This presentation describes an evidence-based workflow for evaluating AI model behavior during development, integration, and pre-deployment review activities. The approach applies structured analysis to vision models and large language models using representative evaluation inputs, including mission-relevant datasets, test vectors, prompt sets, and scenario-driven assessments. The resulting outputs help engineering and program teams better understand model weaknesses before operational integration or major test events.

The workflow evaluates factors that may affect dependable performance, including adversarial susceptibility, instability across operating ranges, sensitivity to environmental conditions or input patterns, and model components associated with unreliable outcomes. Findings are packaged with supporting evidence, reproduction steps, and documented assessment boundaries so technical reviewers can independently validate and assess the results.

The session will discuss how this type of AI assurance workflow can support emerging modeling and simulation environments, including simulation-enabled decision support, digital engineering pipelines, synthetic training environments, and AI-enabled mission rehearsal activities. As AI components become more common within these environments, there is increasing need for repeatable methods that help characterize model behavior prior to integration into operationally relevant simulations or mission-support systems.

Examples from FortiLayer, an AI assurance implementation developed by ObjectSecurity, will be used to illustrate how evidence-backed model evaluation can be incorporated into existing engineering and test workflows. Discussion topics will include practical evaluation methods, evidence artifacts useful to reviewers and engineers, lessons learned from pilot-style assessments, and current limitations associated with automated AI assurance techniques in defense environments.

This work presents AI assurance as an additional analysis capability that can help identify model weaknesses earlier in the lifecycle and support more informed engineering and risk-management decisions for emerging AI-enabled mission capabilities.