CALYPSOAI

The Rise of Adversarial ML

and how independent testing and validation supports secure AI deployments.

Evasion Attacks

What it means

Models are fed misleading information during deployment, potentially tricking them and changing their predictions. Untargeted evasion attacks do not care how the model prediction changes.

The risks

In a target identification mission, a model is tricked into interpreting a school bus as an enemy humvee.

How independent testing supports mission success

Gain advance knowledge on your model's resilience to these attacks, so predictions are not impacted by outside activity during deployment.

Model Inversion Attacks

What it means

The attacker accesses an ML model and seeks to infiltrate its private training data. This data can be used to circumvent the models following deployment.

The risks

In an ISR mission, an attacker learns what objects the model is capable of detecting, and what it cannot detect. This intelligence can help adversaries avoid detection.

How independent testing supports mission success

Keep your training data secure by ensuring robustness against model inversion attacks, and build stakeholder confidence that models cannot be bypassed or tricked.

Be prepared for the unexpected

Adversarial ML is an evolving field and the intrusions and methods threat actors use to cause harm are likely to change. Ongoing testing is key to staying ahead of the anticipated evolution of these technologies on the front lines of the AI battlespace.

For more information on how CalypsoAI supports AI security, visit calypsoai.com