[Service Brief]

# AI Security Assurance

## Advisory and Managed Services for AI Security Assessments, Adversarial Testing, Red Teaming, and Shadow AI Oversight

## OVERVIEW

Generative AI, autonomous agents, and retrieval pipelines have introduced attack surfaces that traditional security programs were not designed to address. Most enterprises are deploying AI faster than security teams can assess it, and the controls that exist were built for a different environment.

**SDG's AI Security Assurance gives enterprises the visibility, adversarial validation, and governance structure to operate AI at scale without accumulating exposure they cannot see or defend.** The service redefines AI protection with a dual-track delivery model that combines advisory-led assessments with a managed assurance capability. This delivers continuous visibility, adversarial validation, and production-grade controls across the full AI lifecycle.

By leveraging the AI Security Assurance service, enterprises achieve:

- **Faster AI Risk Visibility:** Move from reactive discovery to proactive identification of sanctioned and shadow AI exposure within the first two weeks of engagement.
- **Adversarial Validated Controls:** Evidence-backed findings across prompt injection, jailbreak, data exfiltration, and agent misuse scenarios.
- **Continuous Assurance at Enterprise Scale:** Managed monitoring of AI posture, misuse indicators, and control drift across models, agents, APIs, and data pipelines.
- **Future-Proofed AI Governance:** Alignment to NIST AI RMF, MITRE ATLAS, OWASP LLM Top 10, and Google SAIF, ready for EU AI Act and emerging regulatory obligations.

## OVERVIEW

Modern enterprises face multiple barriers when trying to secure AI adoption at scale:

1. **Rapid GenAI Rollout Without Security Validation**

   Business units and development teams deploy copilots, agents, and RAG pipelines faster than security teams can assess them. Critical decisions that include model selection, data grounding, access boundaries, and guardrail design are often made case-by-case. As a result, inconsistent risk posture and delayed remediation increases.

2. **Shadow AI Blind Spots**

   Employees routinely use unsanctioned AI tools, paste sensitive data into public models, and adopt browser-based assistants without IT visibility. Most security programs have no inventory of where AI is used within the business.

3. **Novel Attack Surface, Unprepared Defenses**

   Prompt injection, jailbreak, model exfiltration, training data poisoning, unsafe outputs, and agent misuse are not covered by traditional application security testing. Most existing pen-test programs lack the tooling and methodology to assess AI systems credibly.

4. **Limited Runtime Visibility**

   Without AI-aware telemetry, misuse indicators, and posture dashboards, security leaders cannot detect prompt abuse, data leakage, or anomalous agent behavior once AI is in production.

5. **Governance and Regulatory Gaps**

   Manual, one-time reviews miss control drift, new AI adoption patterns, and evolving regulatory obligations. The EU AI Act, NIST AI RMF, and sector-specific guidance require continuous, auditable evidence over point-in-time snapshots.

6. **Change Management Complexity**

   Application teams must adopt secure-by-design patterns, implement guardrails, harden retrieval and integration paths, and instrument AI workloads for monitoring. All require coordinated guidance across AppDev, platform, identity, and SOC teams.

## SOLUTION: THE AI SECURITY ASSURANCE FRAMEWORK

**SDG's AI Security Assurance service** delivers a repeatable, evidence-backed model for securing enterprise AI across the full lifecycle, from discovery and assessment through adversarial testing, control engineering, and ongoing managed assurance. The methodology is aligned to NIST AI RMF, MITRE ATLAS, OWASP GenAI/LLM risks, and Google's SAIF.

### Service Delivery Model

| Phase | Objective | Key Deliverables |
|---|---|---|
| **1. Architect & Align** | Confirm in-scope AI systems, data flows, APIs, owners, and shadow AI signals. Align success criteria and evidence needs. | Approved scope, RACI, success criteria, test authorization. |
| **2. Discover & Assess** | Baseline sanctioned and shadow AI. Review architecture, guardrails, access, configurations, data paths, and trust boundaries. | AI asset inventory, shadow AI signal map, posture baseline. |
| **3. Adversarial Test & Red Team** | Execute prompt injection, jailbreak, exfiltration, unsafe output, and agent misuse scenarios with reproducible evidence. | Findings register, exploit paths, scenario evidence pack. |
| **4. Secure & Enforce** | Prioritize fixes across identity, API, data, infrastructure, and GenAI controls. Define guardrails and validation checkpoints. | Remediation roadmap, control requirements, retest criteria. |

| Phase | Objective | Key Deliverables |
|---|---|---|
| **5. Monitor & Respond** | Define telemetry, misuse indicators, incident triggers, SOC/SIEM integration, and retest scenarios for managed assurance. | Detection content, runbooks, monitoring dashboards. |
| **6. Govern & Evolve** | Refresh risk register, control mapping, KPIs, exceptions, and recurring assurance cadence for advisory closeout or managed handoff. | Executive readout, governance cadence, assurance schedule. |

**This dual-track model moves AI security from a point-in-time assessment into a continuous assurance capability, with visibility, validation, and governance operating as an ongoing function rather than a periodic exercise.**

## SERVICE CAPABILITIES

1. **AI and Shadow AI Discovery**

   Structured discovery of sanctioned AI systems, unsanctioned tool usage, agents, APIs, data paths, embeddings, vector stores, and non-human identities. Delivers a defensible inventory as the foundation for every downstream control.

2. **AI Security Assessment**

   Architecture, posture, trust boundary, and control review across models, agents, RAG pipelines, MCP servers, tool-use integrations, and data exposure paths. All mapped to NIST AI RMF, OWASP LLM Top 10, and client-specific policy.

3. **Adversarial Testing and Red Teaming**

   Scenario-driven testing using GARAK, PyRIT, Promptfoo, and custom SDG test cases. Covers prompt injection, jailbreak, training data extraction, unsafe output generation, agent misuse, and multi-step autonomous agent abuse paths.

4. **Runtime Guardrails and Control Engineering**

   Prescriptive hardening recommendations across input/output filtering, policy enforcement, identity for AI workloads, secrets management, API protection, and SOC integration. Every recommendation is scoped for production environments with a clear path to implementation.

5. **Managed AI Posture and Misuse Monitoring**

   Continuous monitoring of AI posture drift, new AI adoption patterns, misuse indicators, and control health. Delivered through SDG's Cyber Defense Center with telemetry tuned for GenAI-specific threat models.

6. **Governance, Audit, and Regulatory Alignment**

   Recurring control mapping, KPI reporting, exception tracking, and audit-ready evidence aligned to NIST AI RMF, MITRE ATLAS, OWASP, SAIF, and emerging obligations such as the EU AI Act and sector-specific AI guidance.

## KEY BENEFITS

- **Faster Time-to-Insight:** Initial findings typically surface within the first two weeks of engagement, with full adversarial results by week four to six, giving security programs early, actionable evidence.
- **Evidence-Backed Assurance:** Reproducible findings with severity, exploit path, affected controls, and scenario evidence, replacing checkbox attestations with defensible, retestable proof.
- **Reduced AI Risk Exposure:** Systematic reduction of exposure to prompt abuse, data leakage, unsafe outputs, and agent misuse across the application, API, identity, infrastructure, data, and model layers.
- **Operational Scalability:** Managed assurance enables lean security teams to sustain AI oversight across growing estates without added headcount or reduced audit-ready rigor.
- **Stronger Governance and Audit Readiness:** Continuous control mapping, risk register refresh, and recurring assurance reporting that is ready for board, audit, and regulatory scrutiny including EU AI Act obligations.
- **Unified Identity, Risk, and Threat Posture:** Delivered across SDG's three service towers — Identity (including NHI for agents), Risk (governance and GRC alignment), and Threat (red teaming and Cyber Defense Center monitoring) — for a coordinated AI security posture.

## DIFFERENTIATORS

SDG's AI Security Assurance spans the full AI security lifecycle, from initial discovery through continuous managed assurance. The service is built around five capabilities that work together rather than operating as standalone offerings:

- **Dual-Track Advisory and Managed Model:** A single service spanning one-time assessment and continuous managed assurance with a clear, pre-defined handshake between the two.
- **Shadow AI as a First-Class Scope Item:** Discovery of unsanctioned AI usage, data flows, and policy gaps is built into every engagement.
- **Adversarial Validation, Not Paper Assurance:** Hands-on red teaming using GARAK, PyRIT, Promptfoo, and SDG-developed scenario libraries to deliver reproducible, evidence-backed findings.
- **Framework-Anchored Methodology:** Aligned to NIST AI RMF, MITRE ATLAS, OWASP LLM Top 10, and Google SAIF to provide a defensible basis for board, audit, and regulatory conversations.
- **Continuous Improvement Loop:** Findings, detection content, and control patterns from every engagement feed back into the delivery methodology, improving accuracy and coverage over time.

## WHAT SDG DELIVERS | AI VISIBILITY, GOVERNANCE, AND SECURITY

**SDG helps clients see AI, govern AI, and secure AI — with identity at the core.**

### 01 VISIBILITY

#### Know Every AI Agent and its Access

Inventory sanctioned and shadow AI across agents, copilots, models, identities, data, and connected systems.

*Client Value: A prioritized baseline for ownership, risk reduction, and faster decisions.*

### 02 GOVERNANCE

#### Assign Ownership and Enforce Control

Define policy, decision rights, lifecycle governance, and least-privilege access aligned to NIST AI RMF, EU AI Act, and ISO 42001.

*Client Value: An audit-ready AI operating model leadership can scale with confidence.*

### 03 SECURITY

#### Protect AI from Misuse and Attack

Assess, adversarial test, and continuously monitor AI across prompts, agents, models, tools, and human interaction points.

*Client Value: Reduced exposure to prompt abuse, deepfakes, agent misuse, and control drift.*

## ABOUT US

With more than 30 years of experience partnering with global enterprises on complex business and IT initiatives, SDG is a trusted provider of advisory, transformation, and managed services. The firm empowers organizations to strengthen cyber resilience by integrating AI into identity, threat, and risk management solutions that protect digital assets and deliver measurable business value. Learn more at www.sdgc.com.

SDG

75 North Water Street
Norwalk, CT 06854
203.866.8886
sdgc.com

Contact Us: solutions@sdgc.com