ASSESSED INTELLIGENCE

ASSESSED SIGNAL

March 2026

At the Intersection of Policy, Use & Threat Intelligence

In 2026, cybersecurity must transition from traditional patching to a systems theory approach to manage "Agentic AI" and autonomous workflows that operate without human intervention. To maintain control and meet new legal mandates like the EU AI Act, organizations are adopting the ARISE Framework, which integrates human-in-the-loop oversight and "Zero Trust" architectures for data provenance. This "Shift Left" methodology transforms ethics into a functional control variable, ensuring that AI agents remain reproducible and traceable to human-verified sources.

Secure & Responsible Technology.

AssessedIntelligence.com

When Security Becomes Responsible Understanding

Given the recent revelations of the Meta exploit, as well as the Deloitte/Government of Australia debacle, it has become very clear that understanding the threats posed by AI have not only changed how we need to think about cybersecurity, but more importantly – they have changed how we should be approaching the discipline of cybersecurity. It is no longer about OS level patches, CVEs, APTs or phishing. It is about a system of these things that can work behind the scenes to create viral workflows that are not ever touched by a human. The new approaches to cybersecurity require a sanquine look at what AI is, and what AI is not.

The Erosion of the "System of Systems"

Much like the complexities of “the system of systems” interactions we see in Enterprise IT (EIT) and Operational Technology (OT) overlays, the new approach to responsibly understanding AI security requires practitioners to consider not only the system of systems they have in their organization, but also the system of systems AI creates when added to the organizational process workflows.

The Rise of Agentic Attack Surfaces

In 2026, nearly half of cybersecurity professionals identify Agentic AI and autonomous systems as their top security concern, outranking even deepfakes. When an AI system can build in real-time workarounds for kernel level access that it was not initially granted, it’s time to take a hard look at what your agentic AI system is actually capable of accomplishing, without any human interaction.

https://www.kiteworks.com/cybersecurity-risk-management/agentic-ai-attack-surface-enterprise-security-2026/

https://codewithvamp.medium.com/owasp-top-10-for-agentic-applications-2026-the-ultimate-guide-to-securing-ai-agents-2459c39e37a9

In the same way that EIT processes interact with physical processes in EIT/OT environments, we need to take a systems theory approach to analyzing and responsibly securing these “systems of systems.” Instead of investigating the individual parts of the system (i.e., Agents, IT processes, OT processes, etc.), we need to consider the system as a whole, specifically focusing on "Non-Human Identities" that require broad API permissions to function. We must evaluate how much access the parts comprising the whole can be managed—or not managed. This requires considering how to harden an operating system against rogue AI agent operations, such as Agent Goal Hijacking (ASI01) or Tool Misuse (ASI02). It also requires a deep understanding of how much access the agents have to every part of your operating systems as well as how these interactions function in your operating environment.

Control Theory and Feedback Loops

Control theory, along with systems theoretical approaches to this new problem set, offers many responsible and applicable options. Control theory may be used to shed light on the relationship between system components and can help to explain how feedback between system components serves to control a system.

In an agentic environment, we must move from "Black Box" consumption to Methodological Integrity. If an algorithm's internal decision-making logic is opaque, the system is essentially "open-loop"—it lacks the feedback mechanism required to prevent catastrophic drift or "hallucination." As noted in recent research, 73% of experts

https://www.mdpi.com/2571-8800/4/4/41#:~:text=In%20a%20typical%20system%20controlled,serves%20to%20control%20a%20system

identify Human-in-the-Loop (HITL) as the critical safeguard for these systems. The human is no longer just the "prompter"; the human is the Risk Manager acting as the controller within the feedback loop.

Understanding these feedback loops offers a pathway to understanding potential security vulnerabilities. When the feedback mechanism fails, we see "structural collapses" like the Deloitte incident with the Government of Australia. Deloitte's $290 million dollar penalty proves that adding AI to a workflow without a "Zero Trust" architecture for data provenance is building a house on sand.

https://fortune.com/2025/10/07/deloitte-ai-australia-government-report-hallucinations-technology-290000-refund/

The New Regulatory Mandate

As of 2026, "responsibility" has moved from a best practice to a legal requirement. The EU AI Act (fully applicable by August 2, 2026) and the Colorado AI Act (effective June 2026) now mandate that "high-risk" systems undergo rigorous documentation, post-market monitoring, and human oversight. (Sources: European Union AI Act; Colorado SB24-205). These regulations emphasize that organizations are liable for "unintended goal pursuit" or "unauthorized privilege escalation" by their agents.

Shift Left: Ethics as a Functional Control Variable

Organizational trust is a structural requirement that must be engineered into the technology lifecycle to prevent systemic failure. Assessed Intelligence advocates for the "Shift Left" methodology, transposing disciplined software security principles onto AI governance. By integrating the ARISE Framework (Assurance of Responsible, Innovative, and Secure Environments) at the Hypothesis Phase of AI adoption, organizations transform ethics from a theoretical concept into a functional control variable. This rigorous approach ensures that critical infrastructure objectives are satisfied before technical debt or ethical bias is institutionalized.

Operational Integrity via the ARISE™ Framework

The ARISE Framework operationalizes governance across seven integrated pillars to ensure continuous assurance:

Architecture **| ARISE Pillar: GOVERN**

Organizations establish cross-functional oversight through defined System Boundaries. This foundational requirement delineates in-scope components and trust zones, ensuring the model is technically appropriate for the deterministic task before development begins.

Lineage **| ARISE Pillar: IDENTIFY**

Every claim must be traceable to a human-verified source. This is achieved through AI System Inventory and Data Mapping, which record the origin, rights, and licensing of all datasets to establish an authoritative chain of custody.

Reproducibility **| ARISE Pillar: VALIDATE**

To ensure mission continuity, business logic must remain sound even if the AI agent is removed. This is verified through the TEVV (Testing, Evaluation, Verification, and Validation) process, specifically through reproducibility tests for AI workflows prior to production.

Operational Integrity via the ARISE™ Framework

To maintain a defensible posture, leadership must execute the following mandates anchored in ARISE.

Operationalize Ethics as a Control Loop

Ethics must function as a technical constraint via AI Operational Thresholds. By defining mathematical metrics for fairness and safety, organizations can implement automated gates that block non-compliant models.

Enforce Rigorous Data Provenance

Organizations must adopt Datasets & Tagging controls, ensuring 100% of production datasets carry verifiable provenance and sensitivity tags. This prevents automated synthesis from occurring without a manual review of raw data integrity.

Adopt Zero Trust for AI Outputs

Treat all agentic results as probabilistic hypotheses until verified against primary sources. This is supported by Identity Management, assigning unique identifiers to AI agents to track every action back to a human sponsor.

Establish an Independent Governance Layer

Cybersecurity practitioners must function as the Auditor, not just the Pilot. Through formal Internal Audit: AI Audit procedures, organizations independently assess the effectiveness of safeguards and lifecycle evidence

https://ARISEFramework.com
