SAIF 2.0: Secure Agents — Building powerful agents users can trust

Controls

The following sections describe controls available to organizations for specific AI risks.

Each control is mapped onto the corresponding risks it can address, with the exception of Governance and Assurance controls, which apply universally to all risks and every stage of the AI development process.

Data Infrastructure Model Application (Updated) Assuranc

DATA

Privacy Enhancing Technologies	**Control:** Privacy Enhancing Technologies
	Use technologies that minimize, de-identify, or restrict use of PII data in training or evaluating models.
	Who can implement: Model Creators
	Risk mapping: Sensitive Data Disclosure
Training Data Management	**Control:** Training Data Management
	Ensure that all data used to train and evaluate models is authorized for the intended purposes.
	Who can implement: Model Creators

	Risk mapping: Inferred Sensitive Data, Unauthorized Training Data
Training Data Sanitization	**Control:** Training Data Sanitization
	Detect and remove or remediate poisoned or sensitive data in training and evaluation.
	Who can implement: Model Creators
	Risk mapping: Data Poisoning, Unauthorized Training Data
User Data Management	**Control:** User Data Management

AI applications in compliance with user consent.

Who can implement: Model Creators, Model Consumers

Risk mapping: Sensitive Data Disclosure, Excessive Data Handling

INFRASTRUCTURE

Model and Data Inventory Management	**Control:** Model and Data Inventory Management
	Ensure that all data, code, models, and transformation tools used in AI applications are inventoried and tracked.
	Who can implement: Model Creators, Model Consumers (if storing models)
	Risk mapping: Data Poisoning, Model Source Tampering, Model Exfiltration
Model and Data Access Controls	**Control:** Model and Data Access Controls
	Minimize internal access to models, weights, datasets, etc. in storage and in production use.

	Creators, Model Consumers (if storing models)
	Risk mapping: Data Poisoning, Model Source Tampering, Model Exfiltration
Model and Data Integrity Management	**Control:** Model and Data Integrity Management
	Ensure that all data, models, and code used to produce AI models are verifiably integrity-protected during development and deployment.
	Who can implement: Model Creators, Model Consumers (if storing models)
	Risk mapping: Data Poisoning, Model Source Tampering

Secure-by-Default ML Tooling	**Control:** Secure-by-Default ML Tooling
	Use secure-by-default frameworks, libraries, software systems, and hardware components for AI development or deployment to protect confidentiality and integrity of AI assets and outputs
	Who can implement: Model Creators, Model Consumers (if storing models)
	Risk mapping: Data Poisoning, Model Source Tampering, Model Exfiltration, Model Deployment Tampering

MODEL

Input Validation and Sanitization	**Control:** Input Validation and Sanitization
	Block or restrict adversarial queries to AI models.
	Who can implement: Model Creators, Model Consumers
	Risk mapping: Prompt Injection
Output Validation and Sanitization	**Control:** Output Validation and Sanitization
	Block, nullify, or sanitize insecure output from AI models before passing it to applications, extensions or users.
	Who can implement: Model Creators, Model Consumers

	Risk mapping: Prompt Injection, Rogue Actions, Sensitive Data Disclosure, Inferred Sensitive Data
Adversarial Training and Testing	**Control:** Adversarial Training and Testing
	Use techniques to make AI models robust to adversarial inputs (i.e. prompts) in the context of their use in applications.
	Who can implement: Model Creators, Model Consumers
	Risk mapping: Model Evasion, Prompt Injection, Sensitive Data Disclosure, Inferred Sensitive Data, Insecure Model Output

APPLICATION

Application Access Management	**Control:** Application Access Management
	Ensure that only authorized users and endpoints can access specific resources for authorized actions.
	Who can implement: Model Consumers
	Risk mapping: Denial of ML Service, Model Reverse Engineering
User Transparency and Controls	**Control:** User Transparency and Controls
	Inform users of relevant AI risks with disclosures, and provide transparency and control experiences for use of their data in AI applications.

	Consumers
	Risk mapping: Sensitive Data Disclosure, Excessive Data Handling
Agent User Control	**Control:** Agent User Control
	Ensure user approval for any actions performed by agents/ plugins that alter user data or act on the user's behalf.
	Who can implement: Model Consumers
	Risk mapping: Sensitive Data Disclosure, Rogue Actions
Agent Permissions	**Control:** Agent Permissions

the upper bound on agentic system permissions to minimize the number of tools that an agent is permitted to interact with and the actions it is allowed to take. An agentic system's use of privileges should be contextual and dynamic, adapting to the specific user query and trusted contextual information. This design also applies to agents that have access to user information. For example, an agent asked to fill out a form or answer questions should share only contextually appropriate information and can be designed to dynamically minimize exposed data using reference monitors.

Who can implement: Model Consumers

Risk mapping: Insecure Integrated System, Sensitive Data Disclosure, Rogue Actions

Agent Observability (New)	Ensure an agent's actions, tool use, and reasoning are transparent and auditable through logging, allowing for debugging, security oversight, and user insights into agent activity.
	Who can implement: Model Consumers
	Risk mapping: Sensitive Data Disclosure, Rogue Actions

ASSURANCE

Red Teaming	**Control:** Red Teaming
	Identify security and privacy improvements through self-driven adversarial attacks on AI infrastructure and products.
	Who can implement: Model Creators, Model Consumers
	Risk mapping: All
Vulnerability Management	**Control:** Vulnerability Management
	Proactively and continually test and monitor production infrastructure and products for security and privacy regressions.
	Who can implement: Model Creators, Model Consumers

	Risk mapping: All
Threat Detection	**Control:** Threat Detection
	Detect and alert on internal or external attacks on AI assets, infrastructure, and products.
	Who can implement: Model Creators, Model Consumers
	Risk mapping: All
Incident Response Management	**Control:** Incident Response Management
	Manage response to AI security and privacy incidents.
	Who can implement: Model Creators, Model Consumers

Risk mapping: All

GOVERNANCE

User Policies and Education	**Control:** User Policies and Education
	Publish easy to understand AI security and privacy policies and education for users.
	Who can implement: Model Consumers
	Risk mapping: Will vary
Internal Policies and Education	**Control:** Internal Policies and Education
	Publish comprehensive AI security and privacy policies and education for your employees.
	Who can implement: Model Creators, Model Consumers

Product Governance	**Control:** Product Governance
	Validate that all AI models and products meet the established security and privacy requirements.
	Who can implement: Model Creators, Model Consumers
	Risk mapping: All
Risk Governance	**Control:** Risk Governance
	Inventory, measure, and monitor residual risk to AI in your organization.
	Who can implement: Model Creators, Model Consumers

Risk mapping: All