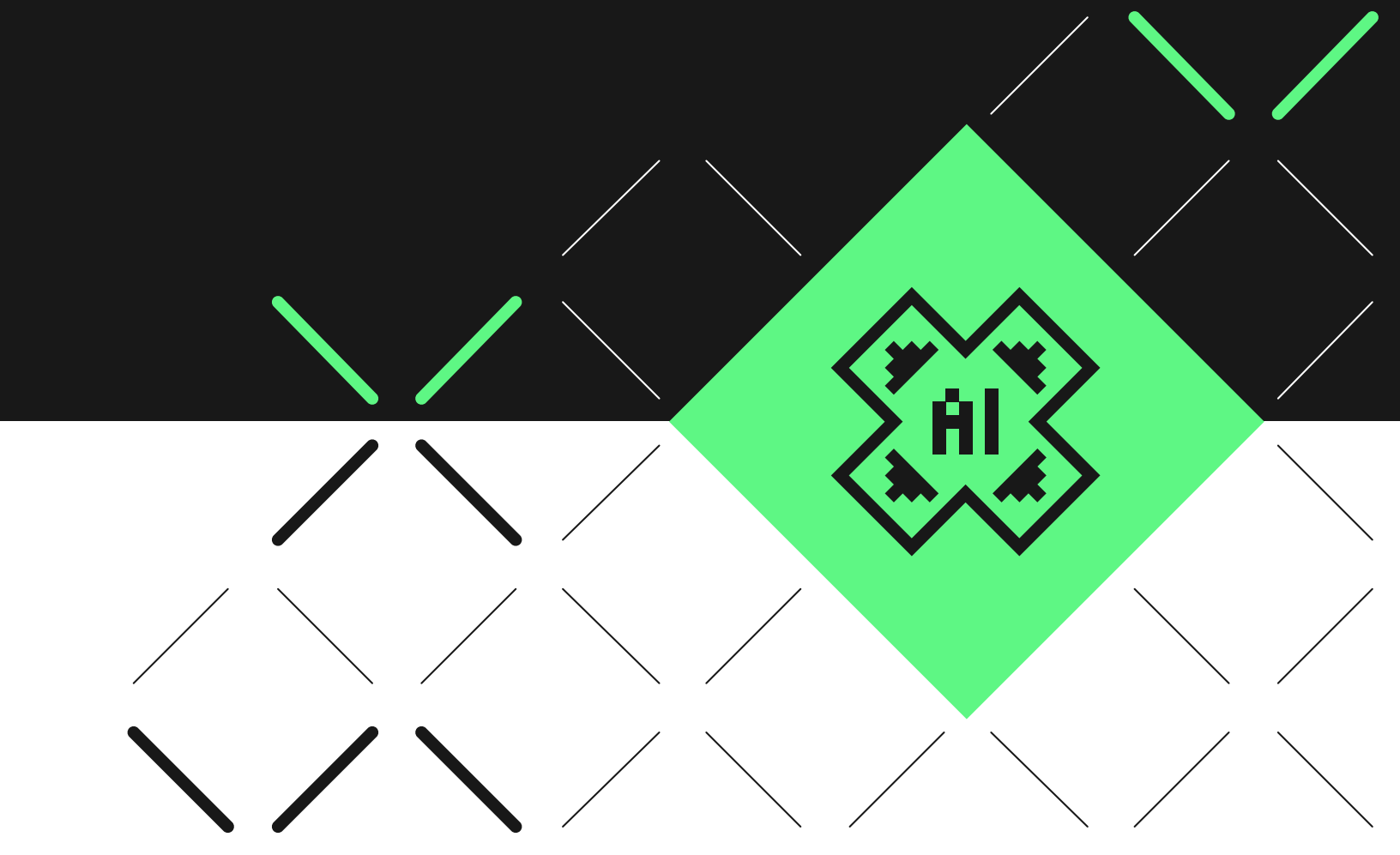
expel

WHITEPAPER

A guide to thinking about AI security risk

On the emergence of AI and the reshaping of security risk



Contents

Executive Summary 2

Overview 3

Why AI requires a new security risk framework 6

How this framework relates to NIST AI RMF and MITRE ATLAS 7

How this framework is structured 7

6 control planes explained 8

The 10 domains 10

Securing the future 21

Appendix 22



Executive Summary

AI is changing the economics of security risk more rapidly than many organizations can move. It introduces uncertainty into existing risk models and creates unknowns where operational controls previously provided confidence.

The changes AI is bringing to our environments does not allow us to pull from our old playbooks as we did with SaaS, BYOD, MDM, and similar technologies. AI alters many of our base assumptions creating uncertainty where we need risk management. This makes it difficult to manage executive and board expectations while also delegating and assessing downstream work.

This framework provides CISOs and enterprise leaders with a structured model to navigate AI security risk. It offers 10 distinct, logical domains to establish a shared organizational language that clarifies ownership, simplifies mitigation efforts, and ties security strategy to clear business impact. The framework moves AI-driven uncertainty into concrete governance, enabling leadership to coherently discuss and cohesively move toward delivering measurable security operations.

Overview

This is not a threat classification model, nor does it compete with existing frameworks that focus on explicit technical or tactical concerns. Instead, this model exists to provide the structural logic to address the ways AI is changing our base assumptions and the economics of attacker and defender operations:

- AI reduces time and costs between exposure and impact for vulnerabilities, exploitation, decisions, and attacks.
- How data and execution are handled is changing, pushing through boundaries not addressed by existing controls.
- Identity is less direct and observability is less clear (and less available).
- Intention, interpretation of intent, and actions are no longer deterministic.

Security outcomes continue to be driven by:

- How quickly risk materializes
- What risks specifically materialize
- How much control is retained across systems (implementation of governance)
- How well impact is observed and contained (observability and response).

6 control planes where control is maintained or lost

This framework is built on six control planes that are becoming a central functional component of the risk domains presented, enabling delegation across cross-functional teams. They answer “Why are you at the table?” when invited representatives come together to address the risk domains themselves.

At a glance: The 10 domains of AI risk

The core offering of this framework is the 10 risk domains. Not all are of equal weight to all organizations, and these are not presented in any priority order. Each organization will need to prioritize their risk mitigation efforts according to their own appetite.

These 10 risk domains also represent ideal reporting lines for board- and executive-level risk reporting. These are intended to be relatively complete for most organizations. They span a range of expertise and can be summarized by three critical shifts:

1. Identity is degraded in terms of boundaries, authorization, accountability, and legal nonrepudiation.

2. Execution is no longer deterministic; it is probabilistic and therefore subject to interpretation.

3. Observability is no longer a simple question of "Do you have the logs?" because logs are no longer sufficient to prove who acted, what they acted on, and what they intended.

Domain 1: Data exposure and control breakdown

Impact: Data flows and exposure within AI tool use impacts regulatory compliance and data sovereignty protections.

Domain 2: Identity, authorization, and boundary degradation

Impact: Identity-based trust is no longer a viable security boundary; authorization must shift to workflow-level verification.

Domain 3: Instruction manipulation (prompt and context injection)

Impact: AI inputs are a programmable attack surface, effectively turning internal data and data processing systems into a vector for system-wide compromise.

Domain 4: Vulnerability discovery and exploit acceleration

Impact: AI-driven exploitation narrows the patching window, forcing a transition from annual remediation to automated assessments and real-time containment.

Domain 5: Attack chain compression and orchestration

Impact: Attack automation eliminates human delays. This means security data and control latency as well as response time delays become a risk amplifier.

Domain 6: Post-compromise intelligence amplification

Impact: AI-augmented analysis allows attackers to pivot internally and understand the attacked environment with ease, shortening the time between breach and impact.

Domain 7: Human trust and business process exploitation

Impact: AI breaks human trust signals, and existing business processes are becoming programmable attack surfaces.

Domain 8: Defender operating model disruption

Impact: Existing manual workflows and processes need automation updates to be fast enough to counter AI-augmented threats, actors, and processes.

Domain 9: External AI and SaaS control loss

Impact: Shifting controls to external entities is happening at an accelerated rate, and decisions are now influenced by external systems with unknowable provenance.

Domain 10: Observability collapse and reconstruction challenge

Impact: Audit trails are now non-deterministic, leaving gaps in forensic reconstruction, accountability, and event validation.

When you apply these domains and the framework that follows, you will have the opportunity to drive meaningful and measurable risk reduction in your organization.



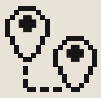
Why AI requires a new security risk framework

AI has changed several of our base assumptions. Because of these assumption changes, some of what we have in place needs review to meet the current security reality.

Some of the key assumptions that are impacted by AI adoption include:

- Humans make decisions.
- Identity maps to a (human) actor.
- Data movement and provenance are clear.
- Execution is deterministic.
- Intent and action are directly linked.
- Observability is sufficient for reconstruction.
- Time exists between attack phases.
- Vendors are infrastructure providers rather than decision-makers.
- There are clear accountable parties.

Any one of these shifts on its own would be a simple change to handle. When all are present and altered by the same introduction of technology, security operations cannot handle the change alone. These broken assumptions call for systems design and architecture changes. In some cases, entirely new protocols, technologies, or controls will need to be created to properly address the concerns.



How this framework relates to NIST AI RMF and MITRE ATLAS

Several frameworks have emerged in the past decade to address aspects of the changes AI brings. This framework provides value by complementing the existing frameworks while addressing a gap not yet covered.

NIST AI RMF is highly governance-focused and process-focused. NIST CSF 2.0 offers broader guidance and taxonomies for generalized security programs. MITRE ATLAS is highly focused on categorizing the way AI and agentic systems attack enterprises. It also covers attacks against AI tooling and systems themselves. The AI Defense Matrix addresses how and where to focus defense of AI systems themselves.

This framework complements these existing frameworks by introducing the domains where enterprise risk surfaces, allowing work delegation and risk mitigation evaluation to proceed.



How this framework is structured

Our approach first seeks to define where control planes exist for implementation of security controls. These planes broadly reflect the monitoring and enforcement locations and represent where the risk exists.

Next, we consider the various large domains where AI has changed moderate to significant portions of the security question space. These domains help us to understand what to measure and where to route work.

The domains are then discussed in a structured way. We highlight the control planes, actor context, impact, loss of control, and strategic reality of that domain within the context of the post-AI world. These get at the heart of what changed, how it changed, and what needs to be addressed.

Big picture, this framework allows a CISO to both know where to measure things and what to report up to the board. We advise using the domains as a guide for an executive summary of how you are addressing the risks AI introduces.



6 control planes explained

The control planes identify where ownership, enforcement, and measurement reside. Because each domain spans multiple control planes, addressing these risks requires cross-functional coordination and executive attention.

The control planes reflect where control exists and where it is lost. These are mapped to each of the domains in a way that helps determine which team or function needs to be involved in addressing the risk. Each organization is different, and these controls may map to different teams in different organizations.

The six control planes are:

1. Data

What is exposed, inferred, retained, classified, or recombined. This control deals directly with the flow and handling of information.

2. Identity

Who or what is authenticated and authorized to act. This control addresses both human and nonhuman identities and the concept of context versus identity.

3. Execution

What actions can be performed and how they propagate. This is directly related to the execution of programs, functions, prompts, and any other actions taken by humans, computers, or AI systems and agents.

4. Human

How trust, decisions, intent, and workflows are influenced. This control plane is outside of the computer and does not overlap with identity. It deals directly with human interpretation, weakness, awareness, perception, and function.

5. External AI and service control

Behavior, data handling, and safeguards are owned, controlled, and/or decided by third parties. This control plane addresses the shift in several controls away from the local environment and into the purview of an external AI vendor.

6. Observability (visibility/audit)

What can be seen, reconstructed, and understood about AI-driven activity. This control plane addresses our ability to see and understand what is going on based on our logging and other telemetry.

Observability increases in its importance as a primary control function. Without it, none of the other planes can be validated or enforced.



The 10 domains

These domains represent where and how the strategic risk shows up. They provide a structured taxonomy for addressing risks introduced by AI. With these, governance, strategic communications, and risk mitigation work can be managed across the organization.

Each domain contains five sections.

Control planes

A list of the control planes commonly included in this domain of risk. These may differ between organizations. What is listed here is the broad set many organizations may find relevant to address that specific domain of risk.

Actor context

A list of the types of actors, systems, or automated processes at play. These are illustrative and not exhaustive.

Impact

A brief description of some impacts commonly associated with this risk domain. These define why it matters your organization. These are illustrative and not exhaustive. Identifying specific impacts in applied environments, industries, and regions is left to the practitioner.

Loss of control

A brief description of how AI changes may result in control being lost associated with this risk domain. This is why this risk domain is in the framework. Individual assessment is necessary to surface applied gaps and aligned controls.

Strategic reality

A bigger-picture summary of our current reality given how AI has changed this risk domain.

Domain 1: Data exposure and control breakdown

This domain covers the impacts to data exposure, data governance, and existing controls. AI changes many distinct aspects of data control within an organization.

Control planes

Data (primary)
External AI/service
Observability

Actor context

Employees using AI
Embedded AI systems
Attackers extracting or inferring data

Impact

AI systems surface and recombine data across boundaries that were not designed for interpretation. Data is exposed through prompts, retrieval systems, and generated outputs.

Exposure occurs without traditional exfiltration. Retrieval systems (MCP servers, for example) surface information outside normal access patterns. This is made worse by deficient implementation of least privilege access. Outputs become a persistent channel for data leaks.

Loss of control

Data leaves controlled environments when sent to external models or services.
Retention and reuse are governed by vendor policies.
Exposure events are often not logged or reconstructible.

Strategic reality

This domain exposes a possible major strategic failure in regulatory compliance, especially when combined with other domains that functionally limit capabilities that enforce Governance, Risk, and Compliance (GRC) intent.

Domain 2: Identity, authorization, and boundary degradation

This domain focuses on how the meaning of identity and authorization is impacted by AI. Across this domain, broad degradation of boundaries is happening and adoption is moving faster than controls and governance.

Control planes

Identity (primary)
Execution
Observability

Actor context

Employees delegating to AI tools
Embedded agents acting across systems
Attackers mapping and exploiting privilege structures

Impact

AI introduces indirect and aggregated authorization.

Agents operate across systems with combined permissions
Actions are executed without clear user attribution
Privilege boundaries are traversed through workflows, not logins
Identity context degrades as actions propagate
Legal impacts emerge as disputes arise over who is responsible for an action

Loss of control

Identity is abstracted at API and agent layers
Attribution becomes ambiguous
Observability does not reliably map actions back to origin and intention

Strategic reality

Authorization is no longer tied to a single identity. This emerges because the systems identity and authorization controls are separated from the user identity at execution time within most systems. Designing specific authorization and identity policies and implementations for AI interactions may be necessary.

Domain 3: Instruction manipulation (prompt and context injection)

This domain focuses on the ability to control, audit, and change the prompts or context an AI system uses. This domain covers the proliferation of AI-driven execution within enterprise workflows.

Control planes

Execution (primary)
Data
Human
External AI/service
Observability

Actor context

Attackers injecting through content
Embedded AI executing manipulated instructions
Employees trusting outputs

Impact

Control over system behavior shifts to control over input.

Malicious instructions can be delivered through documents, web content, internal systems, emails, etc. AI executes these instructions within trusted workflows. Internal data becomes a delivery mechanism.

Loss of control

Execution logic is influenced by untrusted content. Vendor guardrails define behavior but are not transparent. Injection events are not consistently observable. Prompt guardrails are not a solved problem. User input validation processes are incomplete and untrustworthy.

Strategic reality

Execution integrity depends on the integrity of everything the system reads. In some implementations, embedded AI can read and be manipulated by hidden prompts.

Domain 4: Vulnerability discovery and exploit acceleration

AI systems have drastically changed the underlying economics of vulnerability discovery and exploit creation. This domain covers the impacts to security related to this economic change.

Control planes

Execution (primary)
Data
External AI/service

Actor context

Attackers using AI for discovery and exploitation
Employees generating code with AI, often bypassing controls, policies, and systems
Embedded systems increasing code volume

Impact

The vulnerability lifecycle is continuous and scalable.
Discovery shifts to automated processes; exploits are generated with less effort. The time between patch release and exploit creation reduces. Vulnerability volume exceeds triage capacity.

Loss of control

Discovery is no longer bounded by human effort or access to expertise. Models provide attackers with advanced capability. Defensive prioritization cannot keep up with volume.

Strategic reality

Exposure is persistent. Control depends on limiting exploit impact, not preventing discovery.

Domain 5: Attack chain compression and orchestration

This domain covers the shift in attack chains opened by AI being used to make decisions during the attack lifecycle. This also covers the shift that environmental information does not need to be exfiltrated, other than to LLMs, in order for attacks to use it for decision-making.

Control planes

Execution (primary)
Identity
Data
Observability

Actor context

Attackers orchestrating attacks with AI
Systems enabling automated workflows

Impact

Attack execution shifts to be continuous, adaptive, and automated.
More automation; attack phases run via a single process with decisions made in-line with execution. Actions are chained without human actor delay, and attacks adapt dynamically to environment feedback.

Loss of control

Execution visibility becomes deficient because it lacks deterministic outcomes. Response timing becomes insufficient, and containment (even when automated) becomes too slow.

Strategic reality

Security controls, IR processes, and detection and response designed for staged attacks degrade under continuous execution. Capacity is exhausted as attack rates increase.

Domain 6: Post-compromise intelligence amplification

This domain considers the post-compromise rate at which attackers can come to a comprehensive environmental understanding.

Control planes

Data (primary)
Execution
Observability

Actor context

Attackers using AI after initial access

Impact

AI increases attacker effectiveness inside the environment.

Rapid analysis of systems and data with prioritized targeting of high-value assets (sometimes enabled by internal models).
Lateral movement decisions made in real time by AI, with faster and more pervasive persistence strategies.

Loss of control

Internal complexity becomes navigable by actors with automation. Easier hiding of attacker activity because it blends with legitimate use. Observability does not capture intent or decision-making.

Strategic reality

Impact is proportional to how quickly an attacker can understand and act after entry versus the time defenders contain and remediate. Mastery of infrastructure knowledge can lead to directed attacks, minimizing detection opportunities.

Domain 7: Human trust and business process exploitation

This domain covers human trust exploitation and the way humans process, detect, and interact. Deepfakes and improvements in deceptive messaging that simulates trusted connections of the target are within this domain.

Control planes

Human (primary)
Identity
Execution
External AI/service

Actor context

Attackers generating synthetic interactions
Employees' over-reliance on AI outputs

Impact

Trust signals degrade across all human interactions.

High-quality impersonation at scale and synthetic artifacts integrated into workflows. Business processes become entry points (approvals, financial actions, hiring). Decision-making is influenced by generated content (sometimes without known provenance).

Loss of control

Trust, accuracy, and correctness cannot be validated through traditional signals. External AI systems generate indistinguishable artifacts. Human verification becomes unreliable.

Strategic reality

Business workflows become programmable attack surfaces.

Domain 8: Defender operating model disruption

Defense operates within the confines of law, governance, policy, and budgetary constraints. This domain focuses on the gaps between what exists today and what is needed to defend against AI-augmented attackers.

Control planes

Execution (primary)
Human
Observability

Actor context

Defenders operating with or without AI

Impact

Security operations are structurally disadvantaged and cannot match threat velocity.

Detection and response lag behind attack speed. Alert and vulnerability volume increase simultaneously. Triage becomes the limiting factor. Workforce capacity constrains outcomes. Policy and governance become a threat to speed of defender response.

Loss of control

Operational capacity becomes the constraint, observability gaps limit effective response, SOC analysts forced to make decisions with even less context, and tooling does not scale with threat activity.

Strategic reality

Security effectiveness is determined by operational velocity and scalability. This means a need to increase automatic actions through conventional methods or via AI-enabled workflows.

Domain 9: External AI and SaaS control loss

This domain focuses on the fast adoption of AI, particularly external LLM models, and the related loss of control. It mirrors some of the same problems that shifts to SaaS experienced, but at an accelerated rate.

Control planes

External AI/service (primary)
Data
Execution
Observability

Actor context

Employees using external AI services (approved or shadow AI)
Organizations embedding third-party AI
Attackers leveraging the same services

Impact

Movement of critical parts of the security model to external ownership has accelerated.

Data is sent to systems with independent governance, and model behavior can change without visibility. Safeguards are implemented and enforced by vendors. Plugins and integrations extend trust boundaries.

Loss of control

Data retention and usage are externally defined. Execution logic depends on vendor systems.
Observability into model behavior is limited. Failure circumstances are less understood.
Security posture is partially inherited, not controlled.
Users are both relying on output from and pushing data into unapproved models/services.

Strategic reality

Your security boundary includes a growing set of systems you do not control. Your risk posture depends on even more decisions you do not make, may not know about, and/or cannot influence.

Domain 10: Observability collapse and reconstruction challenge

This domain focuses on the changes to observability. Log access and interpretation fall within this domain.

Control planes

Observability (primary across all others)

Actor context

All (attackers, employees, embedded systems, defenders)

Impact

AI-driven systems reduce visibility into cause and effect.

Actions are not deterministic. Prompts, context, and decisions are (very often) not fully logged, reproducible, or fully auditable. Attribution across systems becomes less clear.

Traditional logs do not capture intent, reasoning, or decision paths.

Loss of control

With gaps in observability comes the inability to reconstruct events, explain outcomes, detect subtle manipulation, and validate control effectiveness.

Strategic reality

If activity cannot be reconstructed, it cannot be secured. Observability becomes the foundation for all other controls.



Securing the future

Security control has been distributed, indirect, and partially external for a long time. AI doesn't change this, but it makes these problems harder to manage.

As AI is implemented, data crosses more boundaries. Identity abstracts across systems and is no longer strongly tied to access controls.

Execution remains continuous and adaptive. Trust becomes less explicitly defined. Observability is incomplete, and interpretation of prompts becomes indeterministic.

Realized risk is anchored to how quickly attacks progress to impact. This is influenced by the amount of control maintained and observability available across systems. Mitigation is impacted by how effectively activity can be monitored, understood, and remediated when needed.

Mitigating and remediating risk is cybersecurity's raison d'être. What we do next is fulfill our core mission. Take this framework as a guide and commission teams to assess and address the risk your organization is facing.



Appendix

Domain 1 walk-through

To present an example of how to work with these domains within cross-functional working groups, consider the following for Domain 1: Data exposure and control breakdown.

Identifying stakeholders

This work starts by identifying the stakeholders for the risk exposure within Domain 1. This domain focuses on data exposure, data governance, and existing controls and will need the owners or representatives for each of those functions.

Key cross-functional stakeholders and owners include:

- **Chief Information Security Officer (CISO)/Security Architecture:** Owns data flow security, model boundaries, and prevention controls.
- **Data governance and privacy officers:** Owns compliance, regulatory requirements, data sovereignty, and privacy consent constraints.
- **Legal and compliance lead:** Owns contractual terms, vendor data retention policies, and potential liabilities.
- **AI/Engineering Product Owners:** Owns integration pipelines, retrieval context (RAG/MCP), and system prompt controls.

Defining key dependencies

When the team meets, they may find it necessary to define what data governance policies depend on in their organization. Identifying these dependencies and defining them may produce the following as a key output of the team.

Data governance (in implementation) depends on:

- **Interpretation and presentation:** Dictating how the system translates and displays data.
- **Processing location:** Identifying where data manipulation and interpretation occur.
- **Process visibility:** Ensuring transparency throughout the entire operational pipeline.
- **Data provenance and lineage:** Tracking origin, ownership, and modifications of training data.

- **Quality and bias control:** Monitoring for representative datasets, historical bias, and data poisoning.
- **Privacy and consent management:** Ensuring compliance with regulations (like GDPR) and managing opt-out mechanisms.
- **Security and access control:** Defining who can access, modify, or retrain models with specific datasets.
- **Retention and minimization:** Establishing rules for deleting stale data and limiting unnecessary collection.
- **Feedback auditing:** Governing how user interactions and model outputs are reused for retraining.

Creating key metrics and success criteria

Once the team establishes these definitions, they decide to forego prescriptive definitions to allow teams of SMEs to handle what SMEs are best positioned to handle. They decide to set core targets by defining operational metrics and success criteria. These create clarity from the abstract for SMEs to understand what they are being asked to address. This work may take place via subcommittees.

Operational metrics and success criteria:

- **Unsanctioned data exposure rate:** Percentage of governance-scoped data assets accessible to external or unauthorized AI endpoints without effective controls.
- **Tracking exposure:** A count of the incidents resulting in data exposure over time to assess trends and evaluate efficacy of control changes.
- **Governance and compliance posture index:** Aggregate tracking enterprise-wide alignment with data sovereignty, regulatory mandates, and internal consent policies across all AI integrations.
- **Third-party risk mitigation coverage:** Tracking of vendor AI workflows, data retention policies, and non-training contractual protections. Process defined for regular review of the risk register.

Specific mitigation directives

Given the understanding of risk appetite and concern, the team decides to prescribe some specific directives that must be met by the SMEs. The team determines these must be coupled with a specific reporting cadence.

Mitigation checkpoints:

- **Pre-ingestion filtering:** Strip sensitive attributes and PII before text/embeddings reach context stores or retrieval pipelines. GRC requires monthly update cadence.

- **Context-aware and mandatory access controls:** Enforce fine-grained permissions on AI search and retrieval layers, preventing over-privileged data surface. CISO requires monthly update cadence.
- **Output validation and sanitization:** Scan generated responses in real time to detect and block unintentional leaks of confidential information. GRC and Legal request quarterly updates.

Wrapping it up

With these outputs, the work of making it happen can begin. Work can be sized, scoped, and commissioned.

These examples illustrate some possible outputs from the teams put together to address domain 1. These are meant to be illustrative but not complete nor comprehensive. When working through these exercises, all best practices around scheduling, resourcing, managing, and regular review should be followed.

Cross-domain interaction

When developing these domains, it became clear that several interact in a way that amplifies the possible impact. To illustrate this effect, we have picked several interactions to highlight.

- **External AI and service control amplifies data exposure:** Many widely used LLMs implement data storage policies that may not align to corporate policies. Depending on configuration, the LLMs can train on user data. This means that your corporate data is now stored by the LLM company. Your data may surface in other people's answers, depending on how the model is trained and how it functions.
- **Observability gaps obscure instruction manipulation:** When you have observability gaps to AI surfaces, instruction manipulation can occur unseen and impact your user's results.
- **Attack chain compression and orchestration reduces response opportunity (Defender operating model disruption):** Any changes to the speed of the attack chain or the underground economy efficiency alter the response opportunity. As changes happen, defense teams have to respond. Our current operating model is strained by rapid changes to attacker dynamics.
- **External AI and service control loss reduces observability:** The use of third-party tools reduces the effective observability. This gets worse when combined with the use of shadow AI.

Domain and control plane grid

Risk domain	Primary control plane	Secondary control planes
Domain 1: Data exposure and control breakdown	Data	External AI/service, observability
Domain 2: Identity, authorization, and boundary degradation	Identity	Execution, observability
Domain 3: Instruction manipulation (prompt and context injection)	Execution	Data, human, external AI/service, observability
Domain 4: Vulnerability discovery and exploit acceleration	Execution	Data, external AI/service
Domain 5: Attack chain compression and orchestration	Execution	Identity, data, observability
Domain 6: Post-compromise intelligence amplification	Data	Execution, observability
Domain 7: Human trust and business process exploitation	Human	Identity, execution, external AI/service
Domain 8: Defender operating model disruption	Execution	Human, observability
Domain 9: External AI and SaaS control loss	External AI/ service	Data, execution, observability
Domain 10: Observability collapse and reconstruction challenge	Observability is primary across all control planes.	