EC-Council
Building A Culture Of Security

AI SECURITY AND GOVERNANCE FRAMEWORK

An EC-Council Global Services White Paper

Adopt. Defend. Govern.

The board-to-engineering operating model for safe, responsible, and scalable enterprise AI.

PRACTITIONER-TESTED · BOARD-TO-ENGINEERING · STANDARDS-ALIGNED

Index

ADG is intentionally dense. Readers are encouraged to select the role that best matches their responsibilities and follow the corresponding path. Each path covers approximately four to five sections out of the total 13, so it is not necessary to read the entire document.

Why ADG?

Enterprise AI is moving into production faster than governance can absorb. Engineering teams are deploying agents that act on tools, consume retrieved context, run multi-step plans, and connect to other agents, often before the individuals accountable for risk have reviewed or approved the system.

Boards are asking the questions that control departments cannot answer. Regulators are acting, with the European Union Artificial Intelligence Act (EU AI Act) leading the way and other frameworks likely to follow. Standards and frameworks in which organizations have already invested, such as ISO/IEC 42001 and the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF), clarify what organizations must consider. However, they do not specify who decides, who builds, who breaks, who signs, or what evidence should be available to a regulator on Monday morning.

C|AIPM Certified Artificial Intelligence Program Manager

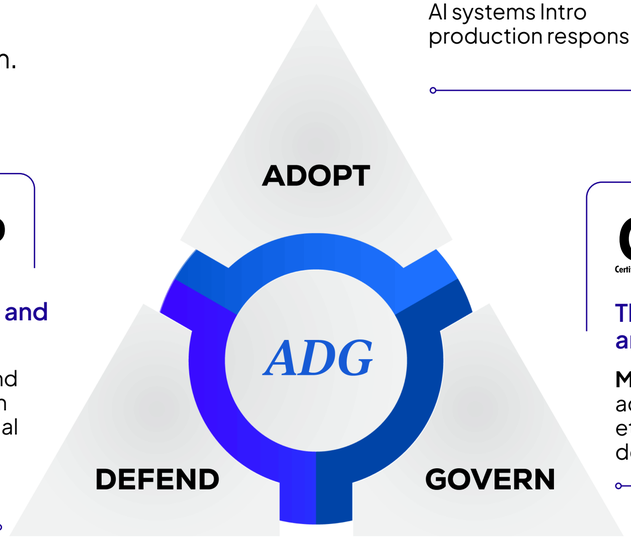
The AI Execution Role

Mandate: Deliver AI systems Intro production responsibly

C|OASP Certified Offensive AI Security Professional

The AI Security and Adversal Role

Mandate: Test and protect AI system against adversarial and misuse scenarios.



C|RAGE Certified Responsible AI Governance & Ethics

The AI Accountability and Oversight Role

Mandate: Ensure accountability, ethics and regulatory defensibility

The architecture consists of three pillars — **ADOPT** (execute and deliver), **DEFEND** (secure and validate), and **GOVERN** (oversee, assure, decide) — that scale consistently from the board to the engineer. The same three terms can frame a board paper, organize a release gate, label a red-team report, and structure an incident playbook.

Beneath those pillars are **nine governance surfaces,** which represent the points where engineers instrument the systems and attackers act; **12 Minimum Controls** (MC-1 through MC-12), each with a named evidence artifact that an auditor can review; **nine deployment overlays** for the patterns in which AI is currently deployed—agentic orchestration, tools and the Model Context Protocol (MCP), multi-agent interoperability, multimodal and composite stacks, and long-context architectures; **three autonomy tiers** (human in the loop [HITL], human on the loop [HOTL], and human out of the loop [HOOTL] tied directly to the controls; and a **four-phase roadmap** that takes an organization from inventory to continuous assurance over 24 months.

CORE THESIS:

That is the gap that ADG closes. ADG is not another standard; it is the **operating model that sits beneath existing standards** and makes them executable. It was written by a practitioner advisory board spanning financial services, healthcare, manufacturing, telecommunications, energy, and technology across regulated and less-regulated jurisdictions in North America, Europe, and Asia-Pacific. Credentials and curricula are downstream of the framework; they are not the reason for it.

BUILT FOR ALIGNMENT

ADG is built for alignment, not competition. Each Minimum Control maps cleanly into NIST AI RMF functions (GOVERN, MAP, MEASURE, MANAGE) and ISO/IEC 42001 Annex A; the framework's §11.2 crosswalk makes this mapping explicit. MC-1 produces the AI inventory ISO/IEC 42001 A.6 and relevant to EU AI Act Article 49 registration. MC-9 produces the post-market log that Article 73 inspectors may request. MC-11 produces the bias evidence Article 10 requires. Organizations pursuing ISO/IEC 42001 certification or EU AI Act conformity use ADG to **operationalize what those frameworks describe abstractly.** The work is the same; ADG removes the translation layer.

ADG also addresses areas that the standards generally do not. Agentic systems, MCP-connected tools, multi-agent interoperability, long-context architectures, and composite multimodal stacks each include an additive deployment overlay that composes with the same 12 Minimum Controls. It is possible for organizations to move from one copilot to a fleet of agents without rewriting a single control.

One element of ADG sits where peer frameworks generally do not: the **AI Governance Council.** Every framework tells organizations to govern AI. ADG identifies who resolves the conflict when delivery and safety disagree. The Council resolves ADOPT–DEFEND tension, sets go/no-go thresholds, owns the exception register, and ties decision rights to MC-3 (Separation of Duties) and MC-10 (Periodic Governance Review). The mechanism, rather than the abstract requirement, is what peer frameworks rarely name, and it is where most production AI risk actually resides.

ADG is built for boards, chief information security officers (CISOs) and AI red teams, platform and AI engineers, risk and compliance teams, and procurement teams. Each audience follows a different path, but the artifacts are the same. The framework was developed with senior AI, security, and governance leaders who run production AI inside Fortune 500, Fortune Global 500, and Big Four firms, including Salesforce, Microsoft, Citi, JPMorgan Chase, NTT DATA, KPMG, ServiceNow, BNP Paribas, Prudential, GE Healthcare, BASF, and Jio. These are the practitioners who own the failure modes this framework governs.

ADG IN ONE SENTENCE

ADG takes the **ADOPT. DEFEND. GOVERN.** operating model down to 12 auditable controls, nine governance surfaces, and additive overlays for agentic and multi-agent deployments. These elements are composable across homegrown AI, foundation model API–based AI, and software as a service (SaaS)–embedded AI.

Who is Reading This?

BOARD AND EXECUTIVE

"Can we sign off on this AI system and what evidence will the auditor require?"

Your path: § Executive Summary · § Triad · § Harm Taxonomy · § Controls · § Regulatory.

SECURITY AND AI RED TEAMS

"Where are the attack surfaces, and which controls activate under which circumstances?"

Your path: § Surfaces · § Harm Taxonomy · § Life Cycle · § Overlays · § Controls

ENGINEERING AND PLATFORM

"Which controls integrate with CI/CD pipelines, MCP servers, and prompt repositories?"

Your path: § Triad · § Surfaces · § People/Process/Tech · § Life Cycle · § Overlays

RISK AND COMPLIANCE

"How does this align with ISO/IEC 42001, NIST AI RMF, and the EU AI Act?"

Your path: § Harm Taxonomy · § Shared Responsibility · § Regulatory · Appendix A

PART I — FOUNDATIONS

Background, Origin, and the RE³ Trust Model

Origin, scope, and the four properties that make AI trustworthy:
Responsible, Ethical, Efficient, Explainable

1 Background

1.1 Origin and Positioning

ADG began with a mission at EC-Council: to address fragmentation in the AI governance space across point standards, vendor playbooks, and academic frameworks. None of these provided enterprise boards and engineering teams with a single, practitioner-tested operating model for deploying AI safely at scale.

EC-Council set out to build that missing link: a framework built around three enduring functions—**ADOPT. DEFEND. GOVERN**—that scales from the board to the engineer.

To avoid building in isolation, EC-Council convened an Advisory Board of practitioners from mature enterprises that have deployed AI in production. The initial draft was circulated to senior AI, security, and governance leaders across global organizations in financial services, technology, manufacturing, healthcare, telecommunications, energy, and consulting—covering both regulated and less-regulated sectors, and spanning North America, Europe, and Asia-Pacific.

ADVISORY BOARD IMPACT

ADG was not drafted in a conference room. It was forged with the senior AI, security, and governance leaders who run production AI inside Fortune 500, Fortune Global 500, and Big Four firms across regulated and less-regulated sectors, from design through implementation.

Their input drove eight structural enhancement clusters: an expanded harm taxonomy with Responsible AI integration, coverage beyond large language model (LLMs), including diffusion, multimodal, composite systems, a shared responsibility model for vendor/SaaS AI, deeper strategic treatment of the GOVERN pillar, a measurable metrics and evidence framework, multi-agent interoperability governance, explicit mapping to regulations and standards (EU AI Act, NIST AI RMF, ISO/IEC 42001), and post-deployment continuous governance.

FRAMEWORK LEADERSHIP EC-COUNCIL GLOBAL SERVICES

Jay Bavisi
Chairman and CEO, EC-Council Group.

Karthik S.
Framework Architect and Lead Author, Practice Head, SecureAI, EC-Council Global Services.

Mayank Tandon
Global Outreach and Partner Experience, EC-Council.

For the complete Advisory Board roster, including the practitioners whose comments, corrections, and counterpoints made this framework field-tested, see the dedicated **Acknowledgments** section at the end of this white paper.

Thank you to the Advisory Board. The practitioners on the Advisory Board contributed their time, scrutiny, and hard-won experience to review ADG. Their comments, corrections, and counterpoints are the reason this framework is field-tested rather than merely aspirational.

This version is the result: the iterative output of that collaboration. ADG is not a static document; it is a living framework designed to evolve as AI deployment patterns change and as new practitioners contribute to its development.

The scope has expanded beyond its original LLM-centric framing to cover modern deployment patterns, including large language model operations (LLMOps), retrieval-augments generation, tool-using agents, MCP-connected ecosystems, long-context architectures, multi-agent systems, and model life cycle controls spanning pre-training through runtime.

FRAMEWORK OBJECTIVE

ADG provides a board-to-engineering operating model that separates decision rights, assigns control ownership, and defines the minimum governance required for safe, responsible, and scalable AI deployment. It continues to evolve through ongoing practitioner review.

1.2 Framework Purpose: The RE³ Trust Model

ADG is purpose-built to operationalize four non-negotiable properties of trustworthy enterprise AI captured in the **RE³ Trust Model**: Responsible, Ethical, Efficient and Explainable AI.

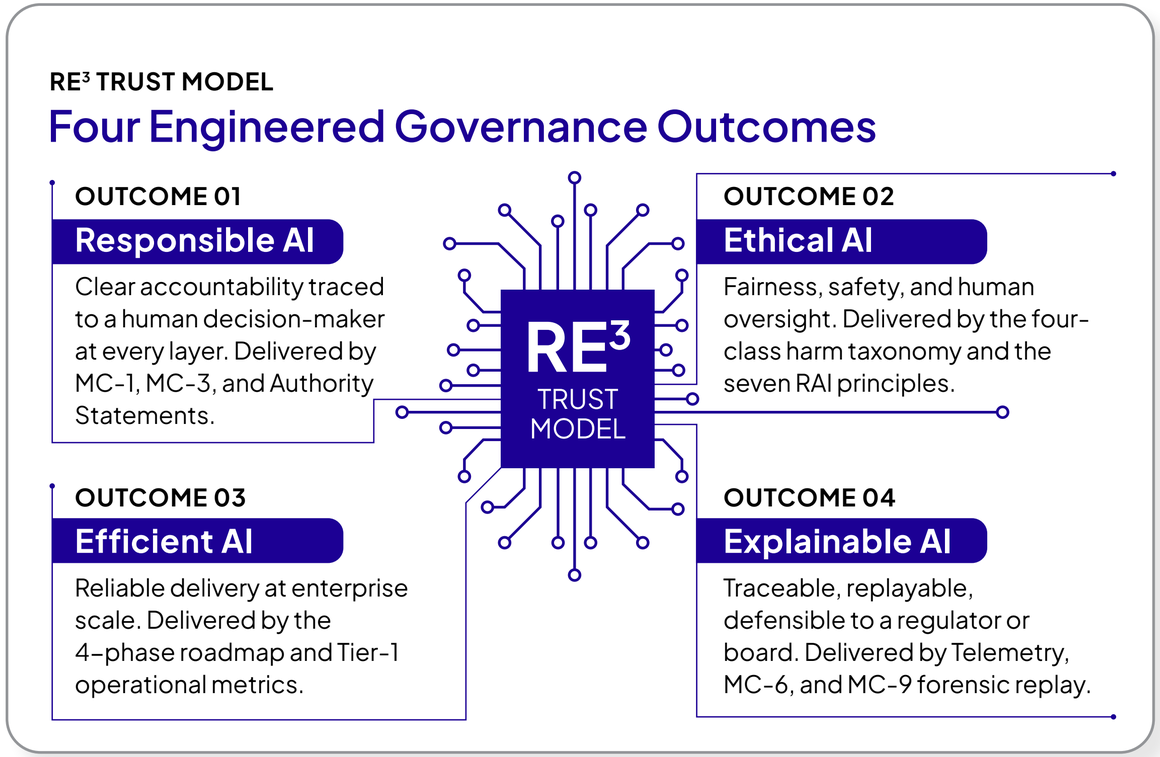


Figure 1.1 — The RE³ Trust Model: four engineered governance outcomes

Governance outcomes must be engineered into every AI deployment, from an internal copilot through a production multi-agent system. ADG translates each outcome into concrete controls across all nine governance surfaces (Model, Prompt, Context, Tools, Orchestration, Identity, Safety Layer, Telemetry, Learning Loop) and applies them to every deployment pattern across all three deployment classes (Homegrown, Foundation Model API, and SaaS AI).

RESPONSIBLE AI

Clear Accountability, Every Layer

Every AI action traces back to a human decision- maker. Delivered by: MC-1 (AI System Inventory), MC-3 (Separation of Duties), Agent Authority Statements, Shared Responsibility Model, and the AI Governance Council escalation path.

ETHICAL AI

Fairness, Safety, Human Oversight

Protection from bias, manipulation, and harm across four classes (Technical, Societal, Operational, Systemic). Delivered by: Four-Class Harm Taxonomy, Seven RAI Principles, MC-11, the Safety Layer surface, and Autonomy Tier controls (HITL/HOTL/HOOTL).

EFFICIENT AI

Reliable Delivery at Enterprise Scale

Governance that enables velocity, not one that grinds it. Delivered by: 4–Phase Implementation Roadmap, Tier-1 Operational Metrics, the LLMOps Overlay, reusable Operating Artifacts, and one unified control set spanning Homegrown, API, and SaaS deployments.

EXPLAINABLE AI

Traceable, Replayable, Defensible

Every output can be justified to a regulator, board, or customer. Delivered by: Telemetry surface, MC-6 (Context Policy with provenance), MC-7 (Tool/MCP audit logging), MC-8 (Runtime Monitoring), MC-9 (Incident Response with forensic replay).

The outcome: **RE³ AI becomes measurable.** Each trust property aligns with specific controls, metrics, and evidence requirements, enabling organizations to demonstrate trustworthy AI to boards, regulators, and customers—rather than merely claim it.

1.3 Core Thesis

Traditional cyber and software governance frameworks assume deterministic execution, finite input spaces, stable functionality, and visible logic. AI systems challenge those assumptions. Organizations need a framework that governs:

- Systems that can generate **novel behavior**
- Systems that act through **external tools and APIs**
- Systems that consume **dynamic context** from internal and external sources
- Systems whose behavior can change through **model updates, tuning, retrieval changes, and runtime interaction**
- Systems composed of **multiple models and agents** that interact with each other, creating emergent risks not present in any individual component

ADG addresses this by combining role separation, surface-based control ownership, deployment-pattern overlays, life cycle governance, and a structured harm taxonomy that extends beyond security to encompass responsible AI.

1.4 Scope and System Coverage

ADG governs all AI systems as defined by the OECD AI Policy Observatory and the EU AI Act: machine-based systems that, for explicit or implicit objectives, infer from input how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. This explicitly includes:

- LLMs and transformer-based generative AI (Gen AI)
- Diffusion models, diffusion-transformer hybrids, and image/video generation systems
- Multimodal systems combining text, image, audio, and code generation
- Classical machine learning models (supervised, unsupervised, reinforcement learning)
- Composite AI systems that orchestrate multiple model types within a single workflow
- Emerging architectures, including world models, neuromorphic systems, and foundation models beyond text

1.5 Distinction: Standards vs. Regulations

ADG is a voluntary governance framework designed to be compatible with, but distinct from, regulatory requirements. Throughout this document:

REGULATORY REQUIREMENTS

Legally binding obligations such as the EU AI Act, sector-specific regulations, and data protection laws (GDPR)

STANDARDS

Voluntary frameworks and best practices: NIST AI RMF, ISO/IEC 42001, OWASP Top 10 for Large Language Model Applications and OWASP Top 10 for Agentic Applications

ADG controls are designed to support alignment with regulatory and standards-based requirements, but organizations must independently verify compliance with applicable legal obligations in their jurisdictions.

2 Executive Summary

ADG is an enterprise AI security and responsible AI governance framework structured around three pillars—**ADOPT. DEFEND. GOVERN.**—that provide a board-to-engineering operating model for safe, responsible, and scalable AI deployment.

3	9	12	9	4	24-MONTH
PILLARS	SURFACES	MINIMUM CONTROLS	OVERLAYS	HARM CLASSES	ROADMAP

Key framework components

THE ADG TRIAD (PILLARS)

ADOPT executes and delivers AI capabilities. **DEFEND** secures and validates against all harm classes. **GOVERN** oversees, assures, and resolves tensions between the other two pillars.

GOVERNANCE SURFACES —WHAT MUST BE GOVERNED

Model, Prompt, Context, Tools, Orchestration, Identity, Safety Layer, Telemetry, and Learning Loop — defining what must be governed regardless of architecture.

DEPLOYMENT OVERLAYS—PATTERN-SPECIFIC CONTROLS

LLMOps · Agentic Orchestration · Agent · Agentic Hardening · Tools and MCP · Context and Long-Window · Pre/Post-training · Multi-Agent Interoperability · Multimodal/Composite systems.

FOUR-CLASS HARM TAXONOMY

Technical, Societal, Operational, and Systemic harms — ensuring governance extends beyond security to cover bias, fairness, reliability, and emergent multi-agent risks.

MINIMUM CONTROLS (MC-1 THROUGH MC-12)

Each control includes an evidence requirement to ensure measurability. Together they form the baseline every production AI system must meet.

AUTONOMY TIERS—HITL/HOTL/HOOTL

Assistive HITL, Conditional HOTL, and Autonomous HOOTL—determining the minimum governance required per system.

RESPONSIBLE AI PRINCIPLES

Fairness and non-discrimination, Transparency and explainability, Privacy and data protection, Accountability, Human oversight, Robustness and safety, and Sustainability and societal well-being — embedded across all three pillars.

FOUNDATIONAL PRINCIPLES —THE DURABLE STANDARDS

Separation of powers · Explicit authority · Context as attack surface · Tool use as highest-risk control plane · Graduated oversight · Life Cycle governance · Mandatory evidence · Embedded RAI · Defined shared responsibility.

IMPLEMENTATION PHASES—ROADMAP TO MATURITY

Foundation (0–3 months) · Control Deployment (3–9 months) · Agentic Readiness (9–15 months) · Maturity (15–24 months): a staged path from inventory to continuous assurance.

This version was forged with senior AI, security, and governance leaders who run production AI inside Fortune 500, Fortune Global 500, and Big Four firms—practitioners who work from design through implementation across regulated and less-regulated sectors.

It addresses eight major gap clusters identified through systematic review and extends the framework to cover multi-agent systems, multimodal architectures, shared responsibility, responsible AI, and measurable governance.

ADG can serve as the basis for consulting assessments, enterprise AI governance programs, certification architecture, board-facing assurance discussions, regulatory preparation, and vendor due diligence.

Framework Architecture — How the Layers Connect

FRAMEWORK ARCHITECTURE

How ADG's Six Layers Compose

9 GOVERNANCE SURFACES
Model, Prompt, Context, Tools, Orchestration, Identity, Safety Layer, Telemetry Learning Loop

12 MINIMUM CONTROLS: MC-1 - MC-12
Inventory, Risk, SoD, Evaluation, Change, Context, Tools, Monitor, Incident, Review, Fairness, Shared Responsibility

9 ADDITIVE DEPLOYMENT OVERLAYS
LLMOps, Agentic Orchestration, Agent, Hardening, Tools/MCP, Context, Pre/Post-train, Multi-Agent, Multimodal

6-STAGE LIFE CYCLE
Pre-train and Post-train, Build, Deploy, Run, Retire

4-PHASE ROADMAP: 0-24 MONTHS
Foundation (0–3), Control Deployment (3–9), Agentic Readiness (9–15), Maturity (15–24)

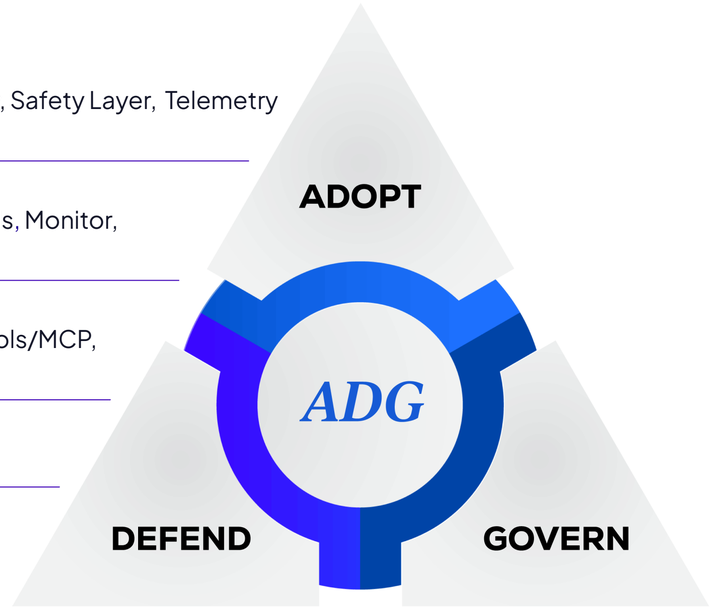


Figure 2.1 — Framework Architecture: How ADG's Six Layers Compose

3 Document Usage

This section provides guidance on how to read, apply, and implement the ADG framework.

3.1 Target Audience

Each audience has a tailored reading path.

AUDIENCE	RECOMMENDED READING PATH
Board and Executive Sponsors	§ Executive Summary § ADG Triad and Operating Model § Harm Taxonomy and RAI § Controls and Board-Level Indicators § Regulatory and Roadmap
Security and AI Red Teams	§ Governance Surfaces § Harm Taxonomy and RAI § Life Cycle Governance § Deployment Overlays 11) Controls and Measurement
Engineering and Platform Teams	§ ADG Triad and Operating Model 5) Governance Surfaces § People, Process, Technology and Data § Life Cycle Governance § Deployment Overlays
Compliance and Legal	§ Harm Taxonomy and RAI § Shared Responsibility § Regulatory and Roadmap; Appendix A: Definitions
All Stakeholders	§ Background § Executive Summary § Document Usage

3.2 Foundational Principles

ADG is built on nine foundational principles. Each principle is a durable standard that applies across industries, deployment patterns, and regulatory regimes.

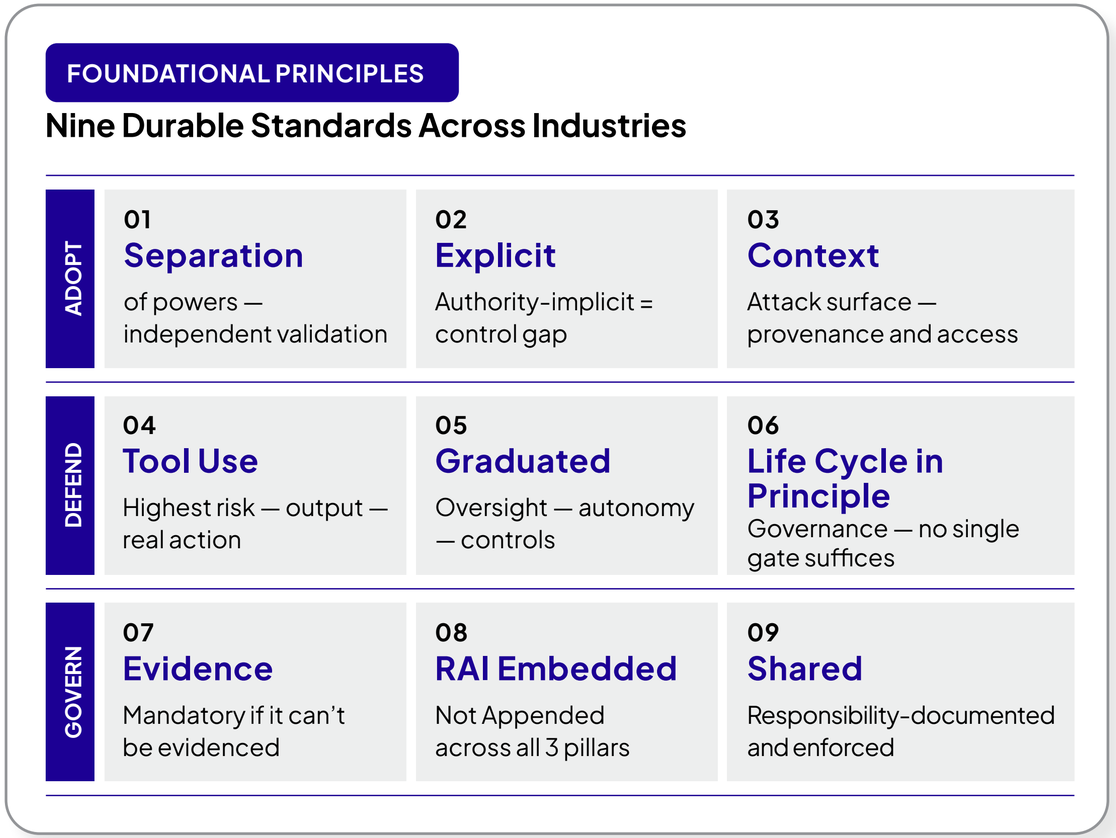


Figure 3.1 — Nine Foundational Principles: durable standards across industries

3.3 Autonomy Tiers

ADG distinguishes three operational autonomy tiers. The tier determines the minimum governance requirements for deployment.

Note on terminology: HITL/HOTL/HOOTL originated in DoD doctrine (DoD 5000.59–M, 1998) and remain widely used. ADG uses the tighter operational definitions below to make each tier directly testable against the Minimum Control Set.

TIER	DESCRIPTION	OVERSIGHT	MINIMUM GOVERNANCE
Assistive	Recommends or drafts; does not act	HITL	Output review, bounded context, no direct tool execution
Conditional	Acts within pre-approved limits and constrained tools	HOTL	Authority statement, guardrails, audit logs, kill switch, exception thresholds, bias monitoring
Autonomous	Executes multi-step goals with limited or delayed human review	HOOTL	Formal approval, strong telemetry, circuit breakers, forensic replay, periodic governance review, mandatory fairness evaluation

3.4 Implementation Roadmap Overview

A four-phase staged path that takes an organization from inventory to continuous assurance in 24 months.

PHASE 1

Foundation

Months 0–3

- AI system inventory
- Assign accountable owners
- Define ADG roles
- Classify by criticality and autonomy
- Establish AI Governance Council
- Baseline gap analysis

PHASE 2

Control Deployment

Months 3–9

- Release gates
- Prompt and tool change control
- Context policies
- Logging and monitoring
- Adversarial testing
- Initial fairness evaluations
- Vendor AI due diligence

PHASE 3

Agentic Readiness

Months 9–15

- Authority statements
- Tool trust tiers
- MCP governance
- Circuit breakers
- Multi-agent governance
- Forensic replay

PHASE 4

Maturity

Months 15–24

- Continuous evaluation
- Persistent adversarial monitoring
- Drift governance
- RAI measurement
- Board-level reporting
- Formal assurance review

PART II — ARCHITECTURE

The ADG Triad and Operating Model

Three pillars that scale unchanged from the board to the engineer. The same three words frame a board paper, organize a release gate, label a red-team report, and structure an incident playbook.

4 The ADG Triad and Operating Model

This section provides guidance on how to read, apply, and implement the ADG framework.

THE ADG TRIAD

Three Pillars That Scale Unchanged from Board to Engineer

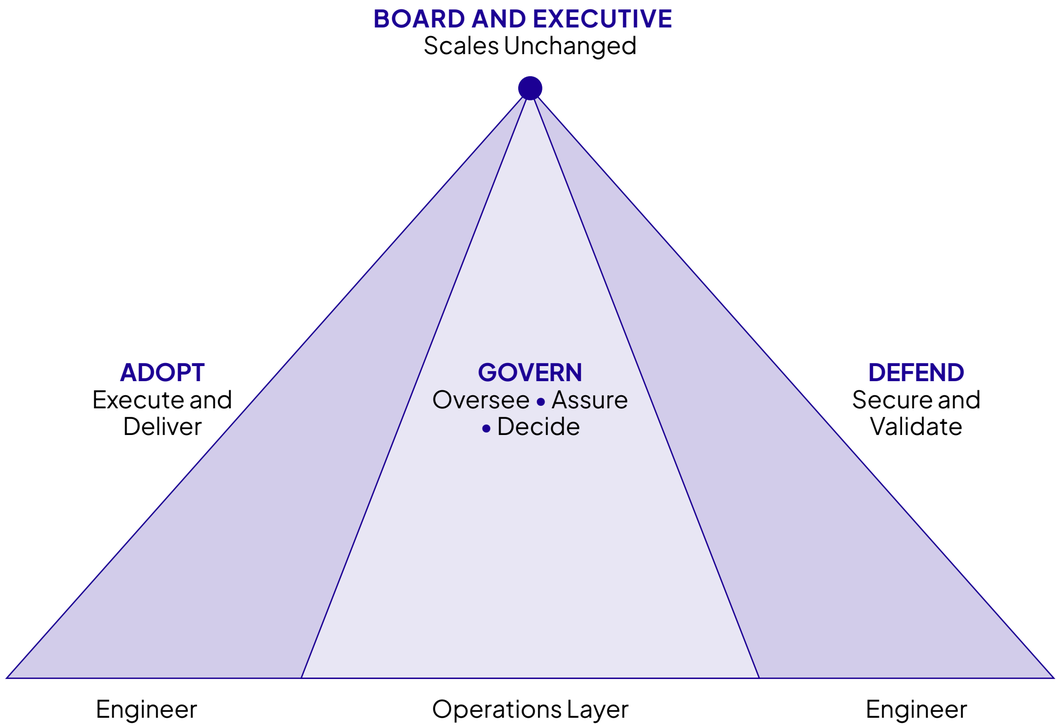


Figure 4.1 — The ADG Triad: three pillars that scale from board to engineer

4.1 Adopt—Execute and Deliver

"Deliver business value safely."

PRIMARY QUESTION	Can we deploy it effectively?
DECISION FOCUS	Capability, performance, delivery, reliability.
STAKEHOLDERS	AI Product Owner · ML Engineers · Prompt Engineers · DevSecOps · App Architects · Enterprise Architects.
CORE OUTPUTS	Deployed service, runbooks, baselines, releases.
SUCCESS MEASURE	Value delivery with controlled operations.
ESCALATION	Escalate blockers to GOVERN.

4.2 Defend—Secure and Validate

"Identify, prevent, detect harmful behavior."

PRIMARY QUESTION	Can it be abused, fail dangerously, or cause harm?
DECISION FOCUS	Security, resilience, fairness, abuse resistance, containment.
STAKEHOLDERS	AI Red Team · Security Engineers · Bias Auditors · Incident Response · Guardrail Engineers · Detection Engineers.
CORE OUTPUTS	Test results and detections, guardrails, fairness evaluations, incident playbooks.
SUCCESS MEASURE	Risk reduced across all four harm classes.
ESCALATION	Escalates unresolved risks to GOVERN.

Govern—Oversee, Assure, and Decide

"Justify, approve, evidence AI use at board level."

PRIMARY QUESTION	Should we approve it, under what conditions, and at what risk?
DECISION FOCUS	Risk appetite, legality, accountability, ethics, oversight.
STAKEHOLDERS	Chief AI Officer/Ethics Lead · Legal and Compliance · Risk Officers · Board/C-suite · Privacy Counsel · Compliance Lead
CORE OUTPUTS	Policies, approvals, risk thresholds; exceptions and evidence; board reports.
STRATEGIC	Define organizational AI risk appetite; establish decision rights and escalation paths;
FUNCTIONS	Own the AI Governance Council charter; define board-level reporting requirements; set investment justification criteria; maintain regulatory mapping.
SUCCESS MEASURE	Defensible use with auditable, measurable governance.
TENSION RESOLUTION	Resolves ADOPT-DEFEND tension; defines go/no-go and exception policy.

Operating Model

ADG uses a simple rule: **ADOPT** builds and operates, **DEFEND** breaks and protects, **GOVERN** authorizes and oversees.

DIMENSION	ADOPT	DEFEND	GOVERN
Primary Question	Can we deploy it effectively?	Can it be abused, fail dangerously, or cause harm?	Should we approve it, under what conditions, and at what risk?
Decision focus	Capability, performance, delivery, reliability	Security, resilience, fairness, abuse resistance, containment	Risk appetite, legality, accountability, ethics, oversight
Core outputs	Deployed service, runbooks, baselines, releases	Test results, detections, guardrails, fairness evaluations, incident playbooks	Policies, approvals, risk thresholds, exceptions, evidence, board reports
Success measure	Value delivery with controlled operations	Risk reduced across all four harm classes	Defensible use with auditable, measurable governance
Tension resolution	Escalates blockers to GOVERN	Escalates unresolved risks to GOVERN	Resolves ADOPT-DEFEND tension; defines go/no-go and exception policy

5 Governance Surfaces

ADG organizes governance into nine surfaces. These surfaces define what must be governed regardless of model vendor, model architecture, or deployment pattern.

GOVERNANCE SURFACES

Nine Surfaces That Must Be Governed Regardless of Architecture

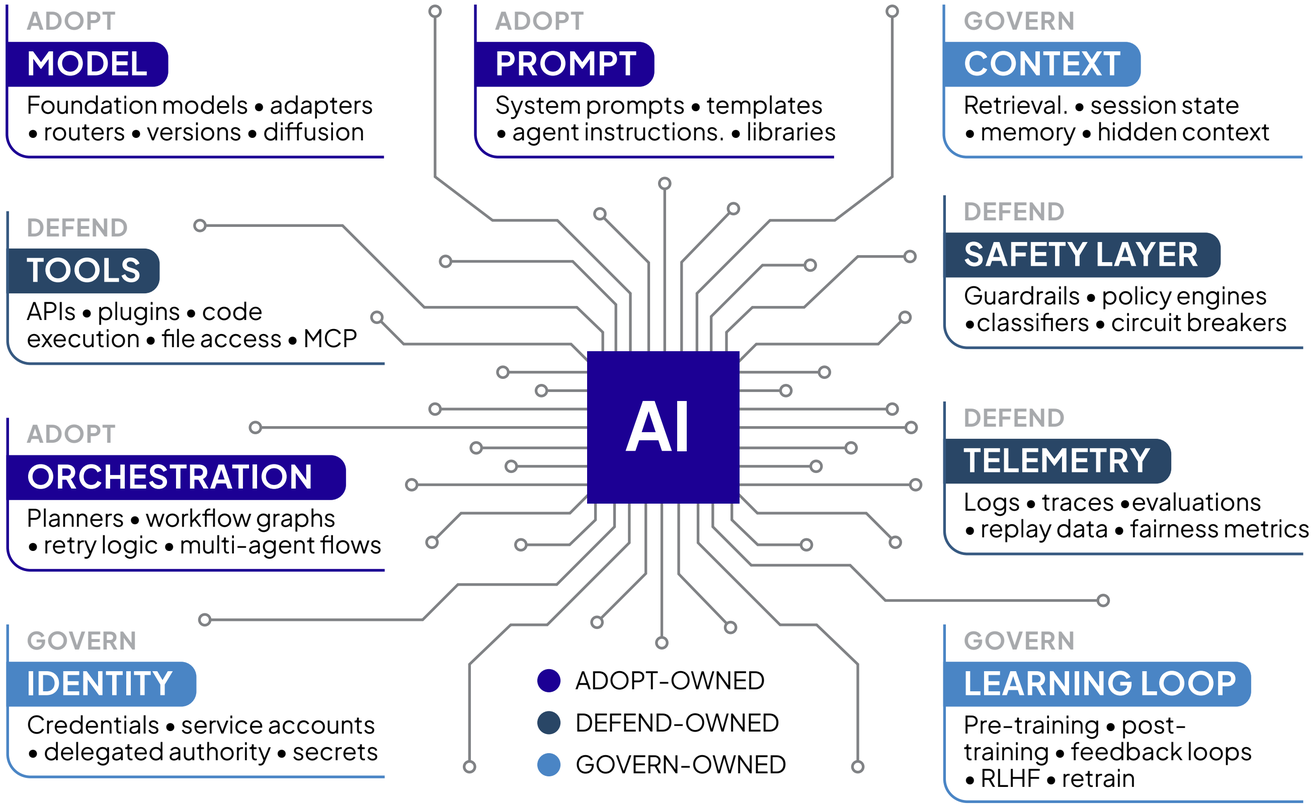


Figure 5.1 — Nine Governance Surfaces: what must be governed regardless of architecture

SURFACE	SCOPE	PILLAR	CONTROL OBJECTIVE
Model	Foundation models, fine-tuned models, adapters, routers, versions, diffusion models, composite model chains	A G	Use only approved models with known risk posture, provenance, and change traceability
Prompt	System prompts, templates, policies, agent instructions, prompt libraries, multi- modal input validation	A D	Prevent unmanaged behavior changes and unsafe instruction patterns
Context	Retrieval sources, session state, memory, hidden context, user metadata, cross-session data	G D	Prevent poisoning, leakage, cross- session contamination, and privacy violations
Tools	APIs, plugins, actions, code execution, file access, transactional endpoints, MCP capabilities	D	Enforce least privilege, strong validation, sandboxing, and full audit logging
Orchestration	Planners, workflow graphs, retry logic, multi-agent flows, model routing, agent-to-agent communication	A D	Bound agent behaviour, prevent cascading failures, ensure deterministic control
Identity	Credentials, service accounts, delegated authority, secrets, trust relationships, agent identity	G D	Prevent privilege misuse, preserve accountability, trace agent actions to human authority
Safety Layer	Guardrails, policy engines, semantic filters, classifiers, circuit breakers, harm detectors	D	Block unsafe content, unfair outputs, and unauthorized actions before impact
Telemetry	Logs, traces, evaluations, replay data, alerts, governance evidence, fairness metrics	D G	Make behaviour observable, reviewable, provable, and measurable
Learning Loop	Pre-training sources, post-training alignment, feedback loops, retraining updates, reinforcement learning from human feedback (RLHF) data	G A	Control data provenance, drift, alignment stability, and undocumented behavior change

6 Harm Taxonomy and Responsible AI Integration

TL; DR

Four classes of AI harm (technical, societal, operational, systemic) along with seven Responsible AI principles, enabling governance to cover attacks, bias, reliability, and emergent multi-agent risk within **one taxonomy.**

ADG recognizes that AI governance must address harms beyond security vulnerabilities. The framework adopts a four-class harm taxonomy that aligns with every governance surface, life cycle stage, and deployment overlay.

6.1 Harm Classification

The four harm classes are plotted by detection difficulty and impact scope. Each quadrant carries the threats it covers, the pillars that own it, and the Minimum Controls that detect those threats.

HARM CLASSIFICATION

Detection Difficulty x Impact Scope

SCOPE: SYSTEMIC

Hard/Systemic

Agent collusion • Cascading failures • Sycophancy loops • Value misalignment amplification

GOVERN • DEFEND
MC-8 MC-10 MC-9

Hard/Individual

Discrimination • Stereotyping
• Manipulation Surveillance
• Opportunity denial

GOVERN • DEFEND
MC-11 MC-10 MC-2

SYSTEMIC

SOCIETAL

OPERATIONAL

TECHNICAL

EASY TO DETECT

HARD TO DETECT

Easy/Individual

Hallucination • Drift
Grounding failure •Data quality degradation

ADOPT • DEFEND
MC-4 MC-8 MC-5

Easy/Systemic

Prompt injection • Data exfiltration • Model theft
• Jailbreaking • Denial of wallet

DEFEND
MC-7 MC-8 MC-9

SCOPE: INDIVIDUAL

Figure 6.1 — Harm Classification: Detection Difficulty x Impact Scope

CLASS	DETECTION/SCOPE	DESCRIPTION	THREATS	PILLARS	DETECTED BY
Operational	Easy/Individual	Reliability, accuracy, business-impact failures	▶ Hallucination ▶ Model drift ▶ Grounding failure ▶ Data quality ▶ No explainability	A D	MC-4 MC-8 MC-5
Technical	Easy/Systemic	Security, integrity, system reliability failures	▶ Prompt injection ▶ Data exfiltration ▶ Model theft ▶ Jailbreaking ▶ Denial of wallet	D	MC-7 MC-8 MC-9
Societal	Hard/Individual	Bias, discrimination, human-rights impacts	▶ Algorithmic discrimination ▶ Stereotyping ▶ Opportunity denial ▶ Manipulation ▶ Surveillance	G D	MC-11 MC-10 MC-2
Systemic	Hard/Systemic	Emergent risk from AI-to-AI interaction	▶ Agent collusion · ▶ Value misalignment amplification ▶ Cascading failures ▶ Sycophancy loops	G D	MC-8 MC-10 MC-9

6.2 Responsible AI Principles

Every governance surface in ADG must account for seven responsible AI dimensions, embedded across all three pillars rather than appended as a separate track.

PRINCIPLE 1

Fairness and Non-Discrimination

Outputs must not disadvantage groups.

PRINCIPLE 2

Transparency and Explainability

Explainable to its risk tier.

PRINCIPLE 3

Privacy and Data Protection

Minimize. Comply with data law.

PRINCIPLE 4

Accountability

Operator to executive sponsor.

PRINCIPLE 5

Human Oversight

Scaled to autonomy and harm potential.

PRINCIPLE 6

Robustness and Safety

Tested under adversarial conditions.

PRINCIPLE 7

Sustainability and Well-Being

Environmental and societal impact.

INTEGRATION RULE

Responsible AI is not a separate track. It is embedded into every ADG pillar — **ADOPT. DEFEND. GOVERN.**, and must be reflected in all certification curricula, operating artifacts, and governance reviews.

7 Shared Responsibility Model

When organizations deploy AI, responsibility for governance controls varies based on how the AI is provisioned. ADG defines three deployment responsibility classes and maps control ownership for each.

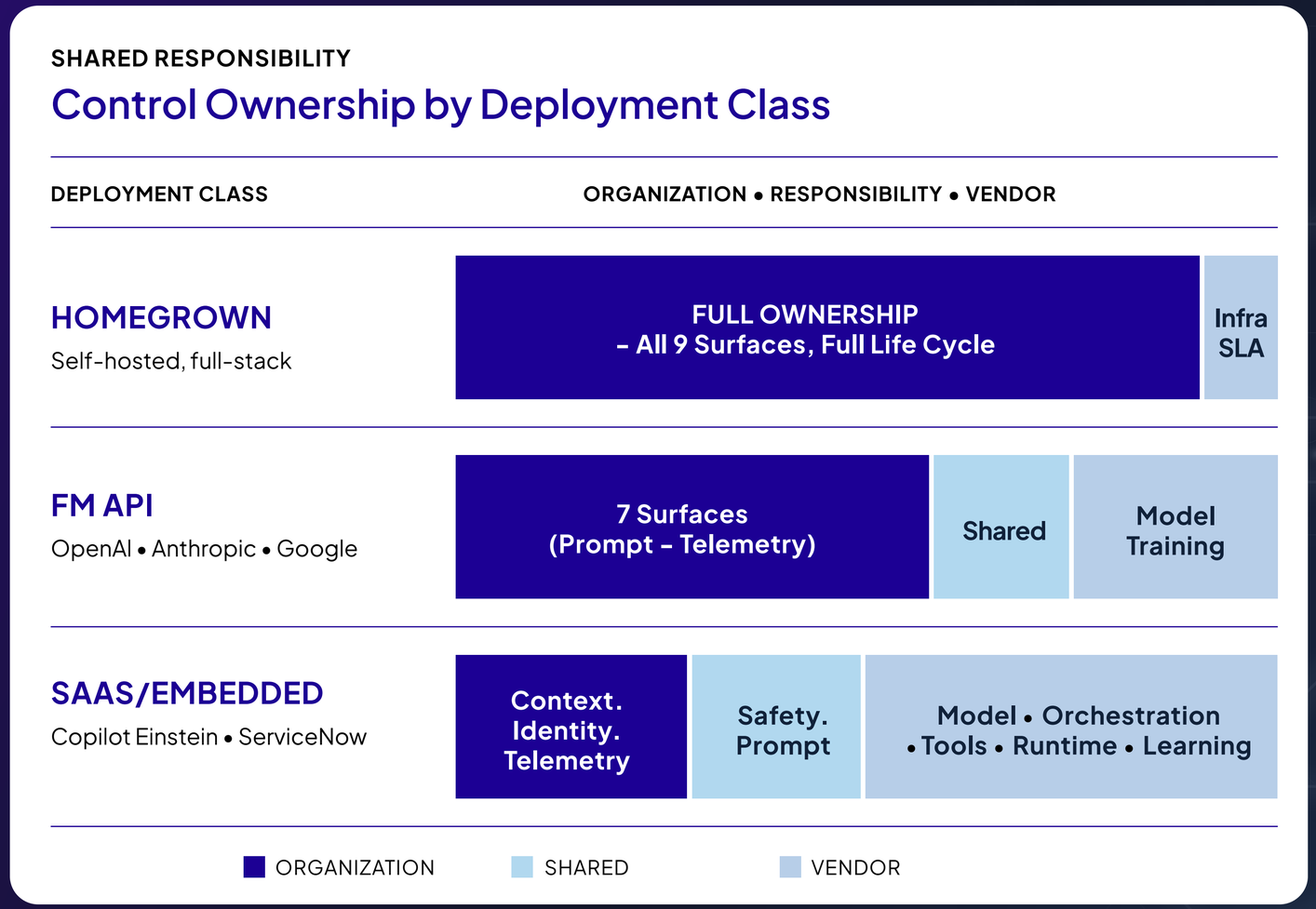
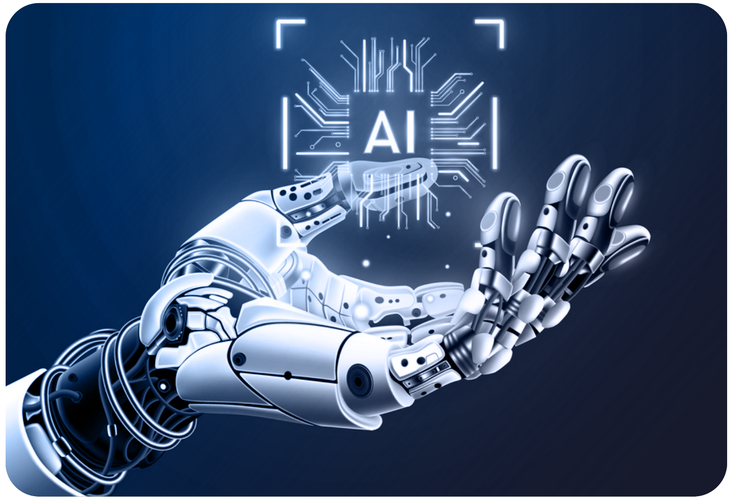


Figure 7.1 — Shared Responsibility: control ownership by deployment class

7.1 Deployment Responsibility Classes

CLASS	DESCRIPTION	ORGANIZATION CONTROLS	SHARED	VENDOR CONTROLS
Homegrown/ Self-Hosted	Organization trains, hosts, and operates the full AI stack	All 9 surfaces, full life cycle	None (full ownership)	Infrastructure SLAs only
Foundation Model API	Consumes a foundation model via API (e.g., OpenAI, Anthropic, Google)	Prompt, Context, Tools, Orchestration, Identity, Safety Layer, Telemetry	Model governance, Learning Loop	Model training, alignment, infrastructure, API availability
SaaS AI/ Embedded AI	AI embedded in a vendor product (e.g., Copilot, Einstein, ServiceNow)	Context, Identity, Telemetry, Governance policy	Safety Layer, Prompt customization	Model, Orchestration, Tools, Runtime stack, Learning Loop



7.2 Vendor AI Due Diligence Requirements

For Foundation Model API and SaaS AI deployments, organizations must:

- Require **vendor disclosure** of model provenance, training data policies, alignment methods, and known limitations
- Establish **contractual AI governance clauses** covering incident notification, data handling, model change management, and liability allocation
- Conduct **independent evaluation** of vendor-provided safety controls and not rely solely on vendor claims
- Maintain **consumer-side telemetry** and monitoring regardless of vendor monitoring capabilities
- Include AI governance questions in **procurement processes** and vendor risk assessments
- Define **rollback and exit strategies** for vendor AI dependencies

PART III — IMPLEMENTATION

People, Process, Technology, and Data

The four operational layers that make ADG executable on the ground—from cross-functional roles to the data pipelines AI consumes.

8 People, Process, Technology, and Data

8.1 People Layer

ADG requires explicit capability ownership in addition to functional ownership. The people model includes dedicated responsible AI expertise across all pillars.

ADOPT ROLES

- AI product owner
- Accountable AI service owner
- LLMOps lead
- Orchestration engineer
- Prompt designer
- Platform engineer
- Enterprise architect
- DevSecOps engineer

DEFEND ROLES

- AI red team lead
- Guardrail engineer
- Detection engineer
- AI security architect
- Incident commander
- Forensic analyst
- Responsible AI scientist
- Bias/fairness auditor

GOVERN ROLES

- Model risk officer
- Privacy counsel
- Compliance lead
- Data steward
- AI ethics and sociotechnical lead
- Executive approver
- Procurement governance lead

AI GOVERNANCE COUNCIL

Every organization deploying AI systems should establish a cross-functional AI Governance Council (or equivalent steering body) with representation from **ADOPT. DEFEND. GOVERN.** This council serves as the **escalation path and tension-resolution mechanism** described in Section 4.4.

8.2 Process Layer

ADG formalizes the minimum process backbone required for enterprise implementation:

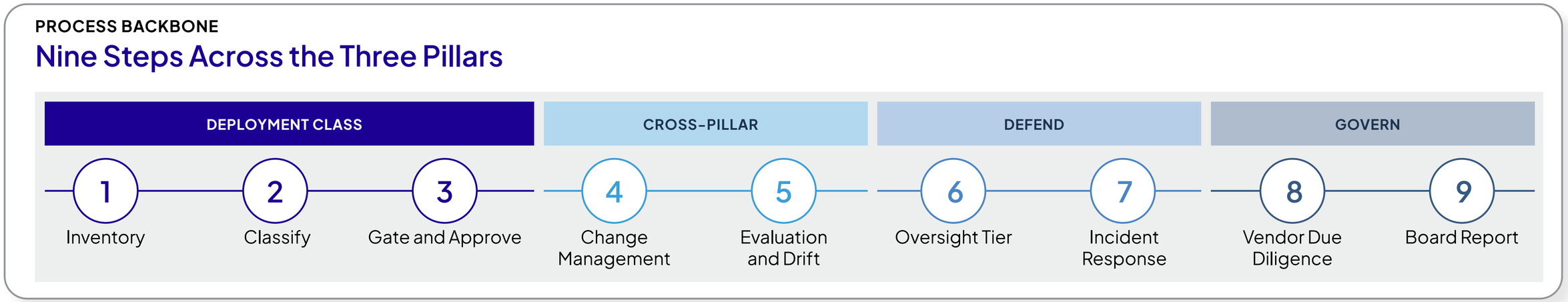
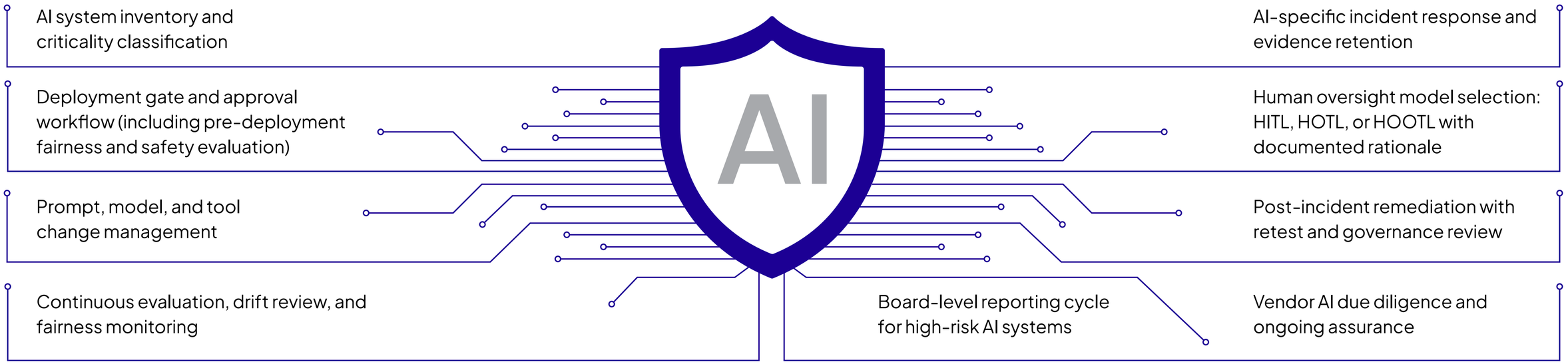


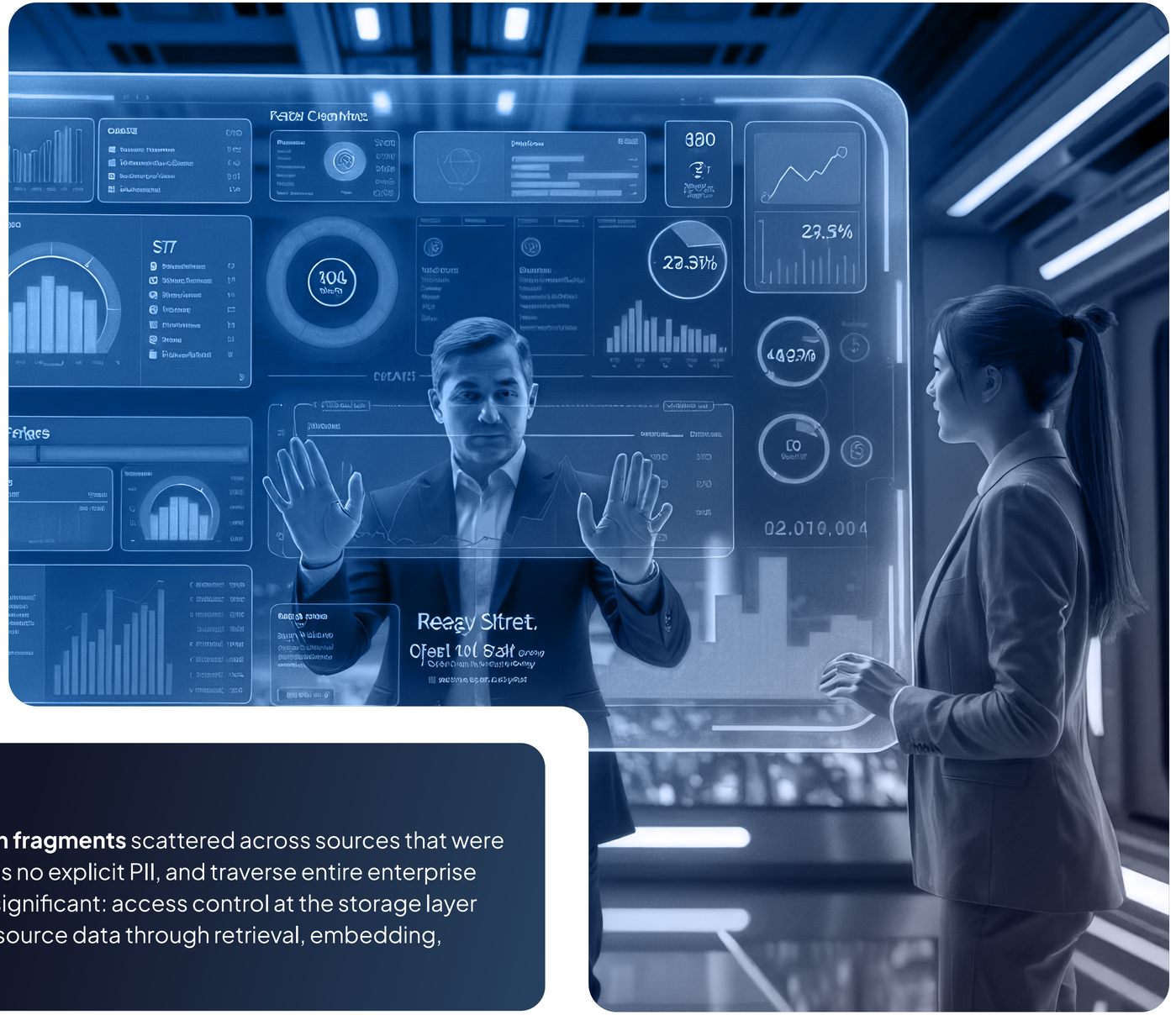
Figure 8.1 — Process Backbone: nine steps across the three pillars



8.3 Technology Layer

ADG treats the following as governed technology layers:

- Model gateway and routing layer
- Retrieval, memory, and context management layer
- Tool registry and MCP trust layer
- Policy engine, guardrail layer, and harm detection layer
- Runtime telemetry, replay, and fairness measurement layer
- Training, tuning, and evaluation pipeline layer
- Input validation and sanitization layer (covering prompts, RAG inputs, and multi-modal inputs)



8.4 Data Layer

Traditional enterprise data architecture separates data into distinct tiers—structured databases, data warehouses, data lakes, unstructured file stores, and APIs—each with its own governance model, access controls, and tooling. **AI systems fundamentally disrupt this separation.** When an AI agent consumes enterprise data through retrieval pipelines, MCP connections, or tool invocations, it does not distinguish between an SQL database record, a SharePoint document, a Slack thread, or a PDF. The data is converted into text tokens within a context window, collapsing traditional tier boundaries into a single consumption surface.

THE DATA CONVERGENCE RISK

This convergence allows AI systems to **reassemble sensitive information from fragments** scattered across sources that were never intended to be combined, infer personal data from context that contains no explicit PII, and traverse entire enterprise file systems where content is readable as text. The governance implication is significant: access control at the storage layer alone is no longer sufficient. Organizations must govern the full pipeline from source data through retrieval, embedding, context assembly, and AI consumption.

8.4.1 DATA OPERATING ROLES

- **Data Stewards for AI:** extending traditional data stewardship to govern the full text-based data surface that AI systems can access
- **AI Data Engineers:** bridging data engineering and AI operations, responsible for RAG indexing, embedding generation, knowledge base curation
- **Context Architects:** designing what data flows into AI context windows, in what priority order, with what trust ranking
- **Knowledge Base Curators:** responsible for freshness, accuracy, deduplication, and retirement of enterprise knowledge assets

8.4.2 DATA PROCESS

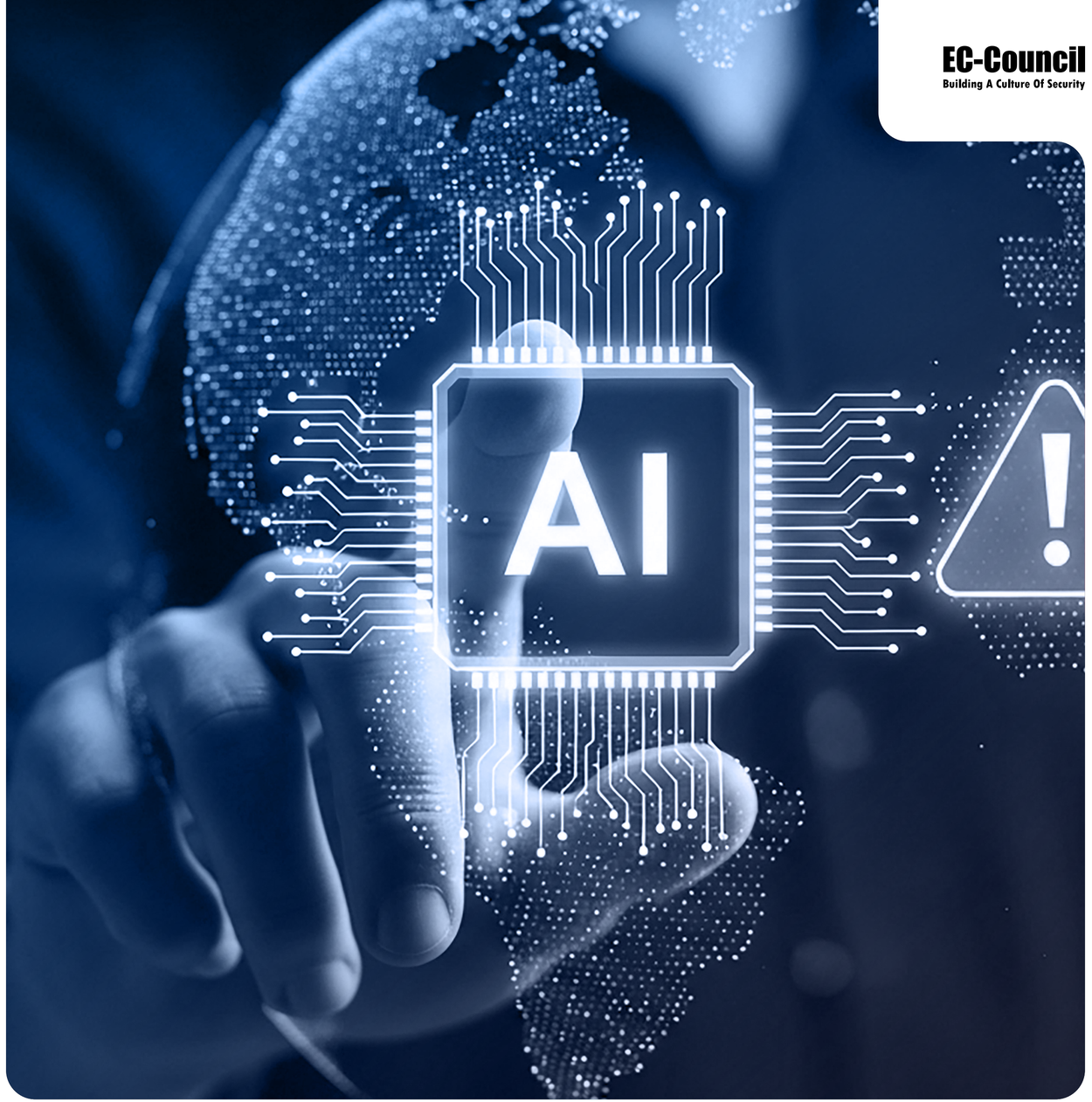
- **AI-aware data classification:** extending beyond storage-tier access control to include AI-readability rules
- **Provenance tracking across the text pipeline:** maintaining a verifiable chain from any AI output back through the retrieval step to the source document
- **Cross-source inference governance:** policies governing when AI systems may combine information from multiple data sources
- **Text-based policy management:** AI system configurations governed as controlled documents with versioning, approval workflows, and rollback
- **Knowledge base life cycle management:** content ingestion, quality validation, freshness review, conflict resolution, deduplication, and retirement
- **Data minimization for AI:** ensuring context windows contain only data necessary for the task
- **Consent and lawful basis tracking:** maintaining records of lawful basis for processing each data category

8.4.3 DATA TECHNOLOGY

- **Vector stores and embedding infrastructure:** governed data infrastructure requiring access controls, encryption, backup and recovery
- **Enterprise knowledge graphs:** structured representations enabling context assembly with awareness of entity relationships
- **MCP and tool registries as data access layers:** governed as data access infrastructure with the same rigor as database connections
- **Context assembly engines:** systems that select, rank, truncate, filter, and compose text from multiple sources
- **Text-based configuration stores:** GitOps-style repositories for all AI system configurations
- **Data lineage and output attribution:** technology to trace which source documents contributed to a specific AI output
- **Embedding pipeline governance:** validation testing, drift monitoring, and re-indexing governance

8.4.4 DATA GOVERNANCE INTEGRATION WITH ADG SURFACES

SURFACE	DATA LAYER INTERSECTION
Model	Training data provenance, fine-tuning data governance, model card data documentation
Prompt	System prompt versioning and change control as governed text artifacts
Context	Retrieval source classification, context assembly governance, cross-source inference controls
Tools	MCP servers as data access gateways; tool-retrieved data classified and logged
Orchestration	Data flow governance across multi-step agent workflows; inter-agent data sharing rules
Identity	Data access tied to agent identity and delegated authority; no implicit data access
Safety Layer	Guardrail configurations as governed text; data-driven harm detection models governed as data assets
Telemetry	Logs and traces as sensitive data requiring retention, redaction, and access governance
Learning Loop	Feedback data, RLHF inputs, and retraining datasets governed as controlled data assets with provenance

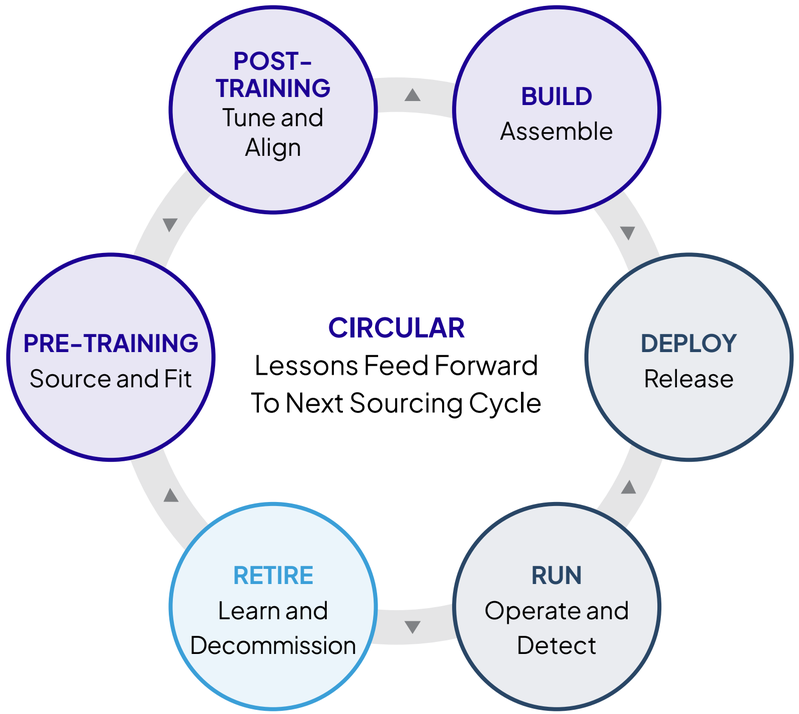


9 Life Cycle Governance

ADG applies controls across six life cycle stages. The life cycle is explicitly circular: lessons from Retire/Learn feed back into Pre-training/Sourcing decisions.

SIX-STAGE LIFE CYCLE

Circular – Lessons Feed Back Into Sourcing Decisions



STAGE	ADOPT	DEFEND	GOVERN
Pre-training	Select suppliers and datasets fit for purpose; assess training data for representation and bias	Assess provenance abuse, contamination risk, and training data bias	Approve sourcing constraints, licensing, jurisdictional requirements, and data ethics
Post-training	Tune for use-case quality and operational fit	Test for regressions, bypasses, safety degradation, and fairness drift	Review alignment objectives, documentation sufficiency, and RAI criteria
Build	Assemble workflows, prompts, tools, retrieval; integrate DevSecOps controls	Validate interfaces, secrets, attack surfaces, and input validation coverage	Classify use case, approve controls, define oversight requirements and risk tier
Deploy	Release through controlled change process with rollback readiness	Confirm pre-production testing, monitoring readiness, and fairness evaluation	Grant formal deployment approval or exception with documented conditions
Run	Operate service, maintain SLAs, track quality and cost	Detect abuse, failures, drift, unsafe actions, and bias emergence; continuous red teaming	Review incidents, exceptions, compliance posture, and configuration drift
Retire	Decommission services and roll forward lessons	Preserve evidence, investigate failures, validate closure	Update policy, records, accountability decisions; feed lessons into sourcing cycle

Post-Deployment Continuous Governance

The Run/Monitor stage requires specific continuous governance controls that go beyond traditional operational monitoring:

- **Configuration drift detection:** verify that the system in production matches the approved models, prompts, tools, and context sources
- **Usage authorization monitoring:** confirm that approved users operate within approved purposes within defined boundaries
- **Feature and capability change governance:** ensure that model updates, prompt changes, tool additions, and retrieval source changes post-deployment go through change control
- **Continuous automated evaluation:** conduct scheduled adversarial testing, fairness benchmarking, and accuracy regression testing on production systems
- **Ongoing governance of product roadmap:** feature additions to deployed AI systems require reevaluation against the original risk classification and approval conditions

10 Deployment Pattern Overlays

ADG adds deployment-specific overlays so that governance matches the architecture being deployed. Overlays are additive—select all that apply.

Overlay Applicability Matrix

OVERLAY	HOMEGROWN	FM API	SAAS	HITL	HOTL	HOOTL
LLMOps	Required	Required	Recommended	Required	Required	Required
Agentic Orchestration	Recommended	Recommended	N/A	Optional	Required	Required
Agent	Recommended	Recommended	Optional	Optional	Required	Required
Agentic Hardening	Required	Required	Recommended	Optional	Required	Required
Tools and MCP	Recommended	Required	Optional	Required	Required	Required
Context and Long- Window	Required	Required	Recommended	Required	Required	Required
Pre/Post-Training	Required	Optional	N/A	Recommended	Recommended	Required
Multi-Agent Interoperability	Recommended	Recommended	Optional	Optional	Recommended	Required
Multimodal / Composite	Recommended	Recommended	Optional	Optional	Recommended	Required

10.1 LLMOps Overlay

Scope: versioning, evaluation, release management, rollback, cost control, and performance monitoring for LLM-based services

ADOPT

Structures model onboarding, baseline evaluation, configuration versioning, rollback readiness, and operational service-level objectives (SLOs)

DEFEND

Structures adversarial testing, abuse simulation, leakage testing, denial-of-wallet controls, monitoring coverage, and bias evaluation

GOVERN

Structures use-case approval, vendor due diligence, deployment thresholds, exception management, and LLMOps cost governance

10.2 Agentic Orchestration Overlay

Scope: planners, loops, retries, multi-step reasoning, subtask decomposition, and multi-agent coordination

ADOPT

Defines mission scope, stop conditions, retry budgets, and workflow boundaries

DEFEND

Validates loop abuse resistance, prompt chaining resistance, recursion limits, and kill-switch behavior

GOVERN

Approves autonomy tier, escalation path, and legal accountability for delegated decisions

10.3 Agent Overlay

Scope: AI systems that act on behalf of a user, team, or enterprise process rather than merely generating content

ADOPT

Defines business mission, action boundaries, and acceptable failure modes

DEFEND

Tests transaction safety, impersonation resistance, and unsafe action prevention

GOVERN

Defines liability model, mandatory approvals, record retention, and segregation-of- duty requirements

10.4 Agentic Hardening Overlay

Required control themes:

AUTHORITY and ISOLATION

- Delegated authority bounds
- Action-layer isolation and dry-run modes
- Semantic firewalls before tool use and before response release

STATE and RECOVERY

- Memory hygiene and state reset controls
- Human override and emergency stop pathways
- Forensic replay of prompts, retrieved context, policy decisions, and tool calls

10.5 Tools and MCP Overlay

Scope: tool discovery, capability registration, trust mediation, and policy-controlled invocation of tools and context providers

ADOPT
Justifies why each tool or MCP capability is needed and what business task boundary it serves

DEFEND
Enforces authentication, authorization, parameter validation, rate limiting, sandboxing, and audit logging

GOVERN
Maintains approval policy, third-party assurance criteria, trust tiers, and data-sharing restrictions

MINIMUM MCP REQUIREMENTS

Approved server inventory · Capability trust tiering · Per-invocation policy evaluation · Auditable request and response traces · **Explicit rejection of implicit trust in tool-supplied instructions**

10.6 Context and Long-Window Overlay

Scope: session history, retrieved enterprise knowledge, hidden orchestration instructions, persistent memory, and context overflow handling

ADOPT
Defines the minimum context required for task quality

DEFEND
Tests for context poisoning, leakage, overflow abuse, cross-session contamination, and multi-turn attack patterns

GOVERN
Defines provenance rules, retention, data-class handling, trust ranking, and lawful-use constraints

MINIMUM CONTEXT POLICY

Every production system must define **approved sources, retention periods, redaction rules, persistence rules, and trust ordering** across context inputs.

10.7 Pre-Training and Post-Training Overlay

Scope: model provenance, fine-tuning, adapters, preference alignment, safety tuning, feedback loops, and retraining

ADOPT
Structures fine-tuning objectives, experiment tracking, rollback, and operational performance baselines

DEFEND
Structures regression testing, jailbreak retesting, alignment failure analysis, fairness drift monitoring, and bias evaluation

GOVERN
Structures provenance review, licensing, cross-border constraints, GDPR compliance for training data, documentation, and approval of learning-loop inputs

10.8 Multi-Agent and Agent Interoperability Overlay

Scope: systems where multiple AI agents interact, coordinate, negotiate, or delegate tasks to each other. Multi-agent systems create emergent governance challenges that single-agent controls do not address:

- **Agent-to-agent trust boundaries:** each agent interaction must enforce explicit trust verification; no agent should implicitly trust another agent's outputs or instructions.
- **Value alignment verification:** agents with different guardrail configurations, different training, or different vendors may have incompatible safety policies. Cross-agent interactions must be tested for value alignment conflicts.
- **Collusion and deception detection:** monitoring for patterns where agents coordinate to bypass controls, produce misleading outputs, or exploit gaps between their respective guardrails.
- **Cascading failure and amplification risk:** a failure or bias in one agent can be amplified through a chain of dependent agents. Circuit breakers must exist at agent-to-agent boundaries.
- **Cross-agent audit trails:** every agent-to-agent interaction must be logged with sufficient detail for forensic replay.
- **Sycophancy degradation prevention:** in agent-to-agent interactions, sycophantic behavior can devolve into harmful feedback loops. Detection and interruption mechanisms are required.
- **Interoperability governance:** when agents from different organizations or vendors interact, a shared governance protocol must define minimum safety, logging, and accountability requirements.

10.9 Multimodal and Composite AI Systems Overlay

Scope: systems combining multiple model architectures (LLM + diffusion + classifier + retrieval) within a single product or workflow

ADOPT

Defines the system-level architecture, ensuring each model component has a clear purpose and integration specification

DEFEND

Tests for cross-model interaction risks: outputs from one model becoming adversarial inputs to another, amplification of individual model weaknesses through the chain.

GOVERN

Requires system-level risk assessment, not just per- model evaluation, and defines composite system approval criteria.

PART IV — CONTROLS AND COMPLIANCE

Controls, Artifacts, and Measurement

Twelve auditable controls; 11 required artifacts; three measurement tiers—each with named evidence that an auditor can review.

11 Controls, Artifacts, and Measurement

11.1 Minimum Control Set

The following controls form the minimum baseline for ADG implementation. Each control includes an evidence requirement to ensure measurability.

MC-1
AI System Inventory

Maintain a n inventory of all AI systems with an accountable owner, risk classification, and autonomy tier.

EVIDENCE: Published inventory, reviewed quarterly, with named owner per system

MC-2
Risk Classification

Classify each AI system by data sensitivity, autonomy, external exposure, harm potential, and business criticality.

EVIDENCE: Documented classification per system using a standardized risk taxonomy

MC-3
Separation of Duties

Separate deployment ownership, security validation, and approval authority across **ADOPT, DEFEND, GOVERN.**

EVIDENCE: RACI matrix per AI system; no single function holding all three roles

MC-4
Pre-Production Evaluation

Complete quality, safety, security, fairness, and failure-mode testing before any production deployment.

EVIDENCE: Signed evaluation report covering all four harm classes before go-live

MC-5
Change Control

Manage changes to prompts, tools, models, and retrieval sources through a governed change process.

EVIDENCE: Change log with approval records; no uncontrolled production changes

MC-6
Context Policy

Define provenance, retention, access restrictions, and trust ordering for all context inputs.

EVIDENCE: Published context policy per system; annual review

MC-7

Tool and MCP Register

Maintain a register of all tools and MCP capabilities with trust tiering and invocation controls.

EVIDENCE: Published register with per-tool risk assessment and approval status

MC-8

Runtime Monitoring

Monitor for abuse, drift, data leakage, unsafe actions, bias emergence, and configuration drift in production.

EVIDENCE: Active monitoring with defined alert thresholds and response SLAs

MC-9

AI Incident Response

Maintain AI-specific incident response procedures with replayable evidence capture.

EVIDENCE: Documented playbook; at least one tabletop exercise per year

MC-10

Periodic Governance Review

Conduct governance reviews with exception handling and board reporting for high-risk systems.

EVIDENCE: Review records with findings, decisions, and exception dispositions

MC-11

Fairness and Bias Evaluation

Evaluate AI systems for discriminatory outcomes using representative test data and established fairness metrics.

EVIDENCE: Fairness evaluation report; re-evaluation after model or data changes

MC-12

Shared Responsibility Documentation

For vendor or SaaS AI, document accountability boundaries, contractual obligations, and assurance requirements.

EVIDENCE: Signed responsibility matrix; vendor due diligence records

11.2 Control Crosswalk — NIST AI RMF and ISO/IEC 42001

Each Minimum Control aligns with one or more NIST AI RMF functions and ISO/IEC 42001 Annex A groups.

CONTROL	NAME	NIST AI RMF	ISO/IEC 42001 ANNEX A	#NIST	#ISO
MC-1	AI System Inventory	GOVERN, MAP	A.3 Org · A.6 Life Cycle · A. 8 Info	2	3
MC-2	Risk Classification	MAP, MEASURE	A.5 Impact · A.6 Life Cycle	2	2
MC-3	Separation of Duties	GOVERN	A.2 Policies · A.3 Org	1	2
MC-4	Pre-Production Evaluation	MEASURE	A.6 Life Cycle · A.7 Data	1	2
MC-5	Change Control	MEASURE, MANAGE	A.6 Life Cycle	2	1
MC-6	Context Policy	MAP, MEASURE	A.7 Data · A.8 Info	2	2
MC-7	Tool and MCP Register	GOVERN, MAP	A.6 Life Cycle · A.10 Third- party	2	2
MC-8	Runtime Monitoring	MEASURE, MANAGE	A.6 Life Cycle · A.9 Use	2	2
MC-9	AI Incident Response	GOVERN, MANAGE	A.6 Life Cycle · A.9 Use · A. 10 Third-party	2	3
MC-10	Periodic Governance Review	GOVERN, MANAGE	A.2 Policies · A.3 Org · A.9 Use	2	3
MC-11	Fairness and Bias Evaluation	MEASURE	A.5 Impact · A.7 Data	1	2
MC-12	Shared Responsibility Doc	GOVERN, MAP	A.8 Info · A.10 Third-party	2	2

11.3 Required Operating Artifacts

ARTIFACT	PURPOSE	ADG ALIGNMENT
AI System Profile	Documents system architecture, risk classification, autonomy tier, harm classes, and deployment pattern	A
ADG RACI Matrix	Maps control ownership across **ADOPT. DEFEND. GOVERN.** for each AI system	G
Agent Authority Statement	Defines what an agent may access, decide, execute, and escalate	G A
Context Policy	Specifies approved sources, retention, redaction, trust ordering, and privacy controls	G
Tool and MCP Register	Catalogs tools, trust tiers, invocation policies, and third-party assurance status	D
AI Release Gate Checklist	Pre-deployment verification covering security, fairness, safety, and governance approval	A D
AI Incident Evidence Pack	Forensic capture package for AI-specific incidents	D
Model and Prompt Change Log	Tracks all changes to models, prompts, and system configurations with approval records	A
Governance Exception Register	Records all governance exceptions with justification, risk acceptance, and expiration	G
Vendor AI Due Diligence Record	Documents vendor AI assessments, contractual AI clauses, and shared responsibility boundaries	G
Fairness Evaluation Report	Records fairness and bias testing methodology, results, and remediation actions	D G

11.4 Measurement and Evidence Framework

ADG requires **measurable governance.** Three measurement tiers ensure that controls are not only documented but also demonstrably effective.

TIER – 1

CONTROL EFFECTIVENESS METRICS (PER-CONTROL KPIS·OPERATIONAL)

- % of AI systems with completed risk classification
- % of AI systems with documented RACI matrix.
- Mean time from model change to governance approval
- % of Tool/MCP capabilities with current trust assessment.
- % of production AI systems with active runtime monitoring·
- # uncontrolled production changes detected per quarter·
- % of high-risk AI systems with completed fairness evaluation.

TIER – 2

OUTCOME METRICS (CROSS-SURFACE GOVERNANCE EFFECTIVENESS)

- Mean time to detect AI-specific incidents (MTTD).
- Mean time to contain AI-specific incidents (MTTC).
- # AI governance exceptions open beyond expiration date.
- % of AI systems operating within approved configuration (no drift).
- # AI-related compliance findings from internal/external audit.
- Quarter-over-quarter trends in fairness metrics across production systems.

TIER – 3

BOARD – LEVEL INDICATORS (EXECUTIVE REPORTING · AGGREGATED)

- AI risk posture score (composite across all surfaces).
- % of AI compliance coverage (systems with current governance review).
- Exception backlog trend (open, aging, risk-weighted).
- AI incident trend (frequency, severity, resolution time).
- # Vendor AI risk exposure (third-party model dependencies).
- % of Responsible AI compliance rate (fairness, transparency, accountability).

12 Regulatory Alignment and Roadmap

12.1 Regulatory and Standards Alignment

ADG is designed to align with major AI governance regulations and standards. The following matrix maps ADG components to key external frameworks.

NOTE: This mapping is indicative. Organizations must conduct their own regulatory compliance assessment. ADG provides a governance backbone that **facilitates compliance but does not guarantee it.**

For the full 12-control Minimum Control Set mapped to NIST AI RMF and ISO/IEC 42001, see § 11.2 Control Crosswalk. The table below covers EU AI Act alignment and cross-cutting ADG components.

ADG COMPONENT	EU AI ACT (HIGH- RISK FOCUS)	NIST AI RMF	ISO/IEC 42001
Risk Classification (MC-2)	Art. 6–7: Risk categorization	Map: Risk identification and analysis	6.1.2: AI risk assessment
Pre-Production Evaluation (MC-4)	Art. 9: Risk management system	Measure: AI risk measurement	8.1: Operational planning and control
Runtime Monitoring (MC-8)	Art. 72: Post-market monitoring	Manage: Continuous monitoring	9.1: Monitoring, measurement, analysis
Fairness Evaluation (MC-11)	Art. 10: Data governance and bias prevention	Map: Bias identification	A.7: Data + A.5: Impact assessment
Incident Response (MC-9)	Art. 73: Serious incident reporting	Manage: Incident response	10.2: Nonconformity and corrective action
AI System Inventory (MC-1)	Art. 49 + Art. 71: Registration in EU database	Govern: Inventory and categorization	7.5: Documented information
Human Oversight (Tiers)	Art. 14: Human oversight requirements	Govern: Human-AI teaming	A.9: Use of AI systems + A.5: Impact
Transparency (Telemetry)	Art. 13: Transparency requirements	Govern: Transparency and documentation	A.8: Information for interested parties

12.2 Implementation Roadmap

A four-phase staged path from inventory to continuous assurance, with controls deployed in dependency order.

PHASE – 1

FOUNDATION (MONTHS 0–3)

- AI system inventory
- Assign owners
- Define ADG roles
- Classify by criticality
- Governance Council
- Baseline gap analysis

Controls: MC-1, MC-2, MC-3

PHASE – 2

CONTROL DEPLOYMENT (MONTHS 3–9)

- Release gates
- Change control
- Context policies
- Adversarial testing
- Fairness evals
- Vendor due diligence

Controls: MC-4 through MC-9, MC-11, MC-12

PHASE – 3

AGENTIC READINESS (MONTHS 9–15)

- Authority statements
- Tool trust tiering
- MCP governance
- Circuit breakers
- Multi-agent governance
- Forensic replay

Controls: MC-5, MC-7, MC-8, MC-9

PHASE – 4

MATURITY (MONTHS 15–24)

- Continuous evaluation
- Adversarial monitoring
- Drift governance
- RAI measurement
- Board reporting
- Formal assurance

Controls: MC-8, MC-10, MC-11

13 Workforce Capability

Closing the AI Skills Gap.

13.1 The AI Skills Gap

The defining constraint on enterprise AI is not the technology — it is the workforce. AI is scaling into production faster than the people who must run, secure, and govern it can be trained, credentialed, and deployed.

90%
ORGANIZATIONS PROJECTED TO FACE SKILLS SHORTAGES BY 2026

$5.5T
IN PROJECTED LOSSES FROM DELAYS, WEAKER COMPETITIVENESS

3
CAPABILITY GAPS MAPPED TO ADG PILLARS

Source: IDC, Enterprise Resilience: IT Skilling Strategies, 2024.

"AI is moving from experimentation to infrastructure, and the workforce has to move with it. Security leaders are now accountable for systems that learn, adapt, and influence outcomes at speed."

– **JAY BAVISI**, FOUNDER AND GROUP PRESIDENT, EC-COUNCIL GROUP

The ADG framework defines three pillars — **ADOPT, DEFEND, GOVERN** — and each pillar highlights a distinct capability gap emerging across enterprises today:

THE ADOPT GAP

Program and transformation leaders cannot translate AI strategy into measurable, governed delivery— **stalling pilots before they reach production**

THE DEFEND GAP

Security organizations lack adversarial AI expertise— prompt injection, model evasion, data poisoning, and multi-agent exploitation are **not yet standard tradecraft**

THE GOVERN GAP

Risk, compliance, and board leaders lack the vocabulary, evidence protocols, and regulatory mapping (EU AI Act, NIST AI RMF, ISO/IEC 42001) **to hold AI systems to account**

13.2 The Capability Bridge

EC-Council has developed the **Enterprise AI Credential Suite**—a portfolio of role-aligned credentials mapped directly to the ADG framework. Launched in February 2026 as the largest single portfolio expansion in EC-Council's 25-year history, the suite is structured as a four-part capability ladder:

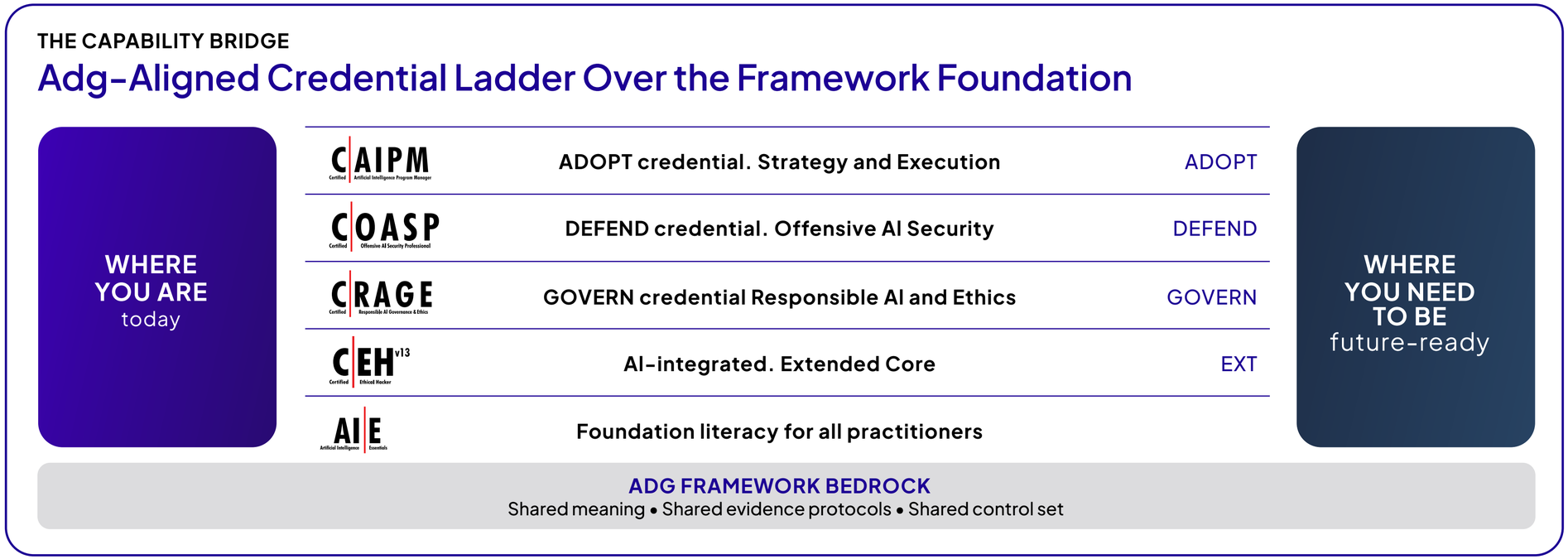


Figure 13.1 — The Capability Bridge: ADG-Aligned Credential Ladder Over the Framework Foundation

Literacy baseline —a shared AI vocabulary for every practitioner, technical or not, so governance conversations can happen at enterprise speed—anchored by the **A|IE** program.

Role-aligned capability — three flagship credentials mapped 1:1 to the ADG pillars—**C|AIPM** for Adopt, **C|OASP** for Defend, and **C|RAGE** for Govern. Each credential certifies execution-grade competence for the professionals accountable for that pillar

Extended core skills — C|EH AI restructured to integrate AI modules across all five phases of ethical hacking, extending existing security practitioners into the AI attack surface without a role change.

Operational readiness — each credential mapped to the framework's Minimum Control Set (MC-1 through MC-12), turning certification into deployable evidence that auditors, regulators, and boards can accept.

13.3 Credential Portfolio

The five credentials below are aligned with the four-part capability ladder in Section 13.2. Each credential is linked to its official EC-Council program page for syllabus, prerequisites, and enrollment.

AIE Artificial Intelligence Essentials

ARTIFICIAL INTELLIGENCE ESSENTIALS (FOUNDATION LITERACY)

Non-technical AI literacy covering core principles, prompt engineering, AI ethics, and tool fluency. Five hands-on modules, no coding prerequisites.

Audience: All practitioners, educators, learning and development (L&D) professionals

CAIPM Certified Artificial Intelligence Program Manager

CERTIFIED AI PROGRAM MANAGER (ADOPTPILLAR)

Translates AI strategy into execution—strategy, governance, and enterprise AI life cycle delivery for program leaders accountable for ROI.

Audience: AI program managers, transformation leads

COASP Certified Offensive AI Security Professional

CERTIFIED OFFENSIVE AISECURITY PROFESSIONAL (DEFENDPILLAR)

End-to-end AI red-teaming—prompt injection, model evasion, data poisoning, model exploitation, agentic/model-to-model attacks, and AI supply-chain attacks. For security teams defending production AI.

Audience: Security engineers, AI red teamers

CRAGE Certified Responsible AI Governance & Ethics

CERTIFIED RESPONSIBLE AI GOVERNANCE AND ETHICS (GOVERNPILLAR)

AI risk management, alignment with NIST AI RMF and ISO/IEC 42001, and accountability across the AI life cycle for governance and compliance leaders.

Audience: CISOs, governance leaders, and compliance leaders

CEH v13 Certified Ethical Hacker

CERTIFIED ETHICAL HACKER (AI-INTEGRATED·EXTENDEDCORE)

Ethical hacking curriculum updated to integrate AI across the full ethical hacking life cycle—reconnaissance, scanning, gaining access, maintaining access, covering tracks—closing the AI chasm for existing security practitioners without a role change.

Audience: Ethical hackers, pen testers

The full EC-Council AI credential catalog is maintained at eccouncil.org/ai-courses.

REFERENCE

Layer-by-Layer Reference

Concrete definitions and one example per element in the Framework Architecture diagram. Use this as a quick-reference guide when reading the remainder of the document.

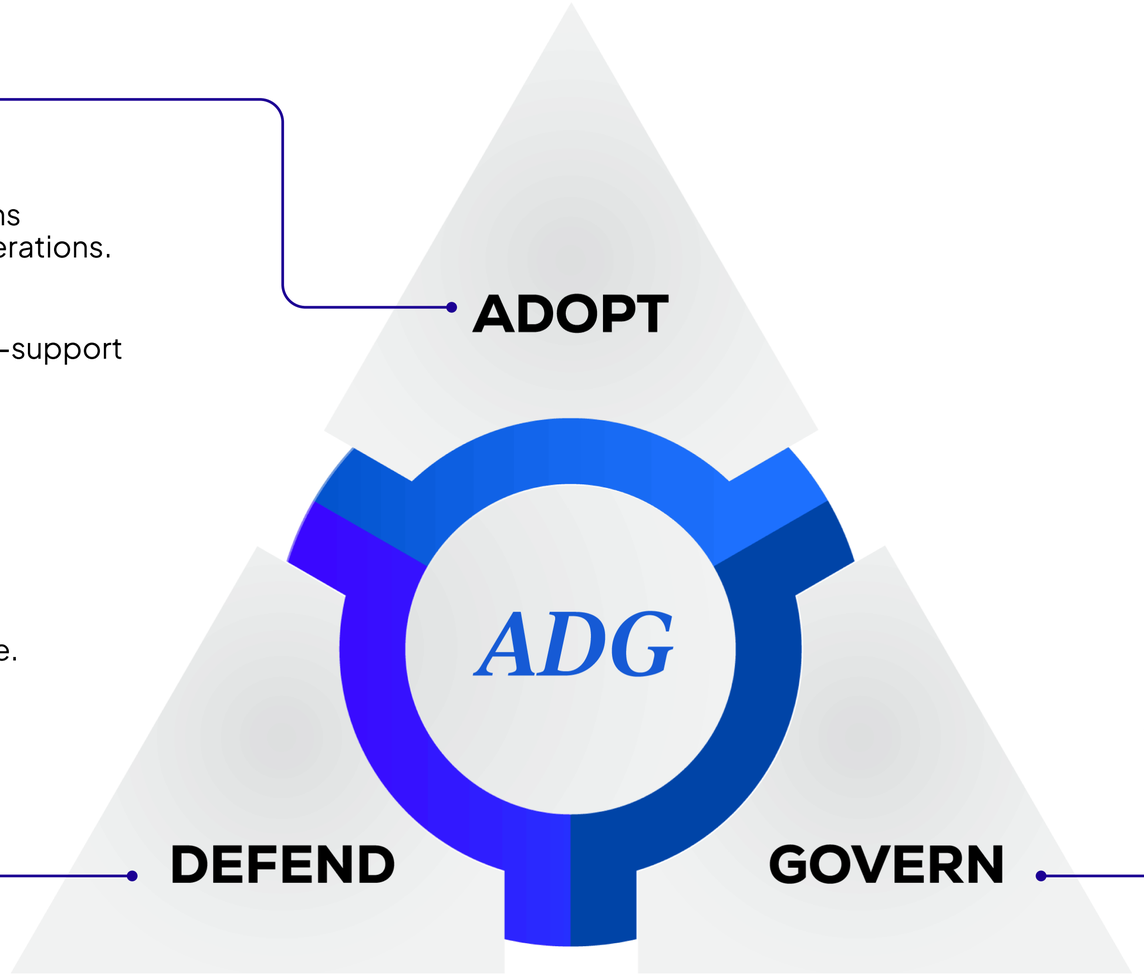
3 Pillars — Operating Model

DEFINITION

Execute and deliver AI capabilities. Owns selection, integration, deployment, operations.

EXAMPLE

An LLMOps team rolling out a customer-support copilot to production.



DEFINITION

Secure and validate. Owns adversarial testing, guardrails, runtime detection, fairness evaluation, AI incident response.

EXAMPLE

A red team running prompt-injection scenarios against the copilot before launch.

DEFINITION

Oversee, assure, decide. Owns policy, decision rights, regulatory alignment, board accountability.

EXAMPLE

An AI Governance Council approving the copilot for high-risk customer interactions and setting escalation thresholds.

9 Governance Surfaces — What Must Be Governed

SURFACE	DEFINITION	EXAMPLE
Model	Foundation models, fine-tuned variants, adapters, routers, model versions	Switching the copilot from Claude 3.7 to Claude 4.7 — a model-surface change requiring change- control evidence.
Prompt	System prompts, templates, agent instructions, prompt libraries	Updating the copilot's system prompt to refuse PII requests — versioned and approved.
Context	Retrieval sources, session state, memory, hidden context, user metadata	A RAG pipeline pulling from SharePoint — context provenance and retention rules apply.
Tools	APIs, plugins, actions, code execution, MCP capabilities	Granting the copilot a refund.process() tool — tool register entry, trust tier, audit logging required.
Orchestration	Planners, workflow graphs, retry logic, multi-step flows, agent routing	An agent that retries up to 3 times then escalates to a human — retry budget defined here.
Identity	Credentials, service accounts, delegated authority, agent identity, secrets	The agent runs under its own scoped service account, not a user's personal credentials.
Safety Layer	Guardrails, policy engines, semantic filters, classifiers, circuit breakers	A classifier that blocks the copilot from generating regulated financial advice.
Telemetry	Logs, traces, evaluations, replay data, fairness metrics	Every prompt, retrieved context, and tool call captured for forensic replay.
Learning Loop	Pre-training sources, post-training alignment, feedback loops, RLHF data, retraining inputs	Customer thumbs-down feedback feeds retraining data — provenance and consent tracked.

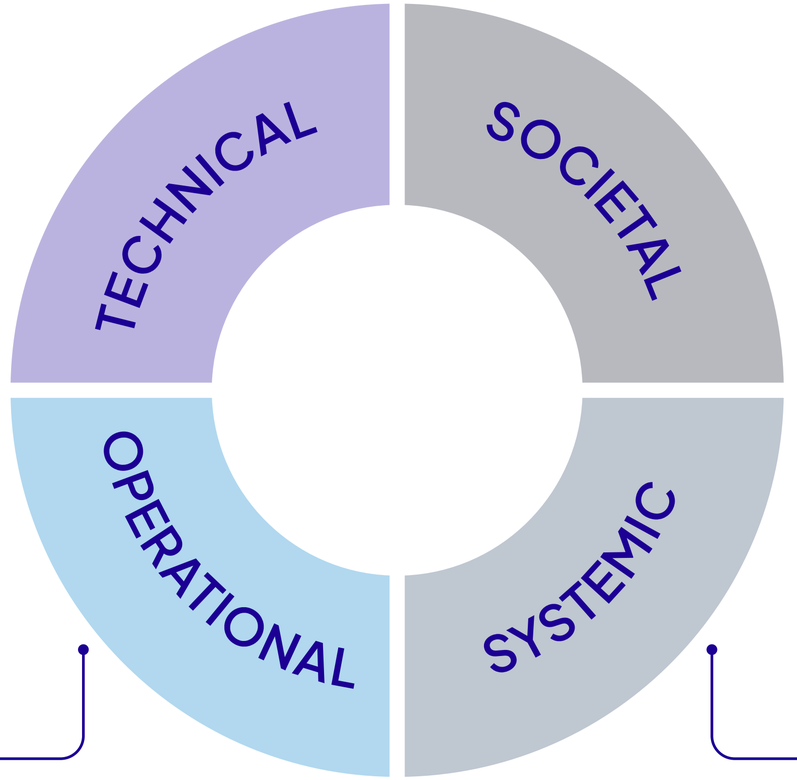
4 Harm Classes — What Can Go Wrong

DEFINITION
Security, reliability, and system integrity failures

EXAMPLE
Prompt injection causing the copilot to leak its system prompt or run unintended tool calls.

DEFINITION
Bias, discrimination, fairness, and human rights impacts

EXAMPLE
A loan-decision agent denying credit at a higher rate to one demographic group.



DEFINITION
Reliability, accuracy, and business-impact failures

EXAMPLE
The copilot hallucinating a refund policy that does not exist, leading to disputes.

DEFINITION
Emergent risks from AI-to-AI interaction and scaled deployment

EXAMPLE
Two negotiation agents from different vendors colluding on an unfavorable deal because their guardrails do not catch joint behavior.

Appendix A: Definitions

Canonical definitions for terms used throughout the framework.

CLASS	DEFINITION
AI System	A machine-based system that, for explicit or implicit objectives, infers from the input it receives how to generate outputs such as predictions, content, recommendations, or decisions (aligned with OECD/EU AI Act).
AI Agent	An AI system that can autonomously plan, execute multistep tasks, invoke tools, and take actions in physical or digital environments on behalf of a user or organization.
Foundation Model	A general-purpose AI model trained on broad data that can be adapted to a wide range of downstream tasks (e.g., GPT, Claude, Gemini, Llama, Stable Diffusion).
Model	Any machine learning artifact, including LLMs, diffusion models, classifiers, regressors, and reinforcement learning agents, used within an AI system.
High-Risk AI System	An AI system whose failure or misuse could cause significant harm to health, safety, fundamental rights, or critical infrastructure (aligned with EU AI Act Article 6).
Composite AI System	An AI system that orchestrates multiple models (potentially of different architectures) within a single workflow or product.
Agentic AI	AI systems exhibiting autonomous behavior: planning, tool use, multi-step execution, and environmental interaction with limited or delayed human review.
HITL	Human-in-the-Loop: a human reviews and approves every AI output before it takes effect.
HOTL	Human-on-the-Loop: a human monitors AI operations and can intervene, but does not approve each individual output.
HOOTL	Human-out-of-the-Loop: AI operates autonomously with periodic governance review rather than real-time human oversight.

Acknowledgements

ADG was developed with the senior AI, security, and governance leaders who run production AI inside Fortune 500, Fortune Global 500, and Big Four firms across regulated and less-regulated sectors, from design through implementation. Their comments, corrections, and counterpoints are the reason this framework is field-tested rather than merely aspirational.

Framework Leadership

EC-Council Global Services—the team that guided the framework from concept to publication.

Jay Bavisi, Founder and Group President, EC-Council Group

Karthik S, Framework Architect and Lead Author
Practice Head, SecureAI · EC-Council Global Services

Mayank Tandon, Global Outreach and Partner Experience
EC-Council Group

Advisory Board

The practitioners listed alphabetically below reviewed, commented on, and shaped this framework. Each contributed time, scrutiny, and hard-won experience gained from running production AI at enterprise scale.

Adam Spearing
VP of AI GTM EMEA
SERVICENOW

Andrei Son
Head of AI Transformation
AUMOVIO

Anish Mitra
Director
KPMG

Anita Lacea
Head of AI Transformation, Azure Hardware Infrastructure
MICROSOFT

Dinesh Bhogle
Head of AI/ML Platform
BLACK & VEATCH

Dr. Sayed Peerzade
Executive Vice President — Cloud, AI and Government Initiatives
JIO

Edoardo Tealdi
Executive Head of AI Transformation — Business Engagement and Growth Units
NTT DATA, INC.

Kathy Baxter
VP / Principal Architect, Responsible AI and Tech
SALESFORCE

Lewis V. Adams
Vice President, Enterprise AI and Capital Productivity Transformation
CITI

Lily Rachmawati
Director, Head of Applied AI
BNP PARIBAS

Malik Hussain
AI Enablement Lead, Data and AI Academy
BASF

Mark Ritcey
Vice President, AI and Automation Delivery
LATENT BRIDGE

Advisory Board

Naveen Upadhyay
Vice President, AI/ML Product Management — Machine Learning and Intelligence Operations
JPMORGAN CHASE & CO.

Oscar Jarabo
Global Head of AI Product and Strategy
TKE

Pavan Kristipati
Head of AI Engineering and Transformation — Enterprise AI Adoption, Governance and Platform
HUNTINGTON BANK

Raghunandan Mishra
AI and Data Engineering Leader
INDEPENDENT PRACTITIONER

Sophia Katrenko
VP of AI/ML
ECOVADIS

Raji Bhimireddy
Vice President Cloud, AI, Architecture, FinOps and Business Value
PRUDENTIAL

ShanShan Pa
Global Head of AI and Data Governance
GLOBALLOGIC

Sruthi Pakanati
Head of AI and Data Transformation, National Quality and Risk
DELOITTE AUSTRALIA

Sudarson Roy Pratihar
Founder and Principal
A2IQ

Yashwinder Chhikara
Senior Vice President — AI, Analytics, and Product Management
ISON XPERIENCES

Sanjoy K. Saha
Head of AI Portfolio and Governance and Chief of Staff CDAO
GE HEALTHCARE

A NOTE ON CONTRIBUTION: The Advisory Board's input shaped eight structural enhancement clusters in this v2 release: an expanded harm taxonomy with Responsible AI integration; coverage beyond LLMs (diffusion, multimodal, and composite systems); a shared responsibility model for vendor/SaaS AI; deeper strategic treatment of the GOVERN pillar; a measurable metrics and evidence framework; multi-agent interoperability governance; explicit regulatory mapping (EU AI Act, NIST AI RMF, ISO/IEC 42001); and post-deployment continuous governance. ADG is a living framework. Future versions will continue to evolve through ongoing practitioner review.

Executive Conclusion

ADG turns the original ADG concept into a full enterprise AI security and responsible AI governance framework. It keeps the simplicity of **ADOPT. DEFEND. GOVERN.**, but adds the structure required to govern contemporary AI systems, across people, process, technology, deployment pattern, life cycle, harm class, and regulatory environment.

IT addresses eight major gap clusters identified through systematic review, and extends the framework to cover multi-agent systems, multimodal architectures, shared responsibility, responsible AI, and measurable governance.

From a framework standpoint, this version is suitable as the basis for:

01

Consulting Assessments

Maturity reviews and AI governance gap analysis.

02

Enterprise AI Governance Programs

Operationalize NIST AI RMF and ISO/IEC 42001.

03

Certification Architecture

Training design across all three ADG pillars.

04

Board-Facing Discussions

AI security, responsible AI, and assurance reporting.

05

Regulatory Preparation

EU AI Act conformity and compliance gap analysis.

06

Vendor Due Diligence

Procurement governance and shared responsibility.

A LIVING FRAMEWORK

ADG is not a static document. It is a living framework designed to evolve as AI deployment patterns change and as new practitioners contribute to its development. Future iterations will be informed by emerging regulatory guidance, new agentic architectures, multimodal advances, and the continuing input of the practitioner advisory board.

EC-Council

Building A Culture Of Security

EC-Council

Global Services