Canonical Reference: AI System Design & Governance

Audience: Enterprise leaders, architects, risk & compliance teams, public-sector policymakers, program owners, system designers

Purpose: This document defines a governance-correct, system-engineering-correct sequence for designing, deploying, and operating any AI-enabled system. It is intended to be used as a **reference standard**, not a vision statement.

The sequence applies equally to: - Enterprise AI products and internal platforms - Public-sector decision-support or service systems - Regulated industry deployments (health, finance, justice, infrastructure) - Hybrid systems combining traditional software with ML / AI components

This framework treats AI as part of a **socio-technical system**, not as a standalone artifact.

Foundational Principle

> **AI systems remain legitimate only when human authority, ownership, and accountability remain explicit, inspectable, and enforceable.**

AI may assist, accelerate, or inform decisions — but it must never become the unowned source of consequential power.

PRE-PHASE: CONSTITUTIONAL GROUNDING

(Establish what may be built before design begins)

0.1 Define the Object of Governance

Real-world meaning

Before any design or procurement activity, the organisation must explicitly define what it considers to be an **AI system** within scope.

Governance meaning

This definition establishes: - regulatory scope - accountability boundaries - enforcement applicability

Without this definition, governance cannot be applied consistently.

System definition

AI system: A socio-technical system in which learned, probabilistic, or non-deterministic computational components materially influence outcomes, decisions, or behaviours affecting people, organisations, or environments.

AI component: Any learned, probabilistic, or adaptive computational element that influences interpretation, evaluation, prioritisation, or decision-making.

This definition applies regardless of: - vendor - model type - deployment environment - scale

---

0.2 AI Component Taxonomy (Functional)

Purpose

Different AI components carry different levels of authority, risk, and governance requirement. Classification is by **function**, not model type.

Component classes

1. **Representational components**
   Examples: embeddings, clustering, feature extraction
   Role: represent information

2. **Interpretive components**
   Examples: LLMs, summarisation, translation
   Role: generate meaning or interpretation

3. **Evaluative components**
   Examples: scoring, ranking, classification, risk prediction
   Role: compare, prioritise, assess

4. **Decisional components**
   Examples: automated approvals, allocations, control policies
   Role: select or execute outcomes

5. **Adaptive / self-modifying components**
   Examples: reinforcement learning, online learning
   Role: change system behaviour over time

Governance obligations increase as systems move from representational toward decisional and adaptive roles.

PHASE 1: PROBLEM RECOGNITION

(Clarify why the system exists)

1.1 Problem Definition

Real-world meaning

Define the underlying organisational, service, or societal problem **independent of AI**.

Governance meaning

Prevents solution-first design and technology justification bias.

System requirement

The problem statement must remain valid if AI components are removed.

1.2 Stakeholder & Impact Mapping

Real-world meaning

Identify: - who uses the system - who is affected by its outputs - who bears risk - who cannot realistically opt out

Governance meaning

Reveals power asymmetries and consent limitations.

1.3 Unacceptable Signals (Early Warnings)

Real-world meaning

Observable indicators that the system may be drifting or harming trust.

Examples include: - confusion or loss of understanding - moral discomfort among operators - over-reliance on system outputs - withdrawal or disengagement - escalation or public backlash

Governance meaning

Signals justify review and intervention before harm becomes entrenched.

Signals are **diagnostic**, not evidence of failure by themselves.

PHASE 1.5: FAILURE & POWER BOUNDARIES

(The constitutional core)

1.5.1 Unacceptable States

Definition

Conditions under which the system is no longer legitimate **regardless of intent or performance**.

Examples: - outcomes cannot be clearly owned by a human or institution - decisions cannot be meaningfully explained, challenged, or appealed - humans cannot realistically override system influence - harm accumulates silently without accountability - stated values collapse under scale or pressure

Governance meaning

These states define revocation conditions.

1.5.2 Authority & Ownership Model

Requirements

For a system to remain legitimate: - outcome ownership must be explicit - decision authority must be human or institutionally grounded - override and shutdown authority must be defined

AI systems may assist but must not become final authorities over consequential outcomes.

1.5.3 Enforcement Posture

Purpose

Establish that boundaries are enforceable in principle.

Governance meaning

This phase does not mandate enforcement mechanisms, but it must confirm that: - alignment claims can be withdrawn - systems can be degraded, halted, or withdrawn

---

1.5.4 Exclusion Criteria

Definition

Explicit identification of system characteristics that are incompatible with legitimate deployment.

Examples: - unappealable AI authority - responsibility laundering to systems or vendors - forced dependency without exit - opaque automation of consequential decisions

---

PHASE 2: VALUE & INTENT ARTICULATION

(Define positive goals within fixed boundaries)

2.1 Values Definition

Purpose

Articulate the outcomes the system is intended to advance.

Values guide behaviour **within** constitutional limits; they do not override them.

2.2 Functional Role Assignment

Explicitly define whether AI components are: - advisory - evaluative - decisional

Role creep must be prevented by design.

2.3 Human Control Guarantees

Requirements

- meaningful human override
- non-AI fallback modes
- explainability appropriate to component role

PHASE 3: SYSTEM DESIGN & ARCHITECTURE

(Translate governance into structure)

3.1 Architecture Principles

Recommended patterns: - AI downstream rather than upstream - optional AI components - separation of reasoning and execution - stateless or bounded-state AI where possible

---

3.2 Data & Feedback Design

Control: - feedback loops - retraining triggers - drift pathways

Prevent silent adaptation.

PHASE 4: IMPLEMENTATION & TESTING

4.1 Scenario & Stress Testing

Test behaviour under: - scale - error - misuse - adversarial conditions

---

4.2 Failure Mode Testing

Explicitly test: - silent failure - partial failure - override failure

---

PHASE 5: DEPLOYMENT & OPERATION

5.1 Monitoring & Review

Track: - dependency signals - override frequency - trust indicators

---

5.2 Revocation & Decommissioning

Ensure the system can be: - modified - rolled back - withdrawn

---

Closing Statement

> **AI governance is not something applied after deployment. It is a structural property of how systems are conceived, built, and operated.**

This reference establishes a repeatable, auditable foundation for responsible AI systems across enterprise and public-sector contexts.