THE AI INSTITUTE / OPERATING INTELLIGENCE

BOARD PAPER · 03 / 2026

Delegation by Design

The board architecture for agentic AI: authority budgets, control evidence and accountability before autonomy scales.

Boards, CEOs, CIOs, CISOs, risk and governance leaders
Research period: January 2024 – July 2026
First published: 5 August 2026
Evidence current to 31 July 2026

PUBLICATION RECORD

An Institute thesis, supported by evidence

Agentic AI changes the central risk question from ‘Can the model produce a bad answer?’ to ‘What is the system authorised to do when it is wrong?’ Autonomy should be earned through evidence and granted incrementally.

WHAT LEADERS SHOULD TAKE FROM THIS PAPER

01	**Define six delegable rights: read, write, communicate, decide, spend and execute.**
02	**Give every agentic use case an authority budget, approval boundary and maximum blast radius.**
03	**Operate governance as a control plane from discovery through retirement.**
04	**Require identity, audit, rollback, evaluation and incident evidence before authority expands.**

Research method

This board paper synthesises internationally relevant risk, identity, incident and management-system work from NIST, OECD and ISO with current deployment evidence. Australian government and cyber-security controls are used as a concrete comparative implementation lens, not as the assumed jurisdiction of every reader. The six-rights and authority-budget frameworks are Institute analytical models intended to make governance decisions explicit.

First published 5 August 2026 · Updated 7 August 2026 · Research period January 2024 – July 2026 · Version 1.2 · Suggested citation: The AI Institute, *Delegation by Design* (2026).

CONTENTS

Inside this paper

01 **Capability is not authority**
Executive brief

02 **Six rights that make delegation visible**
Institute framework

03 **Set an authority budget and blast radius**
Authority architecture

04 **Build an AI control plane, not another policy**
Operating governance

05 **Connected systems turn language into action**
Failure modes

06 **Seven decisions before autonomy scales**
Board decisions

07 **Every material agent needs an assurance pack**
Operating evidence

08 **Not every agent requires the same control intensity**
Counter-case

09 **Establish the delegation baseline**
90-day agenda

M **Method and limitations**

R **References**

How to read the evidence

External research supports the Institute's thesis; it does not substitute for it. Figures retain their original populations and definitions. Institute frameworks and inferences are labelled, and limitations are recorded before the references.

01 EXECUTIVE BRIEF

Capability is not authority

A conversational system proposes. An agentic system can act: retrieve records, update systems, communicate externally, make recommendations, initiate transactions or chain tools across a workflow. The board question is therefore no longer limited to whether an output is accurate. It is whether the organisation has intentionally granted the system the right to act, within boundaries proportionate to the consequence of failure.

Scaled agent deployment remains early. Stanford's 2026 evidence synthesis reports that most business functions still show no agent use and scaled deployment remains in single digits across almost every function. This is a governance window. Organisations can establish identity, authority, evaluation and incident architecture before informal delegation hardens into infrastructure. [1]

The Institute proposes delegation by design: capability does not confer authority; authority is decomposed into specific rights; each right is bounded by context, value, time and consequence; and additional autonomy is earned through operating evidence. The goal is not to prevent agents from acting. It is to make speed safe enough to use.

> **Institute thesis — Never grant an AI system more authority than the organisation can observe, interrupt and recover.**

02

INSTITUTE FRAMEWORK

Six rights that make delegation visible

Agentic authority becomes manageable when it is decomposed. Read is permission to retrieve defined data. Write is permission to create or alter records. Communicate is permission to represent the organisation to a person or external system. Decide is permission to select among consequential options. Spend is permission to commit money or commercial terms. Execute is permission to trigger an operational or technical action.

These rights are not a maturity ladder. A low-risk agent may read and write extensively inside a sandbox but never communicate externally. Another may communicate approved status updates but cannot alter the system of record. A procurement agent might prepare a decision while spend authority remains human. The design objective is least authority for the intended outcome.

Every right needs a scope: the identity under which the agent acts; the data, system and transaction class; the maximum value or volume; the permitted time window; the required approval; the evidence retained; and the recovery action. NIST's 2026 work on software-agent identity and authority highlights why conventional identity and access controls must be adapted when software can act with increasing independence. [2]

EXHIBIT / THE SIX DELEGABLE RIGHTS

R **Read** Retrieve defined information from approved sources.	**W** **Write** Create, amend or delete records within a bounded system.	**C** **Communicate** Represent the organisation to people or external systems.
D **Decide** Select among options with operational or human consequence.	**$** **Spend** Commit funds, pricing, credit or contractual terms.	**X** **Execute** Trigger technical, physical or operational action.

03 AUTHORITY ARCHITECTURE

Set an authority budget and blast radius

An authority budget states how much consequential action a system may take before human approval or automatic suspension. It can be expressed through transaction value, customer count, data sensitivity, external reach, irreversibility, cumulative action or time. The budget should shrink as uncertainty, privilege and consequence rise.

The maximum blast radius is the plausible harm before detection and containment. It is shaped by action speed, tool connectivity, permission breadth, monitoring delay and recoverability. A single incorrect draft has a small radius. An agent with access to a customer database, outbound messaging and refund authority can create compounding harm in minutes.

Controls must therefore work at machine speed. Rate and value limits, allow-listed tools, scoped credentials, separation of duties, approval checkpoints, anomaly detection and circuit breakers are more reliable than a policy instruction inside a prompt. Human review should be placed where consequence or ambiguity is highest, not indiscriminately added after every step.

Australia's Information Security Manual now calls for human approval of organisationally defined risky AI actions and monitoring against behavioural and performance baselines. NIST's GenAI profile similarly emphasises threat modelling, testing, monitoring, acceptable-use boundaries and incident response. [34]

EXHIBIT / AUTHORITY SHOULD CONTRACT AS CONSEQUENCE RISES

04 OPERATING GOVERNANCE

Build an AI control plane, not another policy

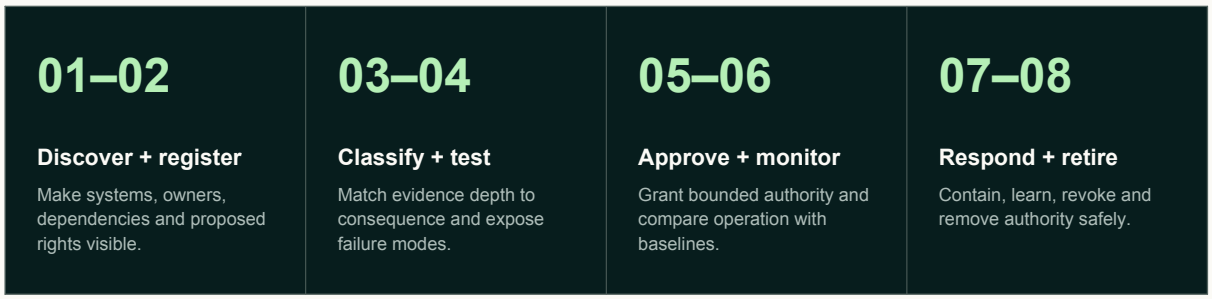
The control plane is the operating system that makes AI use visible and governable. It has eight functions: discover, register, classify, test, approve, monitor, respond and retire. The same workflow should cover vendor features, embedded models, employee-built automations and centrally developed agents; otherwise authority migrates to the least visible channel.

Discovery identifies AI-enabled systems and material changes. Registration records purpose, owners, rights, data, dependencies and evidence. Classification determines consequence and control intensity. Testing evaluates task performance, failure modes, security and human interaction in context. Approval grants a defined authority budget, not a permanent licence.

Monitoring compares behaviour, performance and incidents with approved baselines. Response assigns the ability to interrupt, contain, investigate, notify and learn. Retirement revokes credentials, preserves required records and removes dormant integrations. Australia's 2025 guidance and 2026 public-sector policy updates reinforce accountable ownership, impact assessment, monitoring, incident processes and use-case registers as operating practices; ISO/IEC 42001 places the same disciplines inside a continually improving management system. [569]

A board should ask for evidence from this system: the population of material agents, rights granted, exceptions, control performance, incidents, authority expansions and retirements. A policy completion percentage is not evidence that delegation is controlled.

EXHIBIT / THE CONTROL-PLANE LIFECYCLE



05 FAILURE MODES

Connected systems turn language into action

Prompt injection is not only an output-quality problem. Instructions can enter through retrieved documents, websites, messages or tool responses and steer a connected system toward unintended behaviour. The more systems and rights an agent can access, the more a language-layer failure can become an operational event. [4]

Identity ambiguity compounds the problem. Teams need to distinguish the human sponsor, service identity, model, agent configuration and downstream action. Logs must show who authorised the deployment, which identity acted, what context and tools were available, what the system decided, which approvals occurred and what changed as a result.

Evaluation must include adversarial and operational conditions: missing or conflicting data, malicious content, ambiguous requests, tool failure, approval timeout, repeated actions, privilege escalation and recovery. NIST's 2026 synthesis of agent-security submissions reinforces the need to treat identity, authorisation, tool access and connected-system effects as system concerns. A high task-success score is insufficient if rare failures are irreversible or invisible. [8]

Incidents should be treated as a learning system. The OECD's common reporting framework proposes shared criteria for describing AI incidents across sectors and jurisdictions. Internally, near misses are equally valuable because they reveal control weakness before harm crosses a reporting threshold. [7]

06 BOARD DECISIONS

Seven decisions before autonomy scales

First, define which decisions and actions remain non-delegable. Second, approve the rights taxonomy and authority-budget principles. Third, set materiality thresholds for board and executive visibility. Fourth, require a single accountable business owner for every material agentic workflow.

Fifth, approve the minimum evidence for expanding authority: task performance, control effectiveness, monitoring, identity, recovery and incident readiness. Sixth, define who can suspend an agent and under what conditions. Seventh, agree how customers, workers, regulators and partners will be informed when an agent materially represents or affects them.

These are operating-model decisions, not merely technical standards. They determine how responsibility flows when actions are produced by a chain of people, models, tools and systems. Management can design the controls; the board must ensure authority and accountability remain aligned.

07 OPERATING EVIDENCE

Every material agent needs an assurance pack

An assurance pack is the minimum evidence needed to reconstruct why an agent was allowed to act. It records purpose, owner, affected people, service identity, model and tool chain, data access, six rights, authority budget, test evidence, approvals, monitoring, incidents and recovery procedures.

The pack should distinguish capability evidence from control evidence. Task success shows whether the system can complete representative work. Control evidence shows whether it stays inside authority, rejects or escalates unsafe conditions, produces reconstructable logs and can be interrupted and recovered.

Large-scale red-teaming evidence demonstrates why agent evaluation must include malicious instructions embedded in external data. A system may perform well on benign tasks while remaining vulnerable to indirect prompt injection once connected to email, websites, repositories or tools. [10]

The evidence should travel with the configuration. Expanding a tool right, raising a spend limit, adding a data source or changing the model creates a new assurance question. Versioned evidence prevents an old approval from silently governing a materially different system.

EXHIBIT / SIX ARTEFACTS IN THE ASSURANCE PACK

01 **Purpose + owner** Business outcome, affected population and accountable executive.	**02** **Identity + rights** Service identities, access, authority budget and approval boundaries.	**03** **Evaluation** Representative tasks, adverse tests, quality and control evidence.
04 **Operation** Monitoring, logs, exceptions, incidents and human intervention.	**05** **Recovery** Interrupt, revoke, contain, rollback and notification procedures.	**06** **Change record** Versions, material changes, approvals and superseded evidence.

08 COUNTER-CASE

Not every agent requires the same control intensity

A read-only research assistant using public data does not require the same evidence as a system that sends customer communications, changes records or spends money. Applying the maximum control set everywhere can push use into less visible channels and waste scarce assurance capacity.

Classification should combine consequence, reversibility, detectability, reach and novelty. Low-consequence and reversible use can move quickly with logging and clear boundaries. High-consequence, hard-to-detect or irreversible action requires stronger testing, human approval and recovery evidence.

NIST's preliminary Cyber AI Profile and the DTA assurance pilot both reinforce context-based application rather than a single checklist. The public-sector pilot surfaced fairness, transparency and explainability issues not fully covered by existing arrangements, illustrating the value of structured impact work without proving one universal control design. [11][12]

Proportionality is not permissiveness. It is the discipline of matching evidence and authority to consequence, then increasing control as rights or blast radius expand.

09 90-DAY AGENDA

Establish the delegation baseline

In the first 30 days, identify material agentic and tool-connected AI use. Map the six rights, service identities, data access, approvals and plausible blast radius. Suspend or narrow any use case whose authority cannot be reconstructed.

By day 60, define authority budgets, classification criteria and minimum evidence. Test one material workflow under adverse conditions, including indirect prompt injection, tool failure, repeated action, approval failure and emergency suspension. Confirm that logs support forensic reconstruction.

By day 90, bring the board a delegation register: material systems, owners, rights, authority limits, evidence status, incidents and unresolved exceptions. Expand authority for one use case that meets the standard, constrain one that does not, and exercise the organisation’s ability to interrupt and recover.

RESEARCH RECORD

Method and limitations

Method

This board paper synthesises internationally relevant risk, identity, incident and management-system work from NIST, OECD and ISO with current deployment evidence. Australian government and cyber-security controls are used as a concrete comparative implementation lens, not as the assumed jurisdiction of every reader. The six-rights and authority-budget frameworks are Institute analytical models intended to make governance decisions explicit.

Limitations

Agentic terminology and product architecture are unsettled. Standards and guidance are evolving, and empirical incident data remain limited. Control requirements must be adapted to sector obligations, technical architecture, affected people and the consequence of action; this paper is not legal or cyber-security advice.

Evidence conventions

Source facts are cited to the original publication. Institute analysis reconciles definitions or combines evidence. Institute inference extends that analysis into a management proposition. Frameworks are decision aids, not claims of universal causality. Every important statistic should be read with its population, date, definition and caveat.

Research cutoff: 31 July 2026. Review cycle: six months, or earlier after a material change in evidence, standards or policy.

REFERENCES

Evidence behind the thesis

01 **Stanford HAI, 2026 AI Index — Economy**

Current synthesis of organisational and agent deployment evidence.

02 **NIST, Identity and Authority for Software Agents**

Concept paper on adapting identity and access control to software agents.

03 **Australian Signals Directorate, Information Security Manual — Guidelines for System Hardening**

Current Australian cyber-security guidance for risky AI actions and monitoring.

04 **NIST AI 600-1, Artificial Intelligence Risk Management Framework: Generative AI Profile**

Voluntary risk-management guidance, including prompt injection and connected-system risks.

05 **National AI Centre, Guidance for AI Adoption**

Australian implementation practices for accountable and controlled AI adoption.

06 **Digital Transformation Agency, AI policy update**

Updated public-sector requirements through 2026.

07 **OECD, Towards a Common Reporting Framework for AI Incidents**

Cross-sector incident reporting criteria.

08 **NIST, Security Considerations for Artificial Intelligence Agents**

2026 synthesis of responses on agent security threats.

09 **ISO, ISO/IEC 42001 AI management systems**

Management-system standard using continual improvement.

10 **NIST, Insights into AI Agent Security from a Large-Scale Red-Teaming Competition**

2026 evidence on indirect prompt injection and agent hijacking under adversarial testing.

11 **NIST IR 8596, Cybersecurity Framework Profile for Artificial Intelligence**

Preliminary profile organising cyber risk across securing AI, AI-enabled defence and AI-enabled attacks.

12 **Digital Transformation Agency, Responsible AI Assurance Pilot Findings**

Australian 21-agency pilot evidence on structured impact assessment and assurance gaps.

THE AI INSTITUTE

Evidence should change what you do next.

Research · Readiness · Capability

theai.institute