



From Prompt Injection to Policy

Building a practical AI control framework for
RAG, copilots and agents

David Vohradsky
Founder & CEO, MyRISK

1

One fine day

A bounded, reasonable request with an invalid assumption inside it

1

Large organisation, federated IT

Tens of thousands of users, extended R&D ecosystem, decentralised estate

2

A mature standards set

Risk, data security, IAM, asset management, secure by design, vulnerability, logging, incident, vendor

3

AI arrived through products already bought

Tenanted copilot, RAG assistants in business and IT, staff-built GPTs, business users trying agents

**“We are adopting AI.
Can you write us an
AI security policy?”**

That sentence contains an invalid assumption: that AI security is a discipline that can be written down separately, and that the writing is the work.

2

The abundance problem

Six authorities. Six different kinds of object. None of them a control catalogue.

MITRE ATLAS

Adversary techniques and mitigations against them

OWASP Top 10 for LLM

Application **vulnerabilities** and prevention strategies

OWASP GenAI Project

A **security lifecycle**, extended to agentic architectures

NIST AI RMF

Governance outcomes: GOVERN, MAP, MEASURE, MANAGE

ASD / ACSC

National guidance on supply chain, frontier models, agents

ISO/IEC DIS 27090

Threats, failures, detection and mitigation, separated

A framework that treats these as six comparable catalogues inherits all six structures at once.

Normalise each source on its own terms before comparing — a requirement against a requirement, never a heading against a heading.

3

Coverage improved. Usability collapsed.

Every new publication could add another ten or twenty controls

The same control, six times

Least privilege for agents appeared as privileged-agent permissions, single-user permissions, tool permissions, backend access control, runtime policy enforcement and zero-trust execution.

The MEASURE sinkhole

Almost every control mapped to NIST MEASURE — blurring implementing a control with evaluating whether it works. MANAGE is the treatment; MEASURE is how you know it still holds.

Consolidation destroyed precision

Short-lived agent credentials, safe model-loading flags, two-person approval, parameterised queries, Content Credentials, hallucination logging — all evaporated into broad wording.

Broad controls hid gaps

“Monitor AI systems” overlaps a great deal of guidance while never saying: repeated probing, unusual query distributions, model-extraction behaviour, mitigation deterioration.

Partial coverage is more dangerous than absent coverage.

Having a standalone AI Security Standard was quietly failing - AI API security is still API security, agent permissions are still access control, AI incidents are still security incidents.

4

Rock bottom: the mapping, not the controls

Six sources, several hundred requirements, three months. That workload gets delegated to a model.

What happened

A requirement that the highest-risk agent actions must require dual control was allocated to IAM and described as a rule about granting access. The rationale read well.

It was wrong. Dual control means two parties approve a sensitive action. Allocated that way it gets the wrong owner, the wrong assessor, the wrong mechanism.

It was not caught in review. Reviewers read the mapping artefact, and the artefact was internally coherent.

Why it fails this way

● Compares compressions

Paraphrase is lossy where it matters: qualifiers, scope, authority

● Plausibility is the target

A wrong mapping arrives with a confident rationale

● Terms of art are the trap

Dual control, SoD, four-eyes, access approval sit adjacent

● Fails consistently, not randomly

Same misreading applied to every similar requirement

Quote, never paraphrase

Stable IDs at extraction

Scrutinise “covered” harder
than “gap”

Re-derive a sample
independently

Use AI to propose the mapping. Do not let it own the record.

5

A threat is not a control

And a mitigation is not necessarily a standalone control

Risk scenario

What can go wrong, as an organisational consequence

Changes when the threat environment changes. Owned by risk. Never a control.

Control objective

The security outcome the organisation requires

The durable layer. This is what belongs in policy.

Control requirements

Specific, testable obligations that deliver the objective

Where technical specificity lives and where assurance bites.

Implementation patterns

How a system satisfies the requirement today

Examples and evidence, never obligations. Differential privacy, adversarial training, watermarking live here.

Most AI framework bloat comes from promoting level four into level two.

6

The boundary shift

“What can the model say?” becomes “What can the system cause to happen?”

RAG / retrieval

Influences the information returned to the user. Untrusted retrieved content is treated as instruction.

INFORMATION CONTROL

Copilot

Influences recommendations and generated work. The human is still the actor.

INFORMATION, CONTEXT, RELIANCE

Tool-enabled agent

Influences an ACTION, executed against real systems under inherited authority.

DELEGATED AUTHORITY AND ACTION CONTROL

Same technique. Three architectures. Three different security problems.

user



agent



model



context /
memory



tool → API



downstream
action

Least privilege for AI is not one control. Architecture determines applicability.

7

Put the determinism at the boundaries

Probabilistic in reasoning. Deterministic where security boundaries and material transactions are enforced.

Code layer**Schema-first validation**

Pydantic, Zod, strict JSON Schema on tool contracts, constrained decoding. Failure is forced to be a rejection, not a misinterpretation.

Tool layer**Mediation outside the model**

Allow-listed registry, explicit scopes, parameter validation, destination restrictions, authorisation checked against the acting identity.

Transaction layer**Assume the reasoning was wrong**

Idempotency keys, replay protection, value limits, approval gates, circuit breakers, rollback or compensation.

Business layer**A deterministic record of a probabilistic decision**

Inputs, retrieved sources, model and prompt version, rationale, approving identity, resulting action — captured at the time.

Let the model propose. Make something else decide, record and constrain.

8

Rebuilding: stronger without getting larger

A new source creates new mappings and requirements — not new controls

Strategy, governance, accountability

Risk and trust

Data and knowledge

Model and solution assurance

Orchestration and integration

Runtime and infrastructure

Cross-cutting control planes

- Observability and assurance
- Lifecycle and supply chain
- Identity — users, workloads, agents
- Policy enforcement points

Map in both directions

A standard is not the outer boundary of good practice

Capability, not product identity

MCP turns a text generator into a system that reaches records

Objective, not taxonomy

Entra Agent ID changed mid-engagement. The objective did not.

9

The return

We never wrote the AI security policy

Identity and Access Management

Agent identity types, accountable owners, privileged classification, phishing-resistant auth, scope and SoD, human approval, access reviews

Logging and Monitoring

Logging extended to AI capabilities, with a dedicated AI events schedule

Secure by Design

Hardened baselines, API scopes and short-lived secrets, AI trust boundaries, untrusted content, grounding, memory and media validation, model development controls

Threat and Vulnerability Management

AI capabilities in scope, scanning before production and monthly, full technology stack

Not an AI security policy beside cyber security. A cyber security framework that has evolved to govern AI.

9

The return

We never wrote the AI security policy

Identity and Access Management

Agent identity types, accountable owners, privileged classification, phishing-resistant auth, scope and SoD, human approval, access reviews

Logging and Monitoring

Logging extended to AI capabilities, with a dedicated AI events schedule

Secure by Design

Hardened baselines, API scopes and short-lived secrets, AI trust boundaries, untrusted content, grounding, memory and media validation, model development controls

Threat and Vulnerability Management

AI capabilities in scope, scanning before production and monthly, full technology stack

For the room

How much of your control mapping was done by a model — and how would you know if it was wrong?

Is a standalone AI security standard ever the right answer?

Where does human approval stop being a control and become a rubber stamp?

Are we writing assurance requirements for things almost nobody can test?

Not an AI security policy beside cyber security. A cyber security framework that has evolved to govern AI.

B1

Ten lessons

1

A threat is not a control

Only objectives and requirements belong in the framework

3

The problem is overlap

You need a stable internal control language

5

Partial coverage is the danger

Absent coverage appears on a register; partial does not

7

Agentic AI moves the boundary

From information control to delegated authority

9

Policy is simpler than the framework

Provided traceability from framework to clause survives

2

Compare requirements, not titles

Four impressive names often express one requirement

4

Assess coverage on outcome

Never ask whether a control with that name exists

6

Verify that the AI is mapping correctly

Model errors are systematic, so sampling under-reports them

8

Reassessment is capability-driven

New tools and autonomy change risk without changing the product

10

Policy is the output, not the start

Know what you govern and by whose authority first

B2

The consolidated control domains

Fifteen policy-level objectives carrying several hundred traceable source requirements

01 Acceptable Use, User Obligations and Training

02 Risk Assessment, Tiering and Approval

03 Data Governance, Lawful Use, Residency and Retention

04 Regulatory, Privacy, Ethics and Fairness

05 Identity, Access and Provisioning

06 Agent Authority, Tool Permissions and High-Risk Actions

07 Supply Chain Integrity, Provenance and Asset Governance

08 Input Protection

09 Model Security and Robustness

10 Output Safety, Grounding and Behaviour

11 Output Handling, Encoding and Leakage Resistance

12 Runtime, Environment and Compute Hardening

13 Logging, Monitoring and Traceability

14 Incident Response and Recovery

15 Third-Party, Contract and Supply Chain Assurance

Each domain carries its objective, its generic risk, an organisation-specific risk scenario in plain language, and full provenance to MITRE ATLAS, OWASP, NIST AI RMF, ASD, ACSC and ISO/IEC DIS 27090.



Contact

David Vohradsky

david@myrisk.io

<https://myrisk.io/insights>

myRISK

CYBERCON 2026
AUSTRALIAN CYBER CONFERENCE

Presented by
AISA

Supported by
VICTORIA
State
Government