CiBRAI

CYBER INTELLIGENCE - BEHAVIOURAL RESPONSE

CiBRAI Research Report | 2026

AI TRUST UNDER ATTACK

The CiBRAI Trust Fabric for Cybersecurity Operating Platforms and SOCs

THE CENTRAL QUESTION

What evidence, authority and recovery conditions should exist before an AI system is permitted to influence or execute a cyber defence action?

Operational assurance for digital defenders that can observe, reason and act at machine speed.

CIBRAI RESEARCH

TRUST IS NOT PERMISSION GRANTED ONCE.

IT IS PERMISSION CONTINUOUSLY EARNED.

A trustworthy AI-enabled SOC is not one in which AI never fails. It is one in which failure is bounded, observable, reversible, attributable and learned from.

PUBLICATION DETAILS

RESEARCH LEAD	Andrew Curtis, Founder and CEO, CiBRAI
RESEARCH BASIS	Practitioner-led control design, architecture analysis, threat modelling and standards synthesis
SCOPE	AI-enabled cybersecurity operating platforms, SOC copilots, autonomous agents and multi-agent workflows
DATE	August 2026 | Sydney, Australia
STATUS	Independent industry research | General information, not legal or regulatory advice

REPORT NAVIGATION

Contents

A practical research report for boards, CISOs, SOC leaders, security architects, AI governance teams and platform builders.

Research basis

This report is a practitioner-led research synthesis. It combines CiBRAI design research in cybersecurity operating platforms and SOC workflows with architecture analysis, threat modelling, human factors research, and current public standards and guidance. It does not claim a statistically representative survey. The propositions are intended to be tested in real environments and adapted to organisational risk appetite.

SCOPE BOUNDARY

The report focuses on AI that influences or executes cyber defence activity. It is not a general ethics guide, a product certification, or a substitute for legal, regulatory or safety advice.

EXECUTIVE SUMMARY

Trust must become an operational control plane

AI is moving from describing cyber events to shaping decisions and taking action. That transition changes trust from a user sentiment into a safety, security and accountability requirement.

CENTRAL THESIS

Trust is not a property of a model. It is an operational state produced by evidence, constraint, accountability, transparency, resilience and recovery. In a SOC, it must be action-specific, time-bounded and continuously renewed.

The industry is at risk of asking the wrong question. Model accuracy, benchmark performance and persuasive explanations matter, but they do not answer whether an AI-enabled platform should be permitted to touch an identity store, isolate an endpoint, block traffic, modify detection logic, close an incident or orchestrate a multi-system response. The object of trust is the complete AI action system: the signal, data provenance, model, prompt and policy, retrieved context, agent logic, identity, tools, human gate, action, outcome and recovery path.

CiBRAI proposes an AI Trust Fabric with eight interdependent dimensions: purpose and authority; identity and provenance; data integrity and confidentiality; model behaviour and uncertainty; orchestration, tools and permissions; human control and accountability; resilience and recoverability; and evidence, monitoring and re-authorisation. Trust behaves like a weakest-link system. A highly capable model does not compensate for poisoned context, excessive privileges, a missing kill switch or an analyst who lacks time to intervene.

The report also introduces a Trust Contract for every use case, a six-level Bounded Autonomy Ladder, an AI Decision Evidence Chain, a lifecycle assurance model, a five-level maturity model and a 90-day implementation path. These mechanisms turn broad AI principles into executable controls for security operations.

What leaders should do now

- Govern the action system, not the model in isolation.
- Define an autonomy ceiling and explicit no-go actions for each environment.
- Require a Trust Contract and evidence chain before privileged or consequential action.
- Treat model, prompt, data, permission and tool changes as potential re-authorisation triggers.
- Measure security value and trust integrity together. Governance activity alone is not an outcome.

CIBRAI RESEARCH PROPOSITIONS

Seven propositions that should change the AI trust debate

The propositions below are intentionally direct. Each is designed to expose weak assumptions in current AI assurance and SOC automation practice.

Figure 1. Seven CiBRAI research propositions. Source: CiBRAI analysis, 2026.

A USEFUL TEST

When a control statement contains the words responsible, explainable, human-centred or safe, ask what evidence proves it during a live incident. If the answer is a policy document alone, the control is not yet operational.

CiBRAI
CYBER INTELLIGENCE - BEHAVIOURAL RESPONSE

01.

TRUST IS NOT A FEELING

It is calibrated reliance under uncertainty.

01 / FOUNDATIONS

Trust, trustworthiness and reliance are different things

Precision matters. Organisations often use trust as a single word for several distinct concepts, which creates weak controls and false assurance.

TRUSTWORTHINESS Objective or evidenced system properties that make reliance more or less justified. Examples include validity, reliability, security, transparency, resilience and accountability. NIST treats trustworthiness as multidimensional rather than a single score. [4]	**TRUST** A human or organisational attitude that the system will help achieve a goal under uncertainty and vulnerability. Trust is influenced by performance, context, experience, interface design and institutional confidence. [1][3]
RELIANCE Observable behaviour. The analyst accepts a recommendation, follows a priority, grants authority or permits an action. Reliance can be appropriate even when trust is cautious, and inappropriate even when confidence feels high. [1][2]	**ASSURANCE** The structured evidence and decision process used to justify reliance. Assurance includes evaluation, threat modelling, testing, controls, monitoring, decision records and re-assessment.
AUTHORITY The scope of action delegated to the AI system, including tools, identities, targets, data, time, thresholds and conditions. Authority is a design variable, not an inevitable by-product of capability.	**ACCOUNTABILITY** The allocation of answerability for intent, configuration, approval, residual risk, monitoring, intervention and consequences. Accountability remains human and organisational even when execution is automated.

CIBRAI POSITION

The goal is not high trust. The goal is justified reliance within an explicit operating boundary.

01 / HUMAN FACTORS

Trust should be calibrated, not maximised

Decades of automation research distinguish appropriate use from misuse, disuse and abuse. AI does not remove this problem. It intensifies it because outputs are fluent, probabilistic and context-sensitive. [1][2][3]

Figure 2. Appropriate reliance tracks demonstrated assurance, context and consequence. Source: CiBRAI analysis, informed by human-automation research [1][2][3].

Overtrust creates automation bias, silent propagation of false assumptions and unsafe action. Undertrust creates duplicate work, shadow processes and missed defensive advantage. Both are control failures. The appropriate level of reliance should vary by use case, asset criticality, time pressure, analyst expertise, confidence, reversibility and the quality of the evidence chain.

Overtrust indicators	Undertrust indicators
Analysts accept unsupported claims; confidence displays replace evidence; recommendations are executed without challenge; repeated success becomes assumed infallibility; vendor branding substitutes for local validation.	Analysts repeat every machine task; approved tools are bypassed; low-risk use cases remain trapped in review; defensive speed is lost; teams rely on ungoverned personal tools instead.

01 / OPERATING CONTEXT

Why AI trust is different in cyber operations

A cybersecurity platform operates in an adversarial, privileged and time-compressed environment. Trust patterns from ordinary productivity software are not sufficient.

01 ADVERSARIAL INPUTS

Attackers can intentionally shape logs, tickets, threat intelligence, documents, URLs and system behaviour to manipulate AI reasoning or action.

02 PRIVILEGED TOOLS

The platform may access EDR, identity, network, cloud, email and case-management controls. A small reasoning error can become a material operational effect.

03 TIME PRESSURE

The value of automation is greatest when analysts have the least time to review. Oversight must be designed for real decision windows, not governance workshops.

04 UNCERTAIN GROUND TRUTH

Attribution, intent and incident scope are rarely certain. A plausible explanation can be operationally dangerous when presented as settled fact.

05 CASCADING EFFECTS

An incorrect identity, endpoint or network action can trigger further alerts, agents and automated responses, amplifying the original error.

06 CHANGING SYSTEMS

Models, prompts, retrieval sources, APIs and security products change continuously. Trust expires faster than in conventional static software.

CONSEQUENCE

A more accurate model can be less trustworthy than a weaker one if it has broader authority, poorer observability and no reliable recovery path.

01 / SOC DESIGN

AI authority changes the assurance class

Cybersecurity teams often discuss AI as a single capability. In practice, an AI system that summarises an alert and one that revokes a privileged session belong to different assurance classes.

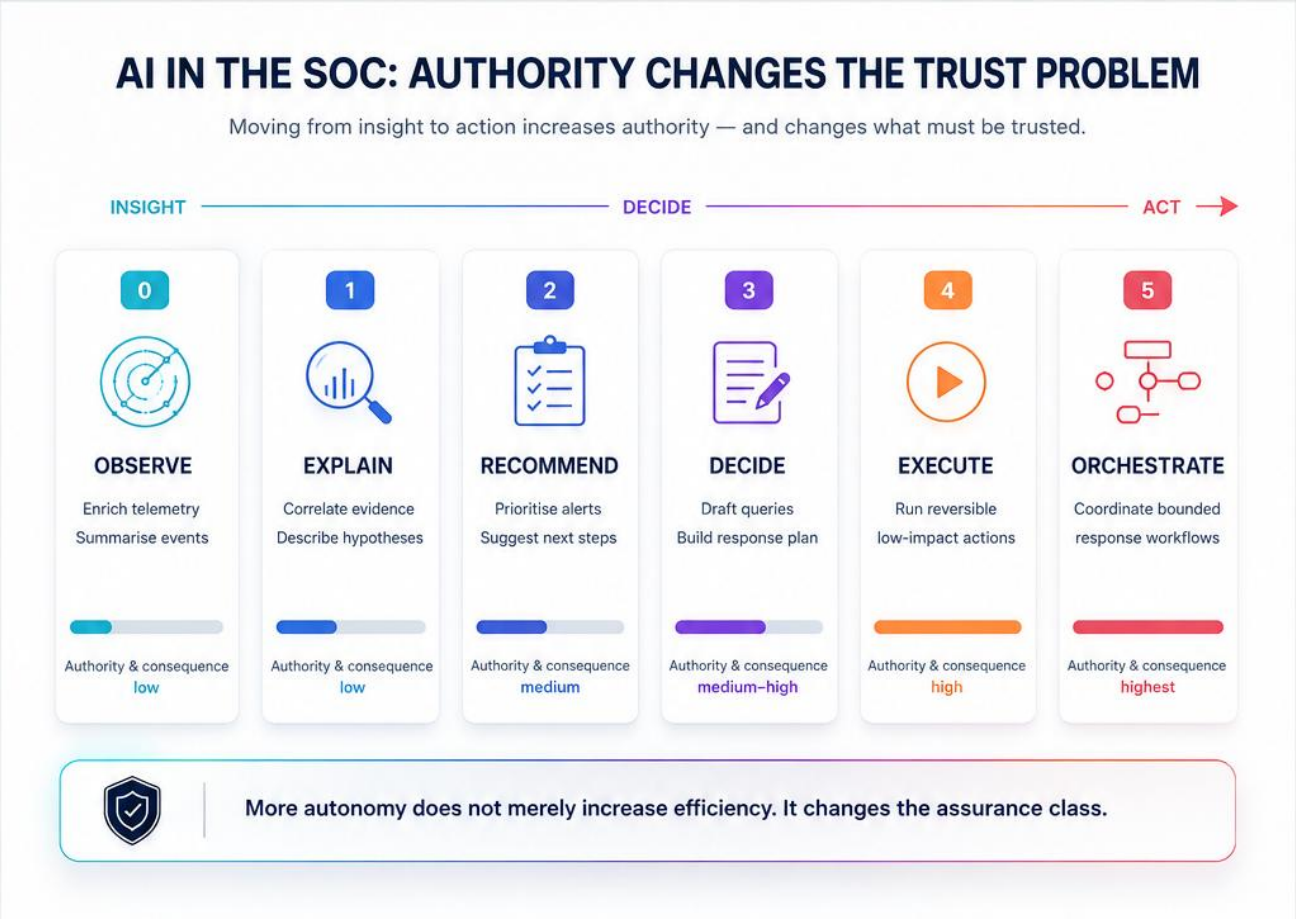


Figure 3. The AI use spectrum in security operations. Source: CiBRAI analysis, 2026.

The progression from observation to orchestration is not a smooth feature upgrade. Each step changes who can be harmed, what can fail, how quickly an error propagates and what evidence is required. Australian cyber guidance similarly emphasises that AI should augment established security tools and processes, with governance, human oversight and secure-by-design controls retained. [12]

OPERATING RULE

Capability may be broad. Authority must be narrow. An agent should receive only the identity, data, tools, targets and time window needed for the current task.

CiBRAI
CYBER INTELLIGENCE - BEHAVIOURAL RESPONSE

02.

THE OBJECT OF TRUST IS THE ACTION SYSTEM

Model, context, orchestration, tools, people and evidence.

02 / ARCHITECTURE

The model is not the system

The operational risk of AI in a SOC is produced by the complete action chain. The orchestration and tool layers can matter more than the base model because they determine context, permissions and effects.

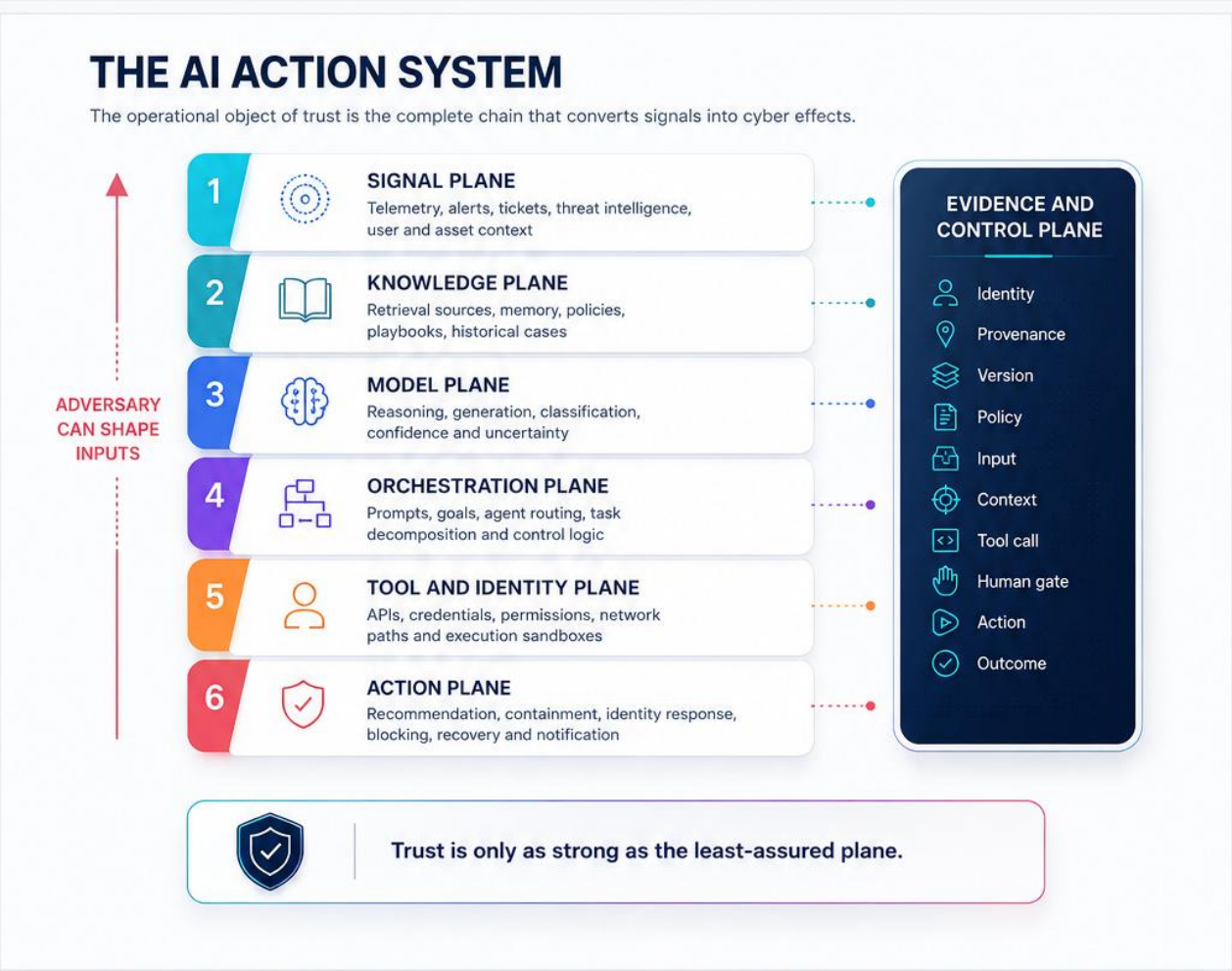


Figure 4. The AI Action System. Source: CiBRAI analysis, 2026.

A well-aligned model inside a poorly designed orchestration layer can still produce unsafe outcomes. Conversely, a probabilistic model can be useful when its inputs are controlled, authority is tightly bounded, uncertainty is exposed, actions are reversible and the entire chain is observable. This is why a model card is not an operating licence.

ARCHITECTURE PRINCIPLE

Separate the signal, knowledge, reasoning, orchestration, identity/tool and action planes. Apply independent controls at each plane and preserve a parallel evidence and control plane across all of them.

02 / ACCOUNTABILITY

Shared responsibility must be explicit

Modern AI systems are assembled across model providers, platforms, integration partners, operators and business owners. Evidence requests should be directed to the party that can genuinely provide them, while use-case accountability remains local.

Figure 5. Shared responsibility and evidence ownership. Source: CiBRAI analysis, 2026.

The model provider can disclose model-level limitations and evaluation evidence. The platform provider can evidence guardrails, tenancy, access control and auditability. The implementation partner can evidence integration and secure configuration. The SOC operator must own data, procedures, prompts, monitoring and response. The business or risk owner must define acceptable use, delegated authority and residual risk. None of these roles can substitute for another.

FAILURE PATTERN

When no party owns the orchestration layer, everyone can produce documents while no one can prove that the deployed action system is safe.

02 / FAILURE ANALYSIS

Four failure patterns that ordinary AI governance misses

The most damaging failures in cyber operations often occur between governance domains, not inside a single model evaluation.

THE POISONED TICKET

A malicious string embedded in an incident ticket instructs an agent to retrieve credentials and ignore policy. The model behaves as designed, but the content boundary fails.

Required response: isolate untrusted content, separate instructions from data, restrict retrieval and tool access, and monitor policy deviation.

THE CONFIDENT FALSE POSITIVE

An agent correlates incomplete identity evidence and recommends disabling a privileged account during a critical service event. The output is fluent and internally coherent.

Required response: expose missing evidence, require corroboration for consequential actions, apply asset-aware thresholds and preserve human veto.

THE QUIET DRIFT

A model or platform update changes severity calibration and tool selection. The service remains available, but operational behaviour moves outside the approved envelope.

Required response: version awareness, change notice, regression scenarios, behavioural baselines and time-bound re-authorisation.

THE MISSING WHY

The SOC can see that an endpoint was isolated but cannot reconstruct the prompt, retrieved context, policy, tool call, human intervention or model version.

Required response: immutable evidence chain, correlation identifiers, action receipts and tested reconstruction procedures.

RESEARCH IMPLICATION

AI assurance must test how the system fails under hostile and ambiguous conditions, not only whether it performs an intended task under clean conditions.

02 / CHANGE

Trust decays and trust debt accumulates

AI systems change at several layers. Assurance that was valid at approval can become stale while the user interface and service availability appear unchanged.

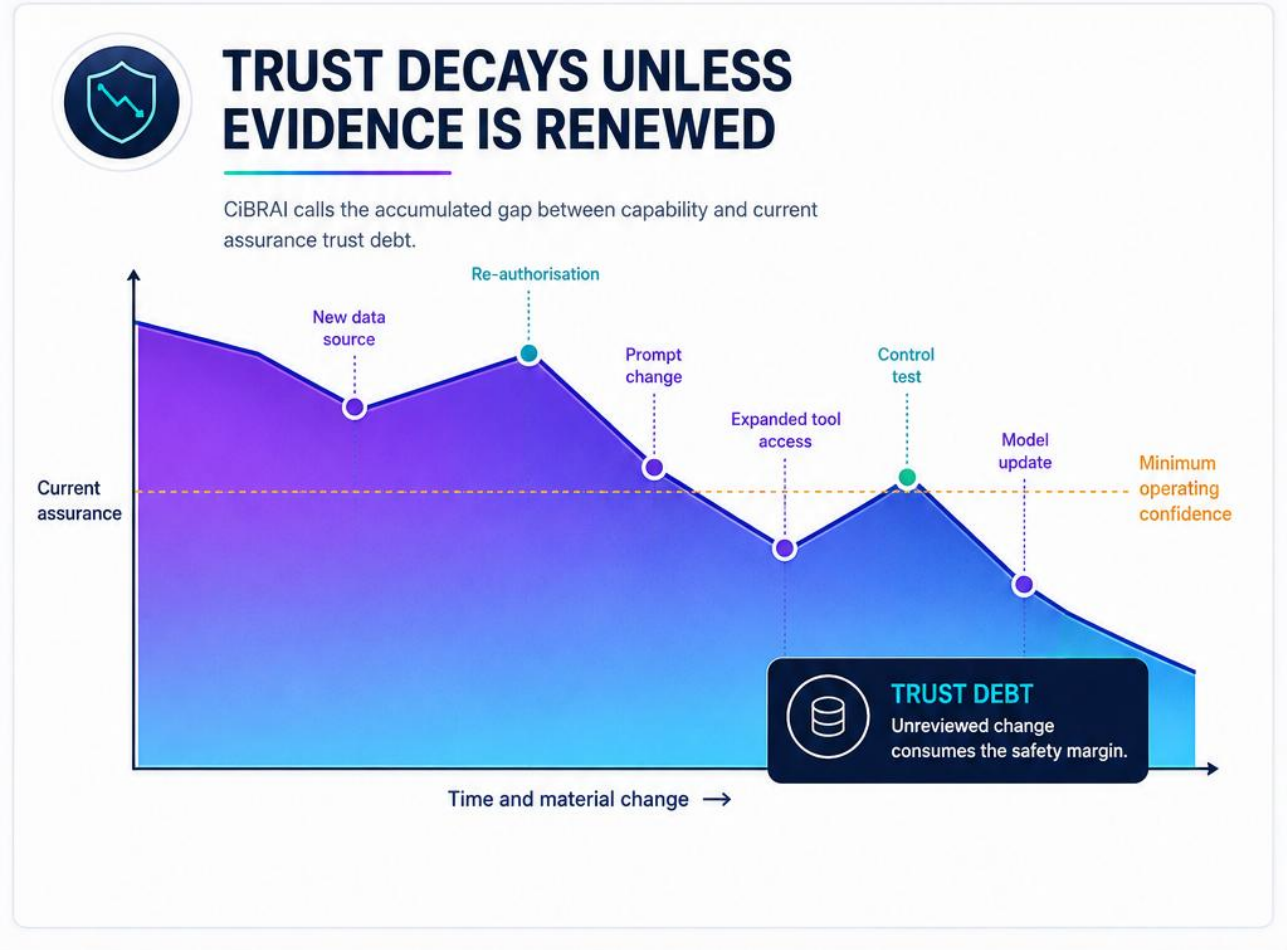


Figure 6. Trust decay and trust debt. Source: CiBRAI analysis, 2026.

Trust debt is the accumulated gap between current capability or authority and current evidence. It grows when models are updated without regression testing, prompts and policies change without review, retrieval sources expand, tool permissions widen, exceptions persist, or monitoring is disabled. It can also grow when the threat environment changes faster than the control design.

Material change triggers

- A model, provider, orchestration framework or major security product version changes.
- New data classes, sources, memory stores or external content are introduced.
- The agent receives new tools, identities, privileges, targets or network access.
- The use case expands to new users, assets, geographies or higher-consequence decisions.
- Observed behaviour, confidence calibration, error rates or adversarial outcomes move outside tolerance.

OPERATING PRINCIPLE

Trust should expire unless renewed by evidence. Re-authorisation can be light for low-risk changes and deep for changes that increase consequence, privilege or autonomy.

CiBRAI
CYBER INTELLIGENCE - BEHAVIOURAL RESPONSE

03.

THE CIBRAI AI TRUST FABRIC

Eight dimensions. One operating licence.

03 / FRAMEWORK

The CiBRAI AI Trust Fabric

The Trust Fabric converts abstract principles into eight interdependent dimensions that can be designed, tested, monitored and re-authorised.

Figure 7. The CiBRAI AI Trust Fabric. Source: CiBRAI analysis, 2026.

The framework is deliberately system-level. It covers organisational intent, technical enforcement, human supervision and operational evidence. It is designed to complement established frameworks such as NIST AI RMF, ISO/IEC 42001, ISO/IEC 23894, Australian guidance, secure AI lifecycle guidance, AI cybersecurity standards and adversarial threat catalogues. [4][7][8][10][12][14][18][19][20]

DESIGN INTENT

The Trust Fabric is not a certification mark or a single maturity score. It is a control architecture for deciding what an AI action system may do, under what evidence and for how long.

03 / DECISION LOGIC

Trust behaves like a weakest-link system

Average scores can conceal catastrophic weaknesses. A single missing control can invalidate an otherwise impressive system.

Figure 8. CiBRAI weakest-link trust and assurance-load heuristics. Source: CiBRAI analysis, 2026.

The weakest-link heuristic means that a system cannot compensate for an unbounded identity by improving explainability, or for poisoned telemetry by adding a policy statement. The assurance-load heuristic directs effort to the use cases that create the greatest operational consequence. It is intentionally qualitative. The purpose is to make consequence, privilege, autonomy, exposure and irreversibility visible before teams debate documentation volume.

COUNTERINTUITIVE RESULT

A high-consequence use case with excellent observability, tight containment and tested rollback may be safer than a medium-consequence use case with opaque actions and no recovery path.

03 / TRUST DIMENSIONS

Dimensions 1 to 4: intent, identity, data and behaviour

The first four dimensions establish what the system is for, who or what is acting, what information can be trusted and how model behaviour is bounded.

1. PURPOSE & AUTHORITY

Define the use case, intended value, affected stakeholders, no-go decisions, autonomy ceiling and residual risk owner. Authority must be explicit, time-bounded and linked to consequence.

Minimum evidence: approved use case, risk tier, action boundary, owner, delegated authority, review date.

2. IDENTITY & PROVENANCE

Assign identities to models, agents, services, tools and human approvers. Preserve source, time, version and custody for inputs, context, outputs and actions.

Minimum evidence: agent/service identities, correlation IDs, source integrity, model and policy version, signed action receipt.

3. DATA INTEGRITY & CONFIDENTIALITY

Control telemetry, threat intelligence, retrieval data, memory and prompts. Address classification, poisoning, leakage, quality, retention and cross-tenant exposure.

Minimum evidence: source allowlist, data classification, integrity checks, retrieval boundary, retention and incident process.

4. MODEL BEHAVIOUR & UNCERTAINTY

Evaluate the model in the actual use context. Test capability, calibration, failure modes, adversarial behaviour, unsupported assertions and known limitations.

Minimum evidence: scenario results, thresholds, uncertainty display, limitations, regression baseline and fallback model or process.

TEST QUESTION

Would the organisation still approve the use case if the same capability were delivered by an unknown contractor with the same data and permissions? If not, brand familiarity may be distorting assurance.

03 / TRUST DIMENSIONS

Dimensions 5 to 8: orchestration, people, resilience and evidence

The second four dimensions determine how capability becomes action, how humans retain meaningful control, how the system fails safely and how trust remains current.

5. ORCHESTRATION, TOOLS & PERMISSIONS

Constrain goals, prompts, agent routing, memory, tools, targets and identities. Enforce least privilege, task-specific access, allowlists and separation of duties.

Minimum evidence: tool inventory, permission map, policy-as-code, sandbox design, action thresholds and two-person rules where required.

6. HUMAN CONTROL & ACCOUNTABILITY

Name the business owner, technical owner, accountable risk decision-maker and live supervisor. Design interfaces for challenge, escalation, veto and intervention.

Minimum evidence: RACI, supervision mode, decision window, workload test, escalation procedure, training and override authority.

7. RESILIENCE & RECOVERABILITY

Limit blast radius and assume unexpected behaviour. Provide isolation, circuit breakers, rate limits, safe defaults, rollback, fallback and continuity arrangements.

Minimum evidence: failure scenarios, kill switch, rollback test, fallback process, recovery time objective and containment validation.

8. EVIDENCE, MONITORING & RE-AUTHORISATION

Log the complete decision chain, monitor behaviour and outcomes, detect drift, retain evidence, trigger reassessment and measure value with risk.

Minimum evidence: evidence schema, live telemetry, behavioural baseline, alert thresholds, review cadence, material-change rules and audit trail.

CRITICAL DISTINCTION

Explainability describes an output. Traceability reconstructs a decision. Cybersecurity operations require both, but traceability is the stronger operational control.

03 / USE-CASE CONTROL

Every use case needs a Trust Contract

The Trust Contract is the operational agreement between business intent, technical enforcement, human authority and live evidence. It should be readable by people and enforceable by the platform.

Figure 9. The AI Trust Contract. Source: CiBRAI analysis, 2026.

A Trust Contract is not a lengthy impact assessment repeated for every workflow. It is a concise operating licence that answers seven questions: purpose, data boundary, decision boundary, action boundary, human authority, evidence duty and change triggers. It should reference reusable evidence packs for common platform components while preserving use-case accountability.

Contract acceptance criteria

- The system cannot access a tool, target or data source outside the declared boundary.
- The required human gate cannot be bypassed through prompt, workflow or API changes.
- Every recommendation and action receives a correlation identifier and action receipt.
- Material changes automatically suspend, narrow or flag the operating licence for review.
- The owner can revoke authority and recover safely without depending on the AI system itself.

03 / TRACEABILITY

The AI Decision Evidence Chain

The SOC must be able to reconstruct why an action was proposed, authorised, executed and judged successful or unsuccessful.

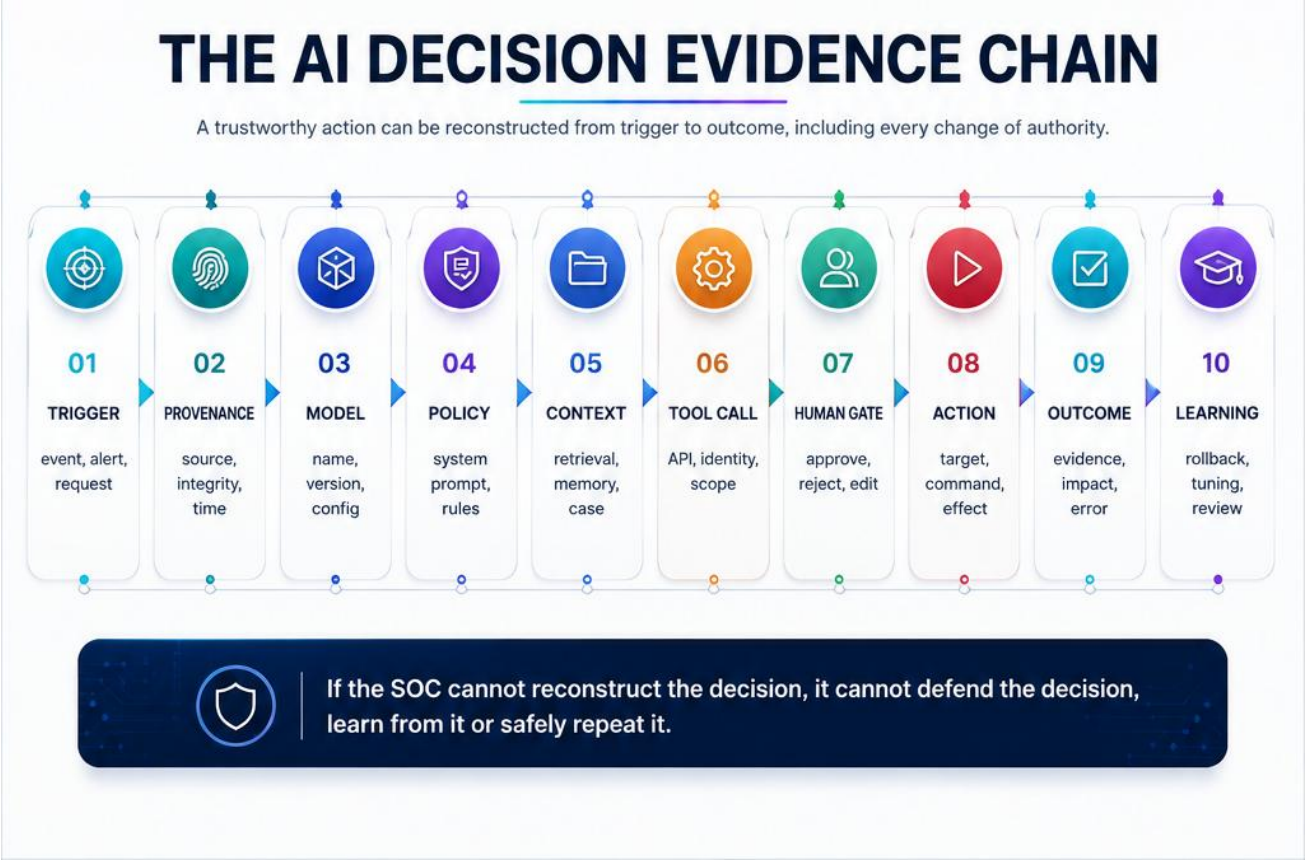


Figure 10. The AI Decision Evidence Chain. Source: CiBRAI analysis, 2026.

The evidence chain should be queryable at incident speed, not assembled manually after an event. It must correlate machine and human activity across multiple systems and preserve the exact context that influenced the decision. This includes model and policy versions, retrieval results, confidence or uncertainty, tool identity, permission scope, human changes, execution result and rollback.

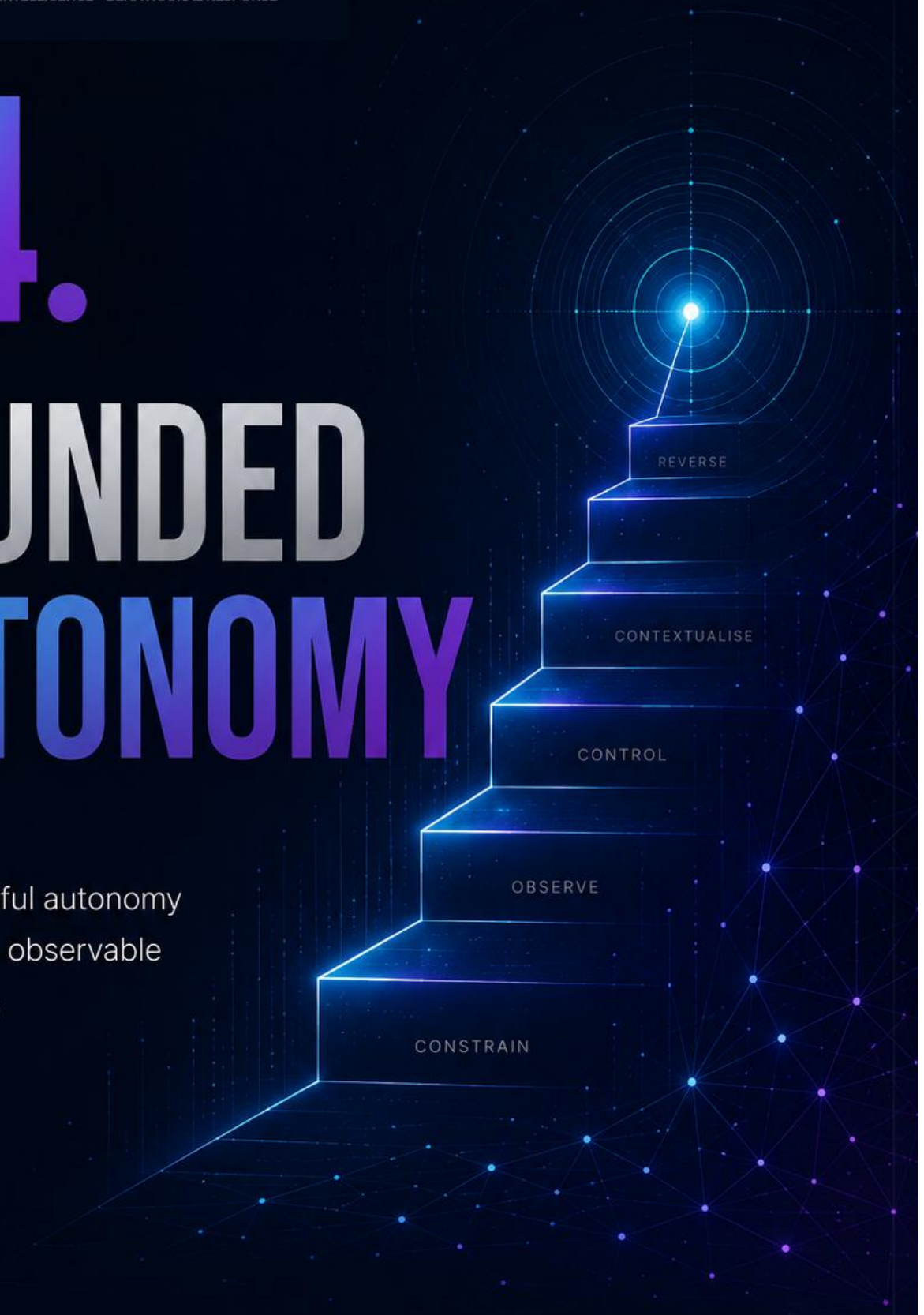
Evidence quality	Evidence use
Complete, time-synchronised, tamper-evident, attributable, searchable, retained according to policy and intelligible to independent reviewers.	Operational monitoring, incident response, assurance review, root-cause analysis, regulatory response, control tuning, supplier challenge and re-authorisation.
	NON-NEGOTIABLE ***If the organisation cannot reconstruct an AI-influenced cyber action, it cannot defend the action, learn from it or safely automate it again.***

CiBRAI
CYBER INTELLIGENCE - BEHAVIOURAL RESPONSE

04.

BOUNDED AUTONOMY

The safest useful autonomy is constrained, observable and reversible.



04 / DELEGATED AUTHORITY

Bounded autonomy is the safe path to useful autonomy

The relevant question is not whether AI is autonomous. It is autonomous over what, using which identity and tools, against which targets, for how long, with what evidence and under whose authority.

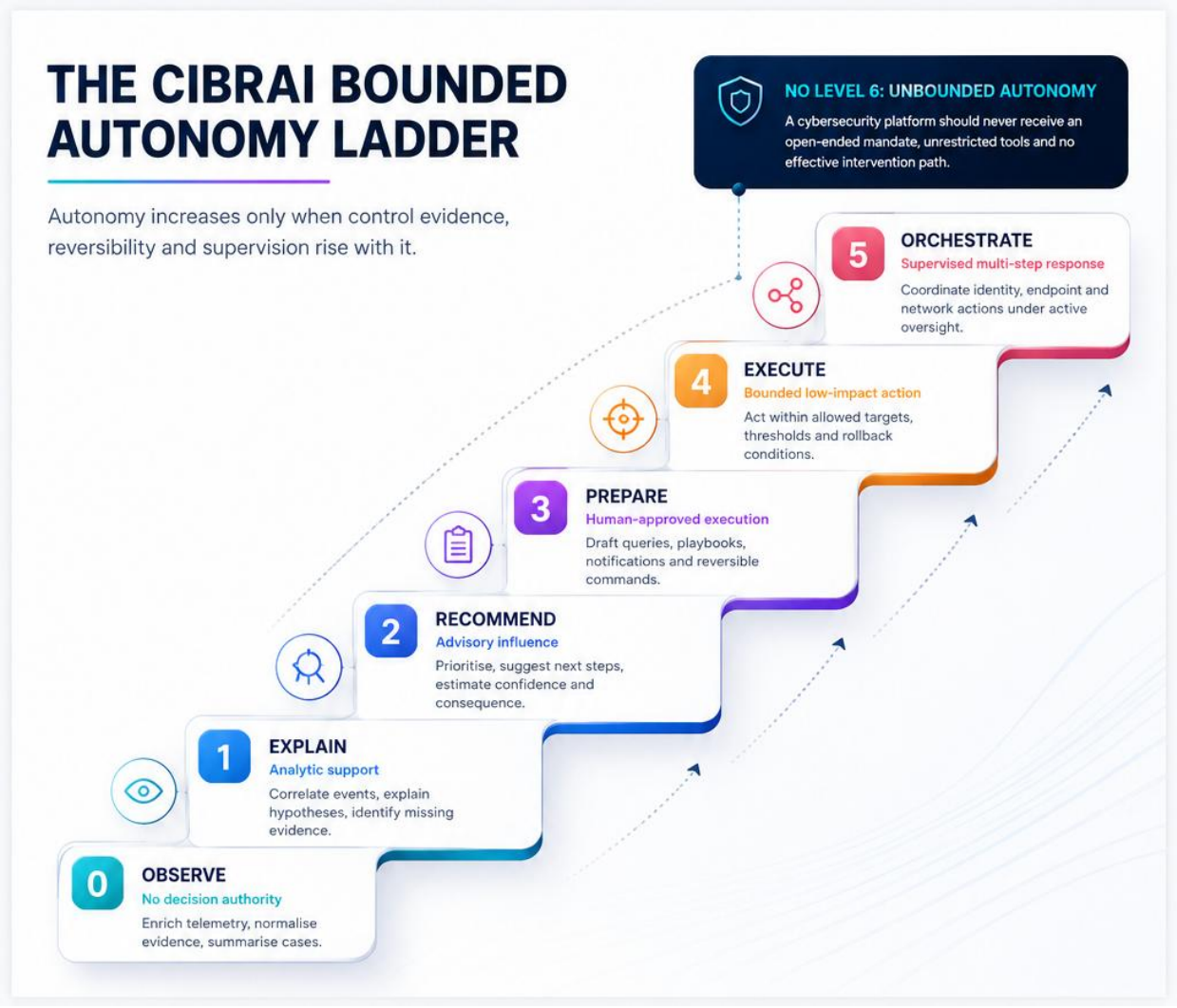


Figure 11. The CiBRAI Bounded Autonomy Ladder. Source: CiBRAI analysis, 2026.

Each increase in autonomy should be supported by stronger evidence, tighter technical controls, tested recovery and clearer human supervision. Levels are assigned per use case and action, not to a platform as a whole. The same platform may operate at Level 0 for one data source, Level 3 for investigation preparation and Level 4 for a small set of reversible containment actions.

NO LEVEL 6

Unbounded autonomy is not a maturity level. An open-ended objective, unrestricted tools, broad credentials and no effective intervention path are an unacceptable architecture for cyber defence.

04 / HUMAN CONTROL

Human-in-the-loop is not a control by itself

Human oversight only works when the person has time, authority, evidence and competence. Otherwise, the human becomes a ceremonial checkpoint or a liability absorber.

Figure 12. Human oversight modes and control conditions. Source: CiBRAI analysis, informed by current agentic AI guidance [15].

Oversight mode should follow consequence and decision time. Human-in-the-loop suits irreversible, high-consequence or rights-affecting actions. Human-on-the-loop can support fast bounded response where live monitoring, stop authority and rollback are reliable. Human-out-of-the-loop should be limited to low-consequence, reversible and highly constrained activities. Current cyber guidance similarly recommends named responsibility, real-time monitoring, human intervention and robust sandboxing for higher-risk agentic use. [13][15]

INTERFACE REQUIREMENT

Show the evidence that matters for the decision: source quality, missing context, uncertainty, policy basis, proposed target, expected effect, reversibility and time available. A confidence percentage alone is not decision support.

04 / ADVERSARIAL RISK

The attacker will target the trust relationship

AI-enabled SOCs create new paths to influence detection, triage and response. Attackers may target the model, but they can also target data, memory, prompts, goals, identities, tools, inter-agent communication and human reliance.

Figure 13. Attack surface of an AI-enabled SOC. Source: CiBRAI analysis, informed by OWASP and MITRE threat research [16][17][18][19].

Threat modelling should assume that malicious content reaches the system through ordinary operational channels. A ticket, log field, malware note, web page or threat intelligence record may contain instructions intended for the agent. The control objective is not to make the model invulnerable. It is to prevent untrusted content from changing authority, policy or tool access, and to detect when behaviour deviates.

PRIORITY RISKS

For cybersecurity platforms, the highest priority risks are often goal hijacking, excessive privilege, tool misuse, poisoned context, cascading action and evidence loss. Hallucination is important, but it is only one failure mode.

04 / DEFENCE IN DEPTH

Agentic risks require layered controls

No single guardrail closes an agentic risk. Effective defence combines provenance, instruction and data separation, policy enforcement, least privilege, sandboxing, human gates, live monitoring, rollback, change control and adversarial testing.

Figure 14. Threat-to-control matrix. Source: CiBRAI analysis, 2026.

The matrix is intentionally many-to-many. Prompt injection is not solved by input filtering alone. Tool misuse is not solved by a better system prompt. Model drift is not solved by an annual review. Controls must be distributed across the action system and tested together under realistic failure conditions.

Minimum adversarial test set

- Direct and indirect prompt injection through operational content.
- Conflicting, stale, poisoned and classification-inappropriate retrieval sources.
- Attempts to expand tool scope, credentials, network access or target range.
- False urgency, authority claims and cross-agent impersonation.
- Cascading errors, partial system failure, monitoring loss and rollback under load.

CiBRAI
CYBER INTELLIGENCE - BEHAVIOURAL RESPONSE

05.

ASSURANCE THAT OPERATES

Lifecycle controls, live evidence and the courage to stop the machine.

05 / LIFECYCLE

Approval is a waypoint, not the end of assurance

AI systems, operational context and adversaries change continuously. Assurance must travel with the use case from discovery through operation, recovery and re-authorisation.

Figure 15. Lifecycle assurance for AI in cyber operations. Source: CiBRAI analysis, aligned with lifecycle approaches in NIST, ISO and secure AI development guidance [4][7][8][14][20].

A staged model avoids false certainty. Low-risk pilots can proceed with tight data and user boundaries, no privileged action and explicit learning objectives. Controlled deployment adds monitoring and stronger evidence. Bounded execution requires proven identity, permission, rollback and supervision controls. Scale should occur only when operational evidence confirms that the system remains within its Trust Contract.

GOVERNANCE OUTCOME

The best assurance process makes low-risk experimentation easier and high-risk action harder. It should reduce duplication while increasing evidence quality.

05 / DECISION GATES

A staged assurance pathway for SOC use cases

The gate model below is designed to be proportionate. It uses reusable evidence for shared platform components, while preserving use-case-specific decisions about data, authority and consequence.

GATE	REQUIRED EVIDENCE	DECISION
G0 PROPOSE	Named owner; purpose; users; affected assets; expected value; initial risk tier; prohibited decisions.	Proceed to controlled discovery or reject the use case.
G1 PILOT	Isolated environment; approved data; no privileged production action; evaluation plan; exit criteria.	Authorise time-bound pilot with explicit learning objectives.
G2 CONTROLLED DEPLOYMENT	Trust Contract; evidence chain; access controls; human gate; monitoring; support and incident process.	Authorise limited users, assets and operating hours.
G3 BOUNDED EXECUTION	Named agent identity; least privilege; sandbox; action allowlist; thresholds; rollback; adversarial tests.	Authorise specific reversible actions under defined supervision.
G4 SCALE	Operational outcomes; drift and error trends; resilience tests; supplier evidence; capacity; training; audit results.	Expand scope without increasing autonomy beyond evidence.
G5 MATERIAL CHANGE	Model, prompt, policy, data, tool, identity, target, user, geography or consequence changes.	Continue, narrow, suspend or re-authorise based on changed assurance load.

Reusable evidence, local accountability

Platform architecture, baseline model evaluations, secure configuration patterns, data-flow diagrams and supplier evidence should be reusable. The deploying organisation must still decide whether the specific use case, data, permissions, supervision and consequences are acceptable. Reuse should remove duplication, not ownership.

ANTI-PATTERN

Do not make every team repeat a full platform review. Do not let a platform review authorise every use case. These are opposite errors.

05 / MATURITY

AI trust maturity is evidence maturity

Higher maturity means the organisation can safely grant more useful authority because controls are integrated, continuously evidenced and recoverable. It does not mean that every process becomes autonomous.

Figure 16. AI trust maturity for cyber operations. Source: CiBRAI analysis, 2026.

Most organisations should aim first for Level 2, controlled copilots, and Level 3, governed agents. Level 4 requires live behavioural monitoring, material-change detection, regular adversarial scenarios and operational re-authorisation. Level 5 is an adaptive control fabric, not permission for unbounded action. Dynamic permissions should narrow or expand within pre-authorised limits based on evidence and risk.

MATURITY TEST

Can the organisation identify every AI agent, its owner, current permissions, approved purpose, model and policy version, live actions, evidence health and revocation method in one view? If not, it is not ready to scale agentic authority.

05 / IMPLEMENTATION

A practical 90-day path

Organisations do not need to wait for a perfect enterprise AI framework before controlling high-risk SOC use. Start with visibility, boundaries and evidence. Expand autonomy only when assurance supports it.

Figure 17. A practical 90-day implementation path, followed by scale over three to twelve months. Source: CiBRAI analysis, 2026.

The first month should expose what already exists, including embedded AI features, personal tools, platform agents, service identities and automation that has quietly become consequential. The second month should create enforceable Trust Contracts and test failure scenarios. The third month should operate a small production cohort, exercise recovery and compare promised value with observed risk. Scale should follow evidence, not enthusiasm.

PROGRAMME DESIGN

Treat AI trust as a cross-functional cyber uplift programme spanning GRC, architecture, SOC, identity, data, privacy, legal, procurement, change management and resilience. Assign one accountable executive and one operational product owner.

05 / MEASUREMENT

Measure trust integrity and security value together

Counting forms, committees and approvals says little about whether the AI system is safe, useful or under control. A balanced scorecard should measure control integrity, behaviour, human control and security outcomes.

Figure 18. The AI Trust Scorecard. Source: CiBRAI analysis, 2026.

Metrics require interpretation. A high human override rate may indicate poor model performance, healthy challenge culture or overly conservative thresholds. A low override rate may indicate excellent calibration or automation bias. Pair quantitative measures with scenario review, analyst interviews, evidence sampling and independent challenge.

Three metrics to avoid using alone

- Analyst acceptance rate. Acceptance is behaviour, not proof of correctness or justified trust.
- Model accuracy. Aggregate accuracy can hide catastrophic failure on rare, high-consequence cases.
- Number of automated actions. Volume rewards authority without proving safety, value or recovery.

EXECUTIVE ACTION

The board and executive decisions that cannot be delegated

Technology teams can design controls, but only executives can define acceptable consequence, delegated authority and the conditions under which the organisation will rely on AI.

01 RISK APPETITE What harms are unacceptable, even if the probability is low?	**02 NO-GO ZONES** Which decisions or actions must remain human, legally reserved or outside the platform?
03 AUTONOMY CEILING What is the highest permitted autonomy level by environment and asset class?	**04 ACCOUNTABLE OWNER** Who accepts residual risk and who supervises live operation?
05 EVIDENCE THRESHOLD What must be proven before advice, execution and orchestration are permitted?	**06 STOP AUTHORITY** Who can suspend the system immediately, and can they do so without relying on it?

Twelve questions for every AI-enabled SOC

01 What decisions and actions can the system influence or execute?	**07** Who has time, authority and competence to intervene?
02 Which actions are irreversible, externally visible or service-affecting?	**08** What happens when monitoring, the model or a dependent service fails?
03 What identities, credentials, tools, targets and networks can it access?	**09** Can the organisation roll back safely and meet continuity obligations?
04 How can an attacker shape its inputs, context, goals or memory?	**10** Which changes trigger automatic narrowing, suspension or re-authorisation?
05 Can every recommendation and action be reconstructed end to end?	**11** What evidence comes from suppliers, and what remains our responsibility?
06 What uncertainty or missing evidence is shown to the analyst?	**12** Are we improving analyst judgement or creating automation dependence?

CONCLUSION

The AI Trust Manifesto for cyber defence

Trust must move from a marketing adjective to an operational discipline. The following principles summarise the CiBRAI position.

01 **Trust is calibrated reliance, not confidence theatre.**
Do not maximise trust. Match reliance to evidence, consequence and context.

02 **Govern the action system.**
Model, data, context, orchestration, identity, tools, people and recovery form one control object.

03 **Capability does not imply authority.**
Grant the minimum authority required for the task and time window.

04 **A human in the loop must be able to act.**
Time, authority, evidence and competence are mandatory.

05 **Explainability is not enough.**
Every action must be attributable and reconstructable.

06 **Assume the system will be attacked and will sometimes be wrong.**
Design for containment, reversibility and recovery.

07 **Trust decays.**
Change and drift consume assurance; re-authorise before the operating envelope is lost.

08 **Reuse evidence, not accountability.**
Shared platform evidence should accelerate review without removing local ownership.

09 **Measure value and risk together.**
A safe system that delivers no value and a fast system that cannot be governed are both failures.

10 **Unbounded autonomy is not maturity.**
The goal is useful, bounded and continuously assured digital defence.

FINAL CHALLENGE

Stop asking whether the model can do the task. Ask under what evidence, authority, constraints and recovery conditions the system should be permitted to do it now.

APPENDIX A

Minimum control catalogue: dimensions 1 to 4

The catalogue is a baseline. Control depth should increase with the assurance load of the use case.

CONTROL	MINIMUM EVIDENCE	ADVANCED EVIDENCE
1.1 Use-case register	Purpose, owner, users, data, model, tools, autonomy level and review date	Live inventory linked to configuration and telemetry
1.2 Risk and consequence tier	Document impact, exposure, privilege, reversibility and affected stakeholders	Dynamic tiering based on asset and threat context
1.3 No-go decisions	Explicitly prohibit reserved, rights-affecting or unacceptable actions	Policy enforcement blocks prohibited action paths
1.4 Delegated authority	Named approver, scope, time limit and conditions	Machine-readable authority token with automatic expiry
2.1 Agent and service identity	Unique non-human identities; no shared credentials	Workload identity, attestation and short-lived credentials
2.2 Provenance	Record source, time, model, prompt, policy and context versions	Cryptographic signing and tamper-evident action receipts
2.3 Correlation	End-to-end identifier across reasoning, human and tool activity	Cross-platform evidence graph and automated reconstruction
2.4 Supplier traceability	Provider, component, dependency and change information	AI bill of materials and continuous dependency monitoring
3.1 Data classification	Approved data classes and prohibited inputs	Context-aware enforcement and automatic redaction
3.2 Source allowlisting	Approved telemetry, CTI, retrieval and memory sources	Source reputation and adaptive integrity scoring
3.3 Poisoning controls	Integrity checks, validation and anomalous-content review	Poisoning detection, canary data and provenance analytics
3.4 Confidentiality	Access, retention, encryption and tenant isolation	Attribute-based access and privacy-preserving processing
4.1 Use-case evaluation	Representative tasks, edge cases and failure scenarios	Continuous evaluation against live and synthetic cases
4.2 Uncertainty and calibration	Display limitations, missing evidence and confidence appropriately	Per-use-case calibration monitoring and threshold adaptation
4.3 Adversarial testing	Prompt injection, context manipulation and goal conflict tests	Independent red team and automated attack regression suite
4.4 Model change control	Version awareness, regression tests and fallback	Behavioural attestation and automatic re-authorisation trigger

APPENDIX A

Minimum control catalogue: dimensions 5 to 8

These controls govern the conversion of capability into action and the ability to intervene, recover and prove what occurred.

CONTROL	MINIMUM EVIDENCE	ADVANCED EVIDENCE
5.1 Tool inventory	Approved tools, APIs, commands, targets and owners	Machine-enforced tool registry with risk attributes
5.2 Least privilege	Task-specific permissions and short-lived access	Just-in-time privilege with dynamic narrowing
5.3 Policy-as-code	Enforced action boundaries, thresholds and prohibited paths	Formally tested policies and independent policy decision point
5.4 Sandbox and network control	Isolated execution, allowlists, rate limits and egress controls	Per-task microsegmentation and blast-radius simulation
6.1 Named accountability	Business owner, technical owner, risk approver and supervisor	Live ownership linked to every decision and action
6.2 Oversight mode	Human-in, on or out-of-loop justified by consequence	Adaptive oversight that tightens when risk indicators rise
6.3 Human factors	Training, workload, interface and intervention testing	Scenario certification and ongoing calibration exercises
6.4 Escalation and veto	Clear pause, reject, edit, revoke and two-person rules	Independent stop authority and out-of-band shutdown
7.1 Containment	Rate limits, target scope, action caps and safe defaults	Dynamic circuit breakers based on live risk telemetry
7.2 Rollback	Defined reversible actions and tested recovery	Automated compensating actions with human confirmation
7.3 Fallback and continuity	Manual or deterministic fallback process	Graceful degradation and resilient alternate control plane
7.4 Incident response	AI-specific detection, triage, containment and notification	Dedicated AI action forensics and supplier coordination
8.1 Evidence schema	Inputs, context, policy, model, tools, humans, actions and outcomes	Normalised cross-platform evidence graph
8.2 Behaviour monitoring	Error, drift, override, exception and outcome monitoring	Real-time policy deviation and anomaly detection
8.3 Material-change triggers	Defined re-assessment events and time-based review	Automatic licence narrowing or suspension
8.4 Value-risk scorecard	Control integrity, behaviour, human control and security value	Board-level trends linked to risk appetite and investment

APPENDIX B

How the Trust Fabric aligns with major frameworks

The CiBRAI framework is an operational extension for cybersecurity platforms and SOCs. It is designed to complement, not replace, established standards and guidance.

FRAMEWORK OR GUIDANCE	PRIMARY EMPHASIS	TRUST FABRIC CONTRIBUTION
NIST AI RMF 1.0 and GenAI Profile	Govern, Map, Measure and Manage; trustworthiness characteristics; lifecycle risk management	All eight dimensions; lifecycle assurance; scorecard and material-change logic
ISO/IEC 42001	AI management system, policy, objectives, roles, risk, performance evaluation and continual improvement	Governance operating model, roles, evidence, monitoring and improvement
ISO/IEC 23894 and 42005	AI risk management and impact assessment across the lifecycle	Assurance load, consequence, stakeholders, lifecycle gates and re-assessment
Australian Guidance for AI Adoption	Fit-for-purpose risk management, acceptable risk, lifecycle checks and supply-chain risk	Trust Contract, risk tiering, shared responsibility and re-authorisation
ASD opportunities and agentic AI guidance	AI aligned to cyber functions; secure-by-design; human oversight; privilege, monitoring and incremental adoption	SOC use spectrum, Bounded Autonomy Ladder, least privilege, sandbox, rollback
NCSC secure AI lifecycle and agentic risk guidance	Secure design, development, deployment, operation; threat modelling; oversight and sandboxing	Action-system architecture, lifecycle, human control and resilience
OWASP GenAI and agentic guidance	Prompt injection, insecure outputs, tool misuse, goal hijacking, memory and inter-agent risks	Threat landscape, layered controls and adversarial test set
MITRE ATLAS	Adversarial tactics and techniques against AI-enabled systems	Threat modelling, test scenarios, detection and response evidence
ETSI EN 304 223 and ENISA guidance	Baseline cybersecurity across the AI lifecycle and supply chain	Control catalogue across data, model, deployment, operation and end-of-life

Use the crosswalk correctly

A crosswalk is not proof of control effectiveness. It shows conceptual alignment. Evidence must still demonstrate that controls are implemented in the deployed environment, operate under realistic conditions and remain effective through change.

CIBRAI POSITION

Compliance can organise trust evidence, but compliance is not the same as trustworthy operation. The control must work when the attacker, the model and the organisation are all under pressure.

REFERENCES

Selected research, standards and public guidance

Sources were selected for their relevance to human-automation trust, AI risk management, secure AI engineering, agentic systems and cybersecurity operations. Accessed 24 August 2026 unless otherwise stated.

[1] Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

[2] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230-253. https://doi.org/10.1518/001872097778543886

[3] Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407-434. https://doi.org/10.1177/0018720814547570

[4] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

[5] National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. https://doi.org/10.6028/NIST.AI.600-1

[6] National Institute of Standards and Technology. (2026). AI Agent Standards Initiative. https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative

[7] International Organization for Standardization. (2023). ISO/IEC 42001:2023, Artificial intelligence management systems. https://www.iso.org/standard/42001

[8] International Organization for Standardization. (2023). ISO/IEC 23894:2023, Artificial intelligence guidance on risk management. https://www.iso.org/standard/77304.html

[9] International Organization for Standardization. (2025). ISO/IEC 42005:2025, AI system impact assessment. https://www.iso.org/standard/42005

[10] Australian Government, National AI Centre. (2026). Guidance for AI adoption: Implementation guidance. https://www.ai.gov.au/staying-safe-and-responsible/essential-ai-practices/guidance-ai-adoption-implementation-guidance

[11] Australian Government. (2024). National framework for the assurance of artificial intelligence in government. https://www.finance.gov.au/government/public-data/data-and-digital-ministers-meeting/national-framework-assurance-artificial-intelligence-government

[12] Australian Signals Directorate and international partners. (2026). Opportunities for AI in cyber defence. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/opportunities-for-ai-in-cyber-defence

[13] Australian Signals Directorate and international partners. (2026). Careful adoption of agentic AI services. https://www.cyber.gov.au/business-government/secure-design/artificial-intelligence/careful-adoption-of-agentic-ai-services

[14] UK National Cyber Security Centre and international partners. (2023). Guidelines for secure AI system development. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

[15] UK National Cyber Security Centre. (2026). Managing the cyber risk of agentic AI. https://www.ncsc.gov.uk/blogs/managing-the-cyber-risk-of-agentic-ai

[16] OWASP GenAI Security Project. (2026). OWASP Top 10 for LLM Applications 2026. https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/

[17] OWASP GenAI Security Project. (2026). OWASP Top 10 for Agentic Applications 2026. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

[18] OWASP GenAI Security Project. (2025). Securing Agentic Applications Guide 1.0. https://genai.owasp.org/resource/securing-agentic-applications-guide-1-0/

[19] MITRE. (2026). MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. https://atlas.mitre.org/

[20] European Telecommunications Standards Institute. (2026). ETSI EN 304 223: Baseline Cyber Security Requirements for AI Systems and Models. https://www.etsi.org/enjoy-magazine/articles/ai-cybersecurity-standard-etsi-en-304-223/

[21] European Union Agency for Cybersecurity. (2023). Multilayer Framework for Good Cybersecurity Practices for AI. https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai

[22] National Security Agency and international partners. (2025). AI data security: Best practices for securing data used to train and operate AI systems. https://media.defense.gov/2025/May/22/2003720521/-1/-1/0/CSI-AI-DATA-SECURITY-BEST-PRACTICES.PDF

[23] European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

ARTIFICIAL INTELLIGENCE DISCLOSURE

AI tools assisted with visual ideation, drafting support and production. The final concepts, analysis, control positions, editorial decisions and report assembly were directed and reviewed as CiBRAI research. The report does not represent certification of any product or deployment.

CiBRAI

CYBER INTELLIGENCE - BEHAVIOURAL RESPONSE

THE INDUSTRY SHOULD STOP ASKING:

"CAN THE MODEL DO THIS?"

AND START ASKING:

"UNDER WHAT EVIDENCE, AUTHORITY, CONSTRAINTS AND RECOVERY CONDITIONS SHOULD THE SYSTEM BE PERMITTED TO DO THIS NOW?"

ABOUT CIBRAI

CiBRAI is an Australian cybersecurity operating platform and research company focused on turning fragmented security evidence into governed, explainable and actionable defence. Our work combines cyber intelligence, behavioural response, continuous compliance and bounded agentic automation.

cibrai.com | support@cibrai.com | 1800 928 982 | +61 2 9000 1137

CLARITY. NOT CHAOS.