ETHIC VAULT
Governed Execution

PUBLIC WHITE PAPER

GOVERN BEFORE CONSEQUENCE

How to Control Agentic AI Before Execution

THE THESIS

AI may propose. It should never inherit authority by implication.

An evidence-led executive and technical guide to pre-execution governance

Ethic Vault | Version 1.0 | July 2026

02 / PUBLIC WHITE PAPER

Executive brief

Agentic AI creates value by acting. Governance must therefore control the transition from proposal to consequence.

Agentic systems can select tools, access data, preserve context, and take multi-step actions with limited continuous supervision. That expands the attack surface and compresses the time available for human review. Public guidance now converges on least privilege, bounded deployment, clear accountability, human oversight, and secure-by-design controls.[4, 5, 7, 9]

- **Content safety is not action authority.** A useful or harmless-looking answer may still request an action the organization has not authorized.
- **A model can propose; a separate control must decide.** The decision point should evaluate identity, scope, policy, evidence, and consequence before any external action is released.
- **High-impact actions require stronger proof.** The level of control should rise with autonomy, irreversibility, legal exposure, and the magnitude of consequence.
- **Evidence labels matter.** Claimed, implemented, measured, reported, independently verified, and production-ready are different statements and must never be collapsed.

PUBLIC	EVIDENCE	POSITION
This paper presents a public conceptual control model and cites external authorities. It does not disclose proprietary enabling details, certify any product, or claim independent verification or production readiness for Ethic Vault technology.		

What the reader will leave with

- A practical definition of the authorization boundary.
- A public, non-enabling reference pattern for pre-execution control.
- A threat-to-control map, evidence taxonomy, pilot plan, and buyer checklist.

03 / PUBLIC WHITE PAPER

Agentic AI moves risk from content to consequence

When a system can choose and invoke tools, the risk question changes from 'What did it say?' to 'What was it allowed to do?'

UK Government guidance describes agentic AI as systems composed of agents that can make decisions and interact autonomously to achieve objectives. The NCSC adds the operational details: agents may access data, remember context, use tools, take actions, and create sub-agents. These capabilities make agents useful, but also more hazardous than non-agentic tools. [7, 9]

Five risk multipliers

- **Access:** The agent may reach data, APIs, accounts, devices, or business systems.
- **Speed:** Actions can occur faster than meaningful human review.
- **Chaining:** One tool result can become the next instruction, propagating error or manipulation.
- **Ambiguity:** A goal can be interpreted in a way the principal did not intend.
- **Persistence:** Memory, credentials, and recurring tasks can extend risk beyond one interaction.

OWASP identifies prompt injection and excessive agency as distinct risks. Its excessive-agency guidance traces damaging actions to excessive functionality, permissions, or autonomy, and recommends minimum necessary functions and privileges, human approval for high-impact actions, and authorization in downstream systems rather than delegation to the language model. [4, 5]

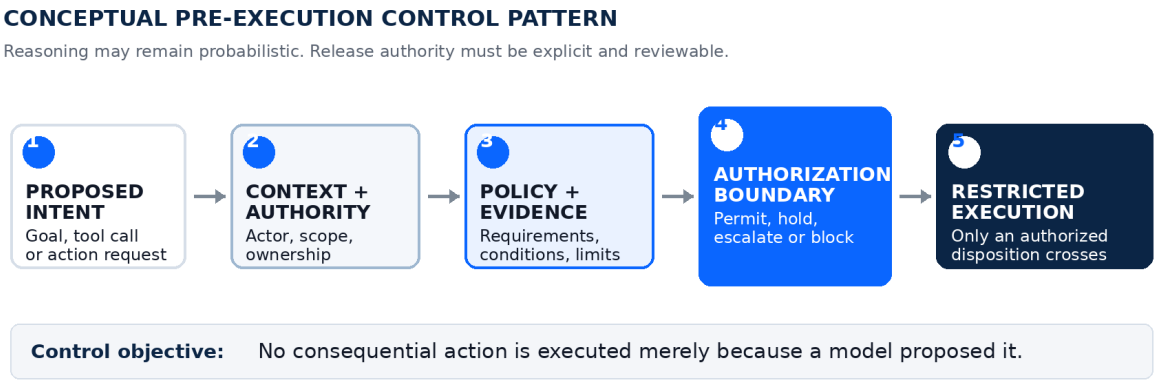
A UK-government-commissioned 2026 review reported that agentic-AI security publications grew 216 percent in 2025 over 2024, while agentic AI remained the least represented AI-system category in its sample. The signal is clear: deployment momentum is rising faster than mature evidence and governance practice. [10]

> **THE GOVERNANCE SHIFT**
> Monitoring remains necessary, but it is downstream. The decisive control is the one that can stop an unauthorized action before the action becomes an operational fact.

04 / PUBLIC WHITE PAPER

The authorization boundary

Ethic Vault working definition: the explicit decision point at which an AI-proposed action is permitted, conditioned, escalated, held, or denied before execution.



Four properties of a credible boundary

- **Separate authority:** The model does not approve its own action.
- **Complete mediation:** Every consequential request crosses the same governed release point.
- **Fail closed:** Missing, conflicting, stale, or unauthorised conditions result in hold or denial, not implied permission.
- **Reviewable evidence:** The organization can later determine which policy, authority, evidence, and disposition applied.

WHAT DETERMINISTIC MEANS HERE

It does not mean the generative model becomes deterministic. It means the release decision is reproducible for the same bounded request, policy state, authority state, and evidence set. This is a control objective, not a claim that all upstream reasoning can be replayed.

05 / PUBLIC WHITE PAPER

A public reference pattern for pre-execution control

The pattern below is intentionally conceptual. It communicates the control objective without exposing implementation details.

1. **Define the action.** Express the proposed external action, target, scope, and expected consequence in a bounded request.
2. **Resolve the principal.** Identify on whose behalf the action is requested and what authority that principal actually holds.
3. **Apply policy.** Evaluate the request against the policy and risk state that is valid for this decision.
4. **Require evidence.** Confirm that mandatory approvals, facts, preconditions, and human attestations are present and current.
5. **Issue a disposition.** Return an explicit permit, hold, escalation, or block decision. Silence is not permission.
6. **Restrict execution.** Allow the downstream system to act only on an authorized disposition, then retain decision evidence for review.

Reasoning plane vs. control plane

Dimension	Reasoning plane	Control plane
Purpose	Generate, plan, retrieve, recommend	Authorize, constrain, escalate, deny
Typical behavior	Probabilistic and adaptive	Explicit and policy-bound
Failure concern	Error, manipulation, ambiguity	Unauthorized release or weak evidence
Proof question	Why did the model propose this?	Why was this action allowed or refused?

For high-risk AI systems within its scope, the EU AI Act requires effective human oversight and the ability, as appropriate, to disregard, override, reverse, interrupt, or stop system operation. The authorization boundary makes those responsibilities operational rather than ceremonial. [8]

06 / PUBLIC WHITE PAPER

Threat-to-control map

The control objective is not to predict every failure. It is to limit what any failure can authorize.

Failure mode	What can go wrong	Boundary response	Evidence to retain
Prompt injection	Untrusted content changes the agent's behavior or instructions.	Treat model output as a proposal; re-check authority and policy outside the model.	Input source, requested action, policy result, disposition.
Excessive agency	The agent has unnecessary functions, permissions, or autonomy.	Minimum functions and privileges; explicit approval for high-impact actions.	Granted scope, tool called, approver, downstream identity.
Credential sprawl	Long-lived or shared credentials expand the blast radius.	Short-lived, task-scoped credentials and rapid revocation.	Credential scope, issue time, expiry, revocation event.
Tool/output abuse	Malformed or unsafe output is passed to a downstream system.	Typed interfaces, validation, allowlisted operations, complete mediation.	Validated parameters, interface version, execution response.
Supply-chain or data poisoning	A model, component, dataset, or tool response is compromised.	Provenance requirements, component inventory, bounded trust, isolation.	Source identity, version, integrity result, dependency chain.
Memory/context corruption	Stale or manipulated context silently shapes later actions.	Context boundaries, expiry, source labels, conflict handling.	Context source, age, conflict status, selected evidence.
Speed and irreversibility	Actions occur before review or cannot be easily reversed.	Risk-tiered human approval, rate limits, staged release, safe stop.	Approval time, release stage, stop event, recovery record.

Source basis: OWASP prompt injection and excessive agency; NCSC least privilege, scope limitation, monitoring, threat modelling, and incident planning; EU AI Act human oversight and cybersecurity provisions. [4, 5, 7, 8]

> **DESIGN** **RULE**
>
> Do not grant a model a capability merely because it can describe when that capability should be used. Description is not authority, and model self-restraint is not access control.

07 / PUBLIC WHITE PAPER

Evidence without exaggeration

Trust rises when the evidence label is as precise as the technical claim.

Evidence status	What it means	What it does not mean
Claimed	A party asserts that a capability or result exists.	No implementation, measurement, or verification is established.
Implemented	A defined artifact or control exists in a stated environment.	Existence alone does not prove effectiveness, security, or scale.
Measured	A metric was captured under described conditions.	The result may not generalize beyond the test setup or workload.
Reported	The originating party communicated a result or outcome.	The reader has not independently reproduced the underlying evidence.
Independently verified	A qualified independent party inspected or reproduced a defined claim within scope.	Verification does not automatically establish deployment acceptance.
Production-ready	The system passed the organization's defined operational, security, governance, reliability, support, and acceptance gates for a named environment.	Readiness is scope-specific; it is not universal certification.

NOT A SINGLE LADDER

These labels can overlap. A result can be measured and reported without independent verification; an independently verified component can still be unready for production. Always state the environment, method, owner, limitations, and decision scope.

Minimum public claim record

01 Claim: the precise capability or result being asserted.

02 Status: one or more labels from the table above.

03 Scope: system, version, environment, workload, and date.

04 Method: test, inspection, reproduction, or acceptance procedure.

05 Limit: what the evidence cannot establish.

06 Verifier: originating party or named independent reviewer.

08 / PUBLIC WHITE PAPER

From policy to a bounded pilot

The first credible deployment is narrow enough to stop, inspect, and learn from.

1. Bound

Choose one consequential action, one owner, one system boundary, and one explicit release authority. Define what the agent may read, propose, and never execute.

Output: boundary statement, authority map, unacceptable outcomes, exit criteria.

2. Instrument

Capture the request, policy state, evidence requirement, disposition, approver, and downstream result. Preserve enough information for review without collecting unnecessary sensitive data.

Output: evidence plan, logging model, retention and access rules.

3. Exercise

Run normal, edge, adversarial, stale-context, missing-evidence, and bypass scenarios. Test the stop path and the human escalation path, not only successful permits.

Output: test matrix, results, unresolved risk, remediation record.

4. Decide

Compare results with pre-agreed acceptance gates. Expand only if the control remains effective and usable at the next consequence level.

Output: accept, restrict, redesign, or stop decision with accountable sign-off.

Acceptance measures

- Unauthorized-action reachability and bypass success rate.
- Disposition reproducibility for the same bounded decision state.
- False-permit and false-hold rates under the defined test set.
- Decision latency, escalation time, evidence completeness, and safe recovery.

> **PILOT BEFORE AUTONOMY**
> A pilot should prove that authority can be constrained and evidenced before it proves that the agent can do more.

09 / PUBLIC WHITE PAPER

Executive and procurement decision guide

Buyers should evaluate the governed boundary, not the sophistication of the demo.

Area	Buyer question	Evidence to request	Red flag
Authority	Who is permitted to approve, override, or stop each action?	Named roles, scoped rights, revocation path.	The model approves its own action.
Mediation	Can any consequential route bypass the control point?	Architecture map and negative-path tests.	Some tools connect directly for speed.
Scope	What data, tools, destinations, and time windows are allowed?	Least-privilege matrix and credential policy.	Shared broad credentials or open-ended tools.
Policy	Which policy version governed a specific decision?	Versioned policy state and change control.	Policy is a prompt with no release control.
Evidence	What proves why the action was permitted or refused?	Decision record with inputs, authority, policy, and outcome.	Logs show activity but not authorization.
Human control	Can an accountable person intervene before consequence?	Risk-tiered approval, safe stop, escalation ownership.	Human review exists only after execution.
Testing	Are holds, blocks, bypasses, and degraded modes tested?	Signed test matrix with limits and unresolved findings.	Only happy-path demonstrations are shown.
Readiness	What exact evidence supports production acceptance?	Scope-specific acceptance decision and operational owner.	Implemented or measured is called production-ready.

HOW WE PRICE THE PROBLEM

The economic unit is not a token. It is the governed boundary: the consequence controlled, evidence required, integration burden, assurance level, service obligation, and cost of failure. Early engagements should be fixed around a boundary assessment or bounded pilot; broader licensing follows validated scope.

NIST distinguishes security function from assurance - what a control does versus the confidence that it works as intended. Procurement should price and test both. [3]

10 / PUBLIC WHITE PAPER

Choose the next conversation

A reader's responsibility should determine the next conversation—not a generic nurture sequence.

Reader	Calendar option	Time	Expected outcome
Executive / Risk	Governance decision briefing	30 min	Decision ownership, consequence boundary, pilot or no-go recommendation.
Technical / Security	Architecture boundary review	45 min	Integration surface, authority separation, control and evidence requirements.
Government / Regulated Pilot	Bounded pilot scoping	45 min	Mission workflow, human release authority, acceptance evidence, exit criteria.
IP / Licensing	Portfolio and field-of-use discussion	30 min	Strategic fit, diligence path, technical-exchange controls, commercial route.

Before the meeting

Executive Bring the workflow with the highest material consequence and the person who owns that risk.

Technical Bring a system boundary diagram, tool list, identity model, and the current approval path.

Government Bring the mission owner, procurement route, evidence obligation, and a reversible pilot candidate.

Licensing Bring the intended field of use, deployment layer, integration capability, and diligence questions.

> **YOUR DATA. YOUR CHOICE.**
> This paper should be delivered immediately. If you consent to follow-up, your work email, role, organization, and optional use case should be used only to offer the conversation most relevant to your responsibility.

Closing position

Agentic AI can expand organizational capacity, but capability is not authority. The durable governance question is whether a proposed action can cross into the world without a separately accountable, policy-bound, evidence-bearing decision. Govern that transition before consequence.

11 / PUBLIC WHITE PAPER

Sources and public disclosure note

Official and primary sources accessed 23 July 2026. Source inclusion does not imply endorsement of Ethic Vault.

[1] NIST AI Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology, 2023. Official source

[2] Artificial Intelligence Risk Management Framework: Generative AI Profile (NIST AI 600-1). National Institute of Standards and Technology, 2024. Official source

[3] Security and Privacy Controls for Information Systems and Organizations (SP 800-53 Rev. 5). National Institute of Standards and Technology, current publication page. Official source

[4] OWASP Top 10 Risk and Mitigations for LLMs and Gen AI Apps, 2025. OWASP GenAI Security Project. Official source

[5] LLM06:2025 Excessive Agency. OWASP GenAI Security Project. Official source

[6] Guidelines for Secure AI System Development. UK National Cyber Security Centre and international partners, 2023. Official source

[7] Thinking Carefully Before Adopting Agentic AI. UK National Cyber Security Centre, 15 May 2026. Official source

[8] Regulation (EU) 2024/1689 - Artificial Intelligence Act, Articles 14 and 15. Official Journal of the European Union. Official source

[9] AI Insights: Agentic AI. UK Government Digital Service, updated 13 March 2026. Official source

[10] Thematic Review and Gap Analysis on AI Security. UK Department for Science, Innovation and Technology, 10 July 2026. Official source

PUBLIC DISCLOSURE BOUNDARY

This document is designed for unrestricted website distribution. It contains no source code, credentials, cryptographic material, customer information, raw internal test artifacts, unpublished patent strategy, enabling implementation details, security weaknesses, detailed thresholds, registries, or bypass instructions. External facts are cited; Ethic Vault definitions and positions are identified as such.

Important: This white paper is informational and does not constitute legal, regulatory, cybersecurity, procurement, certification, or investment advice. Applicability depends on the use case, jurisdiction, and deployment environment.