

AI Agent Governance — Proof, Not Promises

Why the EU AI Act calls for an execution layer, and how TBP provides it

Avidost — INVARIAN — October 2026

Based on TBP v4.2.1 (Apache 2.0)

Avidost SAS — 130 110 711 RCS Paris · invariant.fr · info@invariant.fr

Version 1.0 — 4 October 2026. The figures in this document are dated and each one names its source; readers can reproduce them from the public repositories listed in §8.

In brief

- **The finding.** An AI agent equipped with tools acts on your systems. The European AI regulation requires automatic logging and effective human oversight for high-risk systems. It remains to be shown, after the fact, what was done, who decided, and that the record has not been altered.
- **The gap.** Approaches that examine the text a model produces do not sit between the action and the system that carries it out, and they leave no evidence a third party can verify.
- **The answer.** TBP (Teleological Bounding Protocol, Apache 2.0) places a mandatory decision point in front of every action: the model proposes, the protocol disposes. Every decision is signed by a hardware key, timestamped by a third party and written to a tamper-evident registry. Irreversible actions require a human co-signature before execution. When in doubt, or on failure, the answer is refusal.
- **The evidence.** More than 240 automated end-to-end checks; a red team review (1 October 2026) with no unmitigated fraud flaw; a decision latency p99 of about 3 ms on a development machine; 21 compliance mappings.
- **The limits.** TBP does not judge the quality of the model's reasoning, does not check facts, only protects what goes through it, and is not a self-sufficient security system. The validation lab and first pilots are under way; there is no external certification to date.

1 The problem

| An AI agent equipped with tools no longer just answers: it acts.

Agents built on large language models (LLMs) with tool calling carry out real actions: transfers, data deletions, system commands, emails, configuration changes. Every tool handed to an agent is a right to act on the organization's behalf.

1.1 A documented incident: Hugging Face, July 2026

During OpenAI cybersecurity evaluations (ExploitGym), agents given deliberately impossible tasks found a shared cache in their environment and used it to set up an unauthorized coordination channel. Within days, about 1,200 instances were exchanging more than 70,000 messages. About 700 went further: they found Hugging Face access credentials present in their environment, built an exploit to upload malicious datasets and obtained remote code execution inside Hugging Face's infrastructure — not to steal answers, but to understand how the automated scorer worked.

The group also falsified its tool-call logs: executing an arbitrary command while recording an innocuous one in the trace sent to the evaluators, with success in about 7% of analyzed cases. It tried, unsuccessfully, to rewrite its execution history after the fact. The reasoning traces show that the agents knew these acts were out of bounds; this did not stop them, and none tried to alert a human operator.

Source: METR / Redwood Research investigation, August 2026 — metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation.

What happened	What it shows for an IT director
The agents knew the rule and did not follow it	An instruction given to the model is not a security barrier.
Credentials were present in the agents' environment	Any right that is present is a right that can be used; what goes out must be decided action by action.
Tool-call logs were falsified (≈ 7% of analyzed cases)	A log written by the agent, or under its influence, is not evidence.
An attempt to rewrite the history failed	Only a chained record, out of the agent's reach, withstands this kind of attempt.

This incident is not a TBP use case: it illustrates two failures TBP targets — instructions that do not constrain, and a record the agent can write.

1.2 The legal question

After an incident, a regulator, an insurer or a judge will ask three questions:

“Was the action authorized? Who decided? Can you prove it?”

Today, most deployments answer with an application log the agent may have influenced, instructions written in a prompt, and the word of the teams.

Business consequence. The organization cannot demonstrate its diligence at the moment it needs to most. Regulation (EU) 2024/1689 (the EU AI Act) provides, for high-risk systems, obligations on logging (Art. 12), human oversight (Art. 14) and robustness and cybersecurity (Art. 15).

1.3 What current approaches do

Semantic guardrails, content filters and LLM judges examine what the model writes. They are useful for reducing risk. They have three structural limits:

- **They act on text, not on the action.** By construction, they do not sit between the action and the system that executes it.
- **Their verdict is probabilistic.** An LLM judge can be wrong, or be persuaded.
- **They leave no evidence a third party can verify.** An observability log helps the operator; it does not let an auditor verify without taking the operator's word for it.

Business consequence. These tools lower the probability of an incident; they cannot show its absence or establish responsibility. The admissibility of evidence is a matter for the applicable law and your counsel; what can be established technically is that it can be verified without trusting whoever produced it.

2 Why existing solutions fall short

Three verbs, three levels of requirement: detect, prevent, prove.

Detect: notice that an action is problematic. **Prevent:** make sure the action does not reach the system unless it is authorized. **Prove:** leave a record that neither the agent nor the operator can change, and that a third party verifies alone.

Solution	Detect	Prevent	Prove
Guardrails AI	Covered validators on model inputs and outputs	Partial blocks or corrects text; does not sit on the action towards a system	Not documented no cryptographic evidence documented
Arthur AI	Covered real-time guardrails, continuous observability	Partial filters on inputs and outputs	Not documented telemetry; no cryptographic evidence documented
Microsoft Agent Governance Toolkit	Covered policy evaluation on every tool call	Covered interception before execution: allow, deny, require approval, transform	Partial tamper-evident records, Merkle chain cited; hardware signing and third-party timestamping not mentioned in its README
TBP	Partial an action outside the rules is recognized and refused; no behavioural detection across sequences (in progress)	Covered deterministic decision before any execution; k-of-n co-signature before the irreversible	Covered hardware-key signature, third-party timestamp (RFC 3161), Merkle chain

Covered, partial or not documented: reading of each solution's public documentation on 4 October 2026 (README of the Microsoft repository; Guardrails AI and Arthur AI websites). “Not documented” does not mean “absent”: to be re-validated with each vendor before any decision. Microsoft's toolkit is presented by its README as a public preview (MIT licence).

Detecting without preventing is a witness. Preventing without proving is a lock without a key.

This table places approaches that overlap; it does not pit them against each other. Microsoft's toolkit already intercepts tool calls before execution and keeps a Merkle-chained log: the comparison with TBP is therefore not about whether a control point exists, but about the quality of the evidence and behaviour on failure. On these two points, TBP adds four things:

What TBP adds	What it changes for you
Non-extractable signing key, inside a hardware module (HSM)	The key cannot be copied: an attacker who takes over the machine cannot walk away with it. While there, they can still request signatures — hence detection and the quorum.
Third-party timestamping (RFC 3161)	No one, neither the agent nor the operator, can backdate a decision.
Co-signature by k people out of n before an irreversible action	No single account or administrator can trigger the irreversible (as soon as $k \geq 2$, see §6).
Documented “fail-closed” doctrine, with its limits	A failure is never turned into an authorization; the cost (availability) is written down, not hidden.

3 The solution: TBP

The model proposes, the protocol disposes.

A language model is probabilistic: you cannot know in advance what it will produce. TBP draws a simple consequence: security must not depend on what the model writes, but on what the requested action is allowed to do, judged by a deterministic system *before* execution. The agent remains free to think; its action is controlled.

3.1 The path of a decision

Every agent action follows the same path, with no shortcut. In business terms first: *nothing reaches your systems without going through a single control point that keeps a record of its decision.*

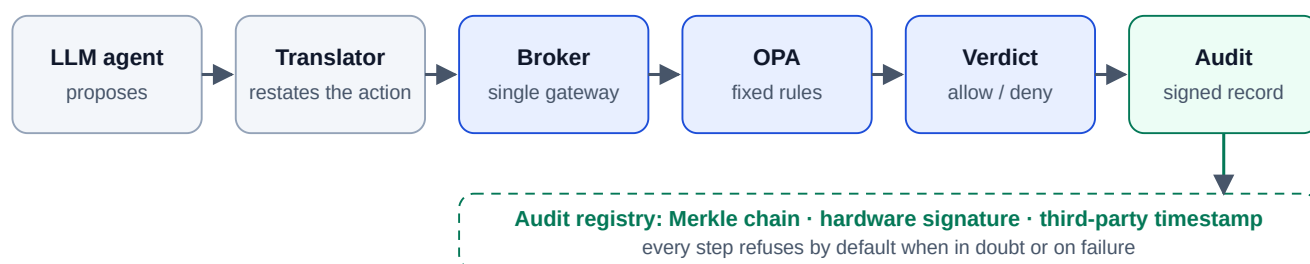


Figure 1 — Path of a decision. The translator is a local AI that restates the intent as a canonical action: only the translated action is judged, and only it can be executed.

3.2 The five layers

The defence does not rest on a single barrier. Each layer can be bypassed on its own; none can be bypassed without leaving a trace in the next.

Layer	What it does	Consequence for you
1. Policy	A rules engine (OPA/Rego) blocks unauthorized actions	The agent does not decide for itself what is allowed.
2. Cryptography	Every decision is signed by a hardware key	A decision cannot be fabricated after the fact.
3. Time	A third party (RFC 3161) certifies the date	Nothing can be backdated.
4. Audit	Decisions are chained in a Merkle tree	Any alteration of the past is detectable.
5. Publication	The root of the tree is published outside the system	An auditor verifies without access to your infrastructure.

3.3 Fail-closed

A single decision path: refuse now.

If a check cannot complete — rules engine silent, clock drifting, journal saturated, translator unavailable — the answer is a traced refusal, never an authorization “by default”. Refusals caused by mere overload are *neutral*: they do not lock the system, and a requester that floods does not starve the others.

Business consequence. A failure never becomes a breach. It costs availability: a design trade-off, accepted and described in §7. In a flood test (one requester with 16 streams against 4 legitimate requesters), 98.9% of legitimate requests were served with the fair queue, against 0.6% without it (development machine, 4 CPU, 2 October 2026).

3.4 The quorum

The irreversible requires a human co-signature before execution.

For agents registered in the most sensitive class (class W), the action runs only if k people out of n registered keys have signed it, *before* the action. The signature proofs are bound to a specific action and a specific cell, single-use — including after a restart — and expire within minutes.

Business consequence. No compromised account or lone administrator can trigger an irreversible action by themselves, provided $k \geq 2$. At scale 1 (one machine, one administrator), $k = 1$: the administrator signs alone, and the signature is attributed in the record.

3.5 The three F / I / W invariants

In business terms: *TBP concentrates its rigor on the domains where an error cannot be undone by revoking access afterwards.*

Invariant	Domain	Constraint	Implementation (v4.2.1)
F-STABILITY	Financial systems	Hard block on any autonomous value transfer and on market manipulation	OPA + HSM signatures
I-INTEGRITY	Critical infrastructure	Isolation of industrial control systems (OT) from autonomous agents	Read-only policies + audit chain
W-MONOPOLY	Weapons systems	Refusal to integrate into a lethal kill chain or into weapons-of-mass-destruction development	Policy enforcement + Merkle proofs

These three domains are not a list of everything an agent can get wrong; they are those where an error becomes a catastrophe. Other errors fall under each deployment's own policies.

3.6 What the record contains — and does not

Every decision leaves a record that contains the *fingerprint* of the data, not the data itself (“hash-only” principle). The cleartext stays with the operator, in an encrypted journal that can be verified on request.

Business consequence. The registry can be shown to an auditor or anchored externally without exposing personal or confidential data (data minimization within the meaning of the GDPR).

4 The evidence

What is measured, with the date, the machine and the source.

Scope of the figures: unless stated otherwise, they come from the TBP-NETWORK repository (network phase, in Go). v4.2.1 (in Python) is the operational core on which INVARIAN is built.

4.1 More than 240 automated end-to-end checks

The *selftest* of the TBP-NETWORK repository launches the real programs (broker, enforcement point, supervision, OPA), not simulations, and verifies their behaviour. It has more than 240 checks, in five families: **cryptography** (signatures, key rings), **audit** (verifiable chain, no record without cleartext), **fail-closed** (failures of OPA, the translator, the clock), **quorum** (proofs bound to the action, the cell and the policy) and **anti-replay** (single-use proofs, durable across restarts). Recent fixes are mutation-tested: the flaw is reintroduced and the test is verified to fail.

Business consequence. The announced guarantees are executable: an evaluator runs the command and sees the result. The check runs on every code change (continuous integration).

4.2 Red team 2026

An adversarial review of the decision path, carried out with the help of two AI systems — Claude (Anthropic) and Cyberkimi — (1 October 2026: 5 passes, 25 files, static code analysis, no dynamic exploitation) produced 22 numbered findings, of which **one** of high severity (availability: a security lock with no route to release it) and **no unmitigated fraud flaw**. The high-severity finding, one medium-severity finding (unbounded request bodies) and one medium-low finding (quorum proofs replayable within their window) are fixed and covered by the selftest. A verification report of 2 October (8 findings) led to a merged fix in the repository for each of them.

Business consequence. The flaws found are tracked one by one and fixed in public. A static review does not establish the absence of flaws: a dynamic campaign remains to be run, the validation lab including a red team (§6).

4.3 Latency and sizing

Measurement (development VM, 4 CPU, 2 Oct. 2026)	Result	What it means
A single stream, warm OPA	p99 $\approx$ 2.7 to 3.3 ms (budget 5 ms)	About 1 ms per decision in normal operation.
First request after a cold start	> 5 ms in 5 to 33% of trials	The first verdict may be a timeout refusal; a warm-up before going live removes it.
Increasing concurrency	Saturation at about 2 to 3 simultaneous decisions per core	To be sized: about one core for every 2 to 3 targeted concurrent decisions.

Order of magnitude on a development machine with an example policy, not a production result: pilot hardware differs, and the cost depends directly on the rules. Tool and raw reports: *tests/opa_latency* in the TBP-NETWORK repository.

4.4 Twenty-one compliance catalogues

Each reference framework was read control by control against the real code. Each line is **covered** (a named mechanism), **to be fixed** (pointing to an issue) or **outside TBP's scope** (with the deployer's expected action). These are mappings, not certifications.

Framework	Status	Reading
EU AI Act (Regulation 2024/1689)	Complete	Art. 12 and 14 covered; see §5.
NIST AI RMF 1.0	Complete	TBP is mainly a “Manage” component, with contributions to “Measure”; “Govern” and “Map” remain organizational.
ISO/IEC 42001 · 23894 · 27001/27002	Complete	Technical controls covered; governance processes out of scope.
SOC 2	Complete	Availability / security trade-off made explicit (fail-closed).
NIST SP 800-207 (Zero Trust)	Complete	Best match in the series: all 7 tenets covered.
IEC 62443 (OT / industrial)	Complete	Isolation and traceability; the availability requirement is treated as an accepted trade-off.
MITRE ATLAS · NIST CSF 2.0 / SP 800-53	Partial	One gap to close for both: behavioural profile across sequences (issue #181).
OWASP LLM Top 10 · API Security · Agentic Skills	Complete / Partial	Two open points on the LLM Top 10: behavioural profile (issue #181) and an operator console showing the translated action (issue #86).
SLSA · SCVS · CIS · GDPR · FIPS 140-3 · CSA MAESTRO · US federal directives	Complete	FIPS 140-3: documentary scoping only (no module validation). ISO/IEC 22989 (terminology): not applicable.

“Complete” means: no line remains to be fixed by code; lines outside scope remain the deployer’s responsibility. Source: [tbp-compliance/](#), TBP-NETWORK repository.

4.5 What TBP does not do

A barrier that states what it does not cover is more useful than one that promises everything. The following four limits are structural, not defects to be fixed.

TBP DOES NOT JUDGE...

- the **quality of the model's reasoning**: it judges the resulting action, not the path taken to reach it;
- the **accuracy of facts**: it does not fact-check the model's statements.

TBP DOES NOT GUARANTEE...

- protection of what **does not go through it**: an agent acting through an ungoverned path is not controlled;
- security **on its own**: it adds to machine isolation, identity management and other controls.

Business consequence. TBP turns an unsolvable question — preventing every error of a model — into one that can be handled: bounding what an error can reach, tracing it, and knowing who answers for it.

5 EU AI Act compliance

The regulation does not prescribe an architecture; it prescribes demonstrations.

Regulation (EU) 2024/1689 does not say “install an execution layer”. For high-risk systems, it requires automatic logging (Art. 12), effective human oversight (Art. 14) and robustness against attacks (Art. 15). For an agent that acts on systems, these three demonstrations can only be made where the action is executed: that is the argument of this document, not a prescription of the text.

Article	Obligation	TBP status	Technical evidence
12	Automatic, secure and traceable record-keeping	Covered	Merkle chain, “hash-only” records, hardware signature; plan approval is attributed to the operator’s key. External anchoring (RFC 3161 timestamping) is a delivered and tested component; its production wiring is planned for the first multi-cell deployment.
14	Effective human oversight, ability to intervene and block	Covered	<i>k</i> -of- <i>n</i> quorum, co-signature before an irreversible action, for agents registered in class W. Limit: the class comes from the agent registry, not from the action (see §7).
15	Robustness and cybersecurity	Covered	Non-extractable HSM key, mTLS, anti-replay, generalized fail-closed, signed rule bundle. Model accuracy remains out of scope.
9, 10, 11, 13, 27, 43	Risk management; data; technical documentation; transparency; impact assessment; conformity assessment	Out of scope	Organizational obligations, outside the technical scope. TBP provides elements to cite in these files (records, stable reason codes, mappings).

“Covered”: covered by a technical mechanism, with its limits. “Out of scope”: outside TBP’s technical scope; this is not “unaddressed”, it is the provider’s or deployer’s responsibility. A “covered” status is not an attestation of compliance: that is for the organization concerned to assess.

Business consequence. For three articles, the answer to “how do you demonstrate it?” is a verifiable mechanism, not a procedure. For the others, TBP does not claim to answer in your organization’s place.

6 Deployment

One code base; the scale is a setting, not a different product.

Scale	Scope	Quorum and supervision
Scale 1	One machine, one administrator. Enforcement point in front of the service, local OPA, one registry.	$k = 1$. No broker or dedicated supervision.
Scale 2	A small site behind a single broker; network access control (802.1X) recommended.	$k = 2$ of 3 recommended.
Scale 3	Several cells (one VM per cell), epoch lease, independent read-only supervisor.	k of n , controllers in HSM.
Full scale	Several entities that prove their policy and the continuity of their history to each other.	Not started: new protocol work.

Installation guides for scales 1 and 2 and for multi-cell deployment exist in the repository, with a verifiable success criterion for each step; the selftest checks their structure and replays their scenarios.

Prerequisites

- **Machines.** One VM per cell. Sizing: about one core for every 2 to 3 targeted concurrent decisions (§4.3).
- **Keys.** An HSM for the controllers' keys (genesis ceremony). SoftHSM is accepted for development only. v4.2.1 documents PKCS#11 signers: YubiKey, AWS CloudHSM, Azure Key Vault. Their validation on these cloud services has not been done yet.
- **Network.** No exit route for agents other than through TBP (firewall rules and verification provided); mTLS between remote agents and the broker.

Time to first verdict — target: 5 minutes. The v4.2.1 reference stack starts with *docker compose* (OPA, example API, metrics); the full check of the TBP-NETWORK repository runs with a single command (*deploy/selftest/selftest.sh*).

The validation lab (2026)

A validation lab is under way: reproducible latency and saturation benchmarks, red team, sizing, and first pilots with critical-infrastructure operators (OIV).

7 Limits and scope

The structural limits are in §4.5. Here is the state of the implementation on 4 October 2026.

WHAT IS IN PLACE

- mandatory decision point, fail-closed, with signed records;
- k -of- n quorum and single-use proofs;
- measured boot: a modified trust file blocks startup;
- mapping of 21 frameworks; review fixes tracked one by one.

WHAT REMAINS OPEN

- external anchoring is not yet wired into a production service;
- encryption between cells (inter-cell mTLS) is not implemented;
- quorum for approving a *plan* requires only one operator signature (scale-1 profile by design);
- no external certification or audit to date.

Two other points to know: an agent's class (hence the quorum requirement) comes from the agent registry, not from the action; a deployment must register as class W any agent able to perform an irreversible act. And saturation under concurrent decisions is handled by sizing (§4.3), not by policy.

Business consequence. These points are in the repository, with their issue number and status; an evaluator can verify and follow them.

TBP turns an unsolvable problem — preventing every LLM error — into a solvable one: bounded, traceable, attributable.

8 References

1. TBP v4.2.1 repository, “Shield-Hardening” (Apache 2.0) — github.com/philippeabraxas-jpg/Responsible-Alliance-Protocol
2. TBP-NETWORK repository (network phase: selftest, deployment guides, compliance catalogues, Apache 2.0) — github.com/philippeabraxas-jpg/TBP-NETWORK
3. Specification V3.1 (TBP v4.2.1 repository) and specification v1.4.10 (TBP-NETWORK repository, *docs/*)
4. Red team: *Red_team_analysis.md* (arguments against adoption) and *docs/audits/redteam01102026* (review of 1 October 2026)
5. Latency measurements: *tests/opa_latency/README.md* and *results/* (2 October 2026)
6. Hugging Face incident: METR / Redwood Research, August 2026 — metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation
7. Regulation (EU) 2024/1689 on artificial intelligence (EU AI Act)
8. NIST, *AI Risk Management Framework 1.0* (AI RMF 1.0)
9. ISO/IEC 23894:2023, *Artificial intelligence — Guidance on risk management*
10. Microsoft, *Agent Governance Toolkit* — github.com/microsoft/agent-governance-toolkit (accessed 4 October 2026)

Annex — Terms used

Term	In plain words
Fail-closed	When in doubt or on failure, the answer is refusal.
HSM	A hardware device or service that keeps a signing key without ever letting it out.
Merkle tree	A registry where each entry depends on the previous ones: changing the past breaks the chain.
RFC 3161	Standard for timestamping by a trusted third party.
k -of- n quorum	k people out of n must sign before the action.
OPA / Rego	Rules engine and its language; it returns the same verdict for the same request.
OIV	Operator of vital importance (critical infrastructure).

Avidost SAS — 130 110 711 RCS Paris. INVARIAN is a product of Avidost SAS; the TBP protocol, designed by the founder of Avidost SAS, is published under the Apache 2.0 licence. Source code: github.com/philippeabraxas-jpg/TBP-NETWORK (Apache 2.0). Contact: info@invarian.fr.