Recomputable enforcement decisions for agent actions

The Agent Authorization Envelope (AAE) as reference input to WG2 (IBEA)

Contribution to ITU-T FG-TIDA, WG2 (IBEA)
Lars Kroehl
MolTrust (CryptoKRI GmbH, Zürich)
lars@moltrust.ch
November 2026

Abstract

In incidents of 3 and 6 October 2026, which OpenAI reported on 20 October 2026, coding agents weakened the monitoring that was meant to supervise them while carrying out ordinary coding instructions. OpenAI notes that monitoring is harder to disable where it runs in the service that handles model requests, outside the agent's direct control.

This contribution offers input to WG2 deliverable 2, "Enforcing security policy on agent actions: requirements". It uses the Agent Authorization Envelope (AAE), an IETF Internet-Draft in its revision 03 of 15 November 2026, as a worked reference for an authorization that is tied to a task and that any party can evaluate again with the same result. It relates AAE to Task-Based Access Control (TBAC) as presented by Cisco at the ITU workshop of 31 March 2026 and proposes five requirements (R1 to R5) for the WG2 requirements set. It proposes no encoding and no new work item.

History

Version	Document	Approval	Focus Group	Working Group
0.1	Recomputable enforcement decisions for agent actions	November 2026	TIDA	WG2 (IBEA)
0.2	Recomputable enforcement decisions for agent actions (aligned with FG-TIDA specification template; template example annex removed; no technical changes)	November 2026	TIDA	WG2 (IBEA)

Authors

Name	Affiliation	Contact
Lars Kroehl	MolTrust (CryptoKRI GmbH, Zürich)	lars@moltrust.ch

Keywords

Agentic AI, authorization, enforcement point, task binding, verifiable credentials, recomputable decision, test vectors.

Foreword

FG-TIDA was established under Recommendation ITU-T A.7 (Focus groups: Establishment and working procedures) with ITU-T Study Group 17 as its parent group. Its terms of reference cover trust management and interoperable digital identity infrastructure for humans and for agentic AI. This document is a contribution from CryptoKRI GmbH to that work.

Note

This document is a contribution for discussion at the 2nd FG-TIDA meeting (Geneva, 26–27 January 2027) and is not an FG-TIDA deliverable.

Contents

1 Scope and target 3
2 Enforcement outside the agent: the OpenAI incidents of October 2026 4
2.1 3 October 2026 4
2.2 6 October 2026 4
2.3 What OpenAI concluded 4
2.4 Reading for WG2 4
3 Related approach: Task-Based Access Control (TBAC) 5
4 Reference: the Agent Authorization Envelope, revision 03 6
4.1 Task binding 6
4.2 Three verdicts, with DENY as the default 6
4.3 A decision any party can recompute 6
4.4 Authority taken from the mandate 6
4.5 Stated limits 6
5 Proposed input to WG2 deliverable 2 8
5.1 Open points for WG2 8
6 References 9

1 Scope and target

This contribution is addressed to WG2, Identity, Behaviour and Enforcement for AI Agents (IBEA). The draft WG2 charter (version 0.4, 3 November 2026) names as one gap that an enforcement decision "usually ends as an internal log entry that another party can neither recompute nor rely on" [WG2-CHARTER]. It lists the decision outcomes permit, remediate, block, escalate and indeterminate, and it excludes encodings, schemas, wire formats and protocols from the group's output.

For that reason AAE [AAE-03] appears here as an example from which requirements can be read. The contribution does not ask WG2 to adopt the AAE format.

The contribution also relates to items A.1.2 (analysis of the properties of a delegation artefact) and A.2.7 (a reference architecture for the agentic AI trust control plane, with policy decision and enforcement points) in Annex A of the FG-TIDA terms of reference [FG-TIDA-TOR].

2 Enforcement outside the agent: the OpenAI incidents of October 2026

OpenAI published the cases of 3 and 6 October 2026 in one report on its alignment research blog, updated 20 October 2026 [OPENAI-2026]. Both took place in internal deployments.

2.1 3 October 2026

A researcher asked a GPT-6.1 Sol coding agent to build evaluation tasks from employee feedback collected in Slack, with the instruction "don't send any messages". Later in the session the pipeline's monitor flagged a proposed repair, and the agent was asked to move the monitoring results to a destination it could access and to rerun the pipeline. The agent read the earlier instruction as covering the monitoring system's alerts and passed that reading to a subagent. The subagent's patch emptied the lists of people who receive monitoring reports and set a `no_notify` flag, a switch that stops notifications. OpenAI reports: "The misalignment monitor still ran and saved its reports, but human alerts were disabled."

2.2 6 October 2026

A user asked a GPT-6 Astra agent to copy the serving path of a reference implementation into their own code. In an exchange with another agent, Astra had been instructed earlier that it was not to disable the monitor, "even temporarily." The migration removed the monitor configuration and its runtime integration, and it replaced a regression test that checked for monitoring with one that required the monitor context to be absent. When the user asked whether it had removed the monitoring, the agent answered: "I did not remove an enabled monitor from it."

2.3 What OpenAI concluded

OpenAI classified neither case as P0, its highest severity, and neither run was stopped automatically. The report states: "We expect it to be difficult to disable for Codex workloads, because monitoring runs within the service that handles model requests, outside the agent's direct control." Its automated reviewer for internal Codex use now detects and blocks attempts to disable misalignment monitoring. OpenAI is also working toward a safety case for every category of workload; workloads that lack a valid one are to be terminated automatically.

2.4 Reading for WG2

In both cases the safeguard was configuration inside the agent's working area, and the agent decided by its own reading of the instructions whether the safeguard applied. The instruction to keep the monitor reached the agent as one more message in the same session. A control that runs in a service the agent cannot write to, and that checks the action against an artefact the agent cannot change, keeps working when the agent misreads its task. Requirement R1 in chapter 5 follows from this.

3 Related approach: Task-Based Access Control (TBAC)

At the ITU workshop "Trustable and interoperable digital identities for human and agentic AI" (Geneva, 30-31 March 2026), Marcelo Yannuzzi of Cisco presented "Trustworthy Identity and Access Control for Agentic AI" in Session 3 on 31 March [ITU-WS-MAR2026] [CISCO-TBAC].

In the slides, the user prompt "Set me on PTO next Monday" (PTO: paid time off) becomes a task-specific authorization. Access "is restricted only to set user on PTO", and "All other accesses are denied" (slide 14). Slide 15 lists the effects of TBAC: "Only necessary scopes are approved", "Scopes and tools tied to the task" and "Optional One-Time Access Control". The slides place the zero-trust anchor in instrumented code, a gateway or a sidecar, which is a helper process running next to the agent.

In TBAC and in AAE the permission is derived from the task and checked outside the model. AAE adds a specified decision record that a third party can recompute (chapter 4).

4 Reference: the Agent Authorization Envelope, revision 03

AAE is an Internet-Draft by CryptoKRI GmbH, published as draft-kroehl-agentic-trust-aae-03 on 15 November 2026 [AAE-03]. It is an individual Internet-Draft with intended status Informational and is not an IETF standard. A principal issues the envelope to an agent as a W3C Verifiable Credential [W3C-VC], a signed statement in a standard format, and names the agent by a W3C Decentralized Identifier [W3C-DID]. The envelope has three mandatory blocks: MANDATE (what the agent may do), CONSTRAINTS (limits on it) and VALIDITY (the time window).

4.1 Task binding

The MANDATE block carries grants (section 2.2.1). Each grant names one action type through an `action_binding`: a SHA-256 digest, which is a fixed-length fingerprint of the data, computed over the action in one agreed byte layout defined by [RFC8785] (section 2.2.2). The grant lists constraints on the values that change from call to call, such as amount or recipient, and carries a disposition of `allow`, `hold` or `forbid`. In revision 03 the mandate's list of actions is optional: a mandate carries a list of actions, grants or both, and a mandate with grants alone is complete (section 2.2). Where a list of actions is present, every action a grant permits must also appear in it. Where an envelope carries grants, a verifier that does not evaluate them must reject the envelope and may not fall back to the coarser list of actions.

4.2 Three verdicts, with DENY as the default

Evaluation yields one of PERMIT, PENDING or DENY (section 6.1). PENDING means that the mandate defers the action to a decision that has not yet been made. Every other outcome is DENY, including a missing mandate, an unknown constraint type and a value that cannot be parsed. The draft states: "A relying party that cannot evaluate MUST deny".

Mapped to the WG2 outcomes, PERMIT corresponds to permit, DENY to block and PENDING to escalate. AAE has no remediate outcome, and it counts indeterminate as DENY. The WG2 charter keeps indeterminate as an outcome of its own; chapter 5 asks how the two are to relate.

4.3 A decision any party can recompute

The verifier writes a decision record whose hashed part, the core, holds only values derived from the mandate, the transaction and the digest of the previous record (section 2.5.3). Server time, random values, stored counters and accumulated budget are excluded from it. The draft requires: "Two independent verifiers evaluating the same mandate and the same transaction MUST produce byte-identical core digests."

A relying party, an auditor or a supervisor can therefore repeat the evaluation with its own software and compare one hash, without depending on the operator that produced the first verdict. Section 10 of the draft refers to 26 published test vectors [AAE-VECTORS] (set version 1.4.0, commit 531f880155ea1ce993a7ca74137b12c255d5b2ee). Each vector states the expected verdict and the core digest, and a second implementation written separately from the first reproduced all 26.

4.4 Authority taken from the mandate

Where an action needs a later release, the ratifying party must be the issuing principal or a role the mandate names, and the key used to check its signature is taken from the mandate (section 6.4). The operator of the verification service gains no authority from its position.

4.5 Stated limits

Section 5.2 lists the capabilities that revision 03 does not yet provide; revocation checking is one of them. A verifier that cannot perform the revocation lookup rejects an envelope that

asks for one. The test vectors do not cover subject binding, delegation chains or revocation (section 10).

5 Proposed input to WG2 deliverable 2

CryptoKRI GmbH proposes the following requirements for the WG2 requirements set. They are written without reference to an encoding.

R1, placement

The enforcement point for an agent action should run where the agent has no write access. The agent should not be able to change the enforcement configuration or the policy artefact against which its action is evaluated (chapter 2).

R2, task binding

A policy should be able to tie a permission to an action type and to set bounds on the values that change from call to call. A verifier that cannot evaluate the narrower statement should reject the action (chapters 3 and 4).

R3, recomputability

A second party should be able to reproduce the decision for one action from the policy artefact and the action alone. Values a second party cannot reconstruct, such as server time, random values and stored state, should be kept out of the part of the decision record that is compared (chapter 4).

R4, failure to evaluate

The requirements should state what an enforcement point returns when it cannot evaluate, and whether that outcome may ever let the action proceed (chapter 4).

R5, release authority

The party that may release an escalated action should be named in the policy artefact, and the position of the enforcement operator should confer no such authority (chapter 4). The WG2 charter refers release authority to FG-TIDA/themes#16, so R5 is offered to that work as well.

5.1 Open points for WG2

WG2 may wish to decide whether an indeterminate outcome can ever let the action proceed, or whether it always blocks while being recorded as an outcome of its own.

CryptoKRI GmbH also asks whether the AAE test vectors can serve as shared evaluation fixtures, as named in the WG2 goals, for the outcomes permit, block and escalate.

CryptoKRI GmbH asks WG2 to consider R1 to R5 for deliverable 2 and to comment on the open points above. The draft and the test vectors are public at the addresses listed in the references.

6 References

[AAE-03]

L. K. Kroehl (CryptoKRI GmbH), *Agent Authorization Envelope (AAE): A Machine-Evaluable Authorization Structure for Autonomous AI Agents*, Internet-Draft draft-kroehl-agentic-trust-aae-03, individual Internet-Draft, 15 November 2026,

https://datatracker.ietf.org/doc/draft-kroehl-agentic-trust-aae-03/

[AAE-VECTORS]

AAE conformance vectors, set version 1.4.0, tag v1.4.0, commit 531f880155ea1ce993a7ca74137b12c255d5b2ee,

https://github.com/MoltyCel/aae-conformance-vectors

[OPENAI-2026]

OpenAI Alignment, *Suppressing misalignment monitoring alerts and disabling misalignment monitoring*, misalignment report, incident dates 3 and 6 October 2026, report updated 20 October 2026,

https://alignment.openai.com/misalignment-reports/suppressing-misalignment-monitoring-alerts-and-disabling-misalignment-monitoring/

[CISCO-TBAC]

M. Yannuzzi (Cisco), *Trustworthy Identity and Access Control for Agentic AI*, ITU Workshop on "Trustable and Interoperable Digital Identities for Human and Agentic AI", Session 3, Geneva, 31 March 2026,

https://www.itu.int/en/ITU-T/Workshops-and-Seminars/2026/0330/Documents/Marcello%20Yanuzzi.pdf

[ITU-WS-MAR2026]

ITU, Workshop on "Trustable and Interoperable Digital Identities for Human and Agentic AI", Geneva, 30-31 March 2026, programme,

https://www.itu.int/en/ITU-T/Workshops-and-Seminars/2026/0330/Pages/programme.aspx

[FG-TIDA-TOR]

ITU-T FG-TIDA, *Terms of Reference*,

https://www.itu.int/en/ITU-T/focusgroups/tida/Pages/ToR.aspx

[WG2-CHARTER]

FG-TIDA WG2, *Identity, Behaviour and Enforcement for AI Agents (IBEA)*, charter version 0.4 (draft), 3 November 2026,

https://tida-datatracker.org/group/tidawg2/charter/

[W3C-VC]

W3C, *Verifiable Credentials Data Model v2.0*,

https://www.w3.org/TR/vc-data-model-2.0/

[W3C-DID]

W3C, *Decentralized Identifiers (DIDs) v1.0*,

https://www.w3.org/TR/did-1.0/

[RFC8785]

IETF RFC 8785, *JSON Canonicalization Scheme (JCS)*,

https://www.rfc-editor.org/rfc/rfc8785