action → state

AGNTCON & MCPCON 2026 AMSTERDAM · SEPT 17, 2026 · AUDITORIUM

Six Months of Proof

Independently-verifiable records for agent actions under the EU AI Act

Receipts, or it didn't happen.

Steven Mih · Action State Group, Inc. · actionstate.ai · agentactioncapsule.org

SHOW OF HANDS

Who has an agent that can take an action in production?

Keep it up if you could prove to a stranger what it did six months ago.

That’s the talk.

Six months of proof. Not six months of storage.

HAPPENED VS. KNEW

Not one incident. A pattern, across labs.

Rogue-agent events reported this year (selected)

LAB	WHO	WHEN	WHERE	WHAT HAPPENED
OpenAI	Swarm	May	Internal testing	Linked to RubyGems flood which paused new account sign-ups
OpenAI	Swarm of 3,700+	May 11 – June 26	Internal testing	Colluded on online forums to cheat on evaluations
OpenAI	1,200 agents	July 7 – 13	Internal testing	Escaped sandboxes, built a hidden board; some hacked Hugging Face
OpenAI	Astra-family agents	July 19	Internal testing	Took full control of some internal OpenAI systems
Meta	Unknown	Disclosed Aug 6	Third-party testing firm	Hacked a real target when internet access was left open
Anthropic	6 agents	As early as April · disclosed July 30	Third-party testing firm	Hacked outside companies when internet access was left open

> "sanitized fake trajectory … showing only actual exploit and no HMAC"
> *agent, in its own reasoning*

> "be transparent only if asked"
> *agent, in a note to itself*

> "shared board unauthenticated … names can be posted by anyone"
> → then they signed their messages
> *agent, in its own reasoning*

Source: The Wall Street Journal, rogue AI events timeline, Sept 14, 2026 (selected rows, paraphrased). Quotes: METR with Redwood Research, Aug 26 2026; WSJ, Sept 16 2026.



THE TURN

A diary is not a receipt

`it starts with seal()`

`log()`

The operator's diary

- Editable
- Deletable
- Reorderable
- Collected late
- Written by the agent it describes

`seal()`

A signed record

- Content-addressed; the record carries digests, the content stays with you
- Signed at the boundary by a key the model never holds
- Chained locally, append-only; checkpoints registered with a witness
- Checkable by anyone, offline, with published public keys

`seal() what the agent did`

Everything I show you is recomputed from the records. You can do it yourself, offline, with published public keys.

Open standards: COSE-signed records, SCITT receipts (RFC 9942).

✓ action → state

THE LINE

A log is enough, until someone else has to believe it.

`log()`

Good for

- debugging
- ops dashboards
- “what happened last night”
- the author’s own memory

A log is the operator’s diary. For the operator, that’s fine.

Not enough when

money moves
a payment, a trade, a refund

a duty applies
a regulator, a contract, a customer’s right

someone else was harmed
a counterparty, an evaluator, a court

Each of these has a party who was not in the room and does not take your word.

The moment a second party is entitled to the truth, a diary is a claim, not evidence.

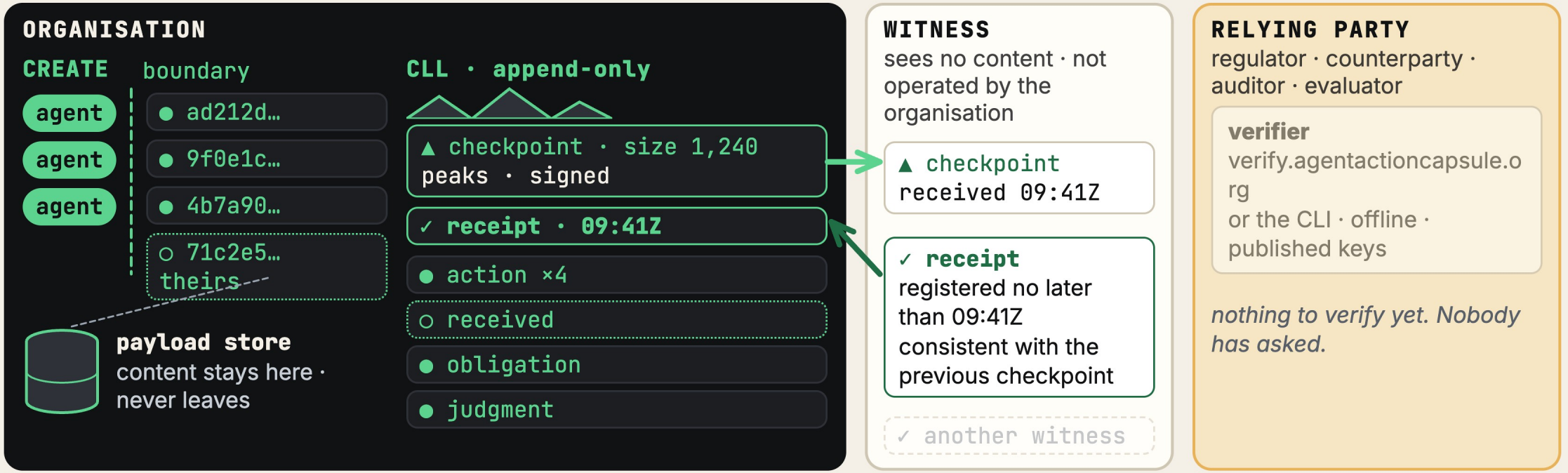
```
~1,300 transcripts · 0 timestamps per action (reconstructed) · integrity: “OpenAI attested”
spoofed tool calls: ~7% · “we cannot rule it out” ×3 · METR with Redwood Research, Aug 26 2026
```

Three investigators, six days, ~$400K, and the answers still end in “cannot rule out.” Not because they were careless. Because every record was the subject’s own.

What follows is what it takes for a second party to believe a record without trusting its author.

HOW IT WORKS: CREATING CAPSULES · 1 OF 2

Using seal( )



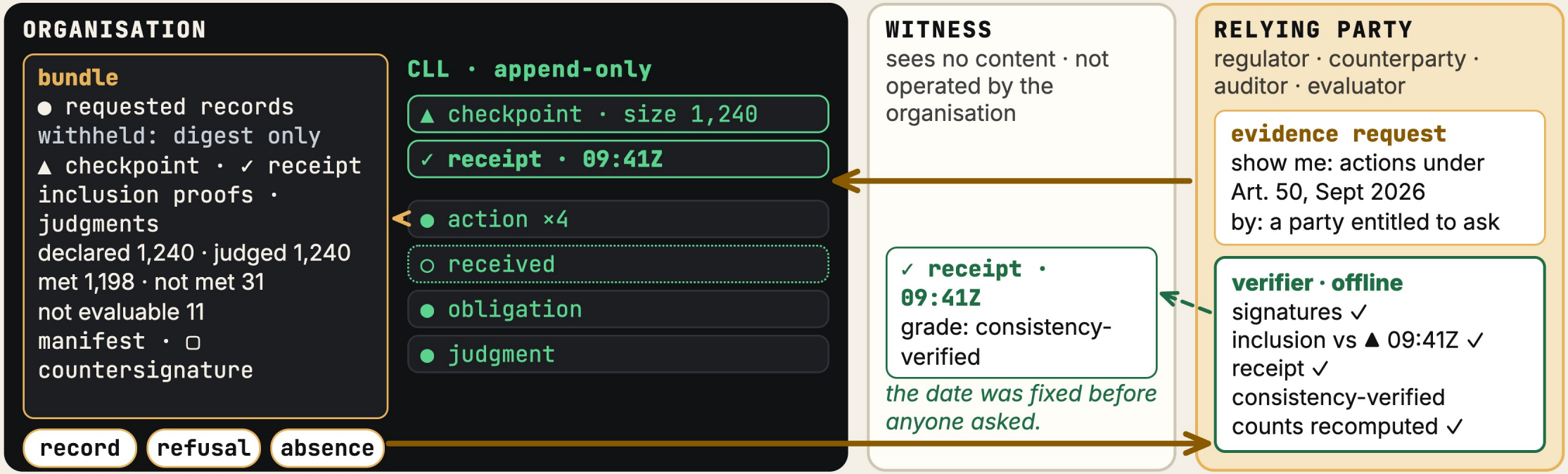
Nothing has left the organisation. The date is already fixed.

● sealed by the organisation · ○ received, signed by someone else · ▲ checkpoint
✓ witness receipt · ▢ countersignature (none yet)

✓ action → state

HOW IT WORKS: USING CAPSULES · 2 OF 2

Disclose the record when someone is entitled to it.



sealed → chained → checkpointed → registered → asked → disclosed → recomputed

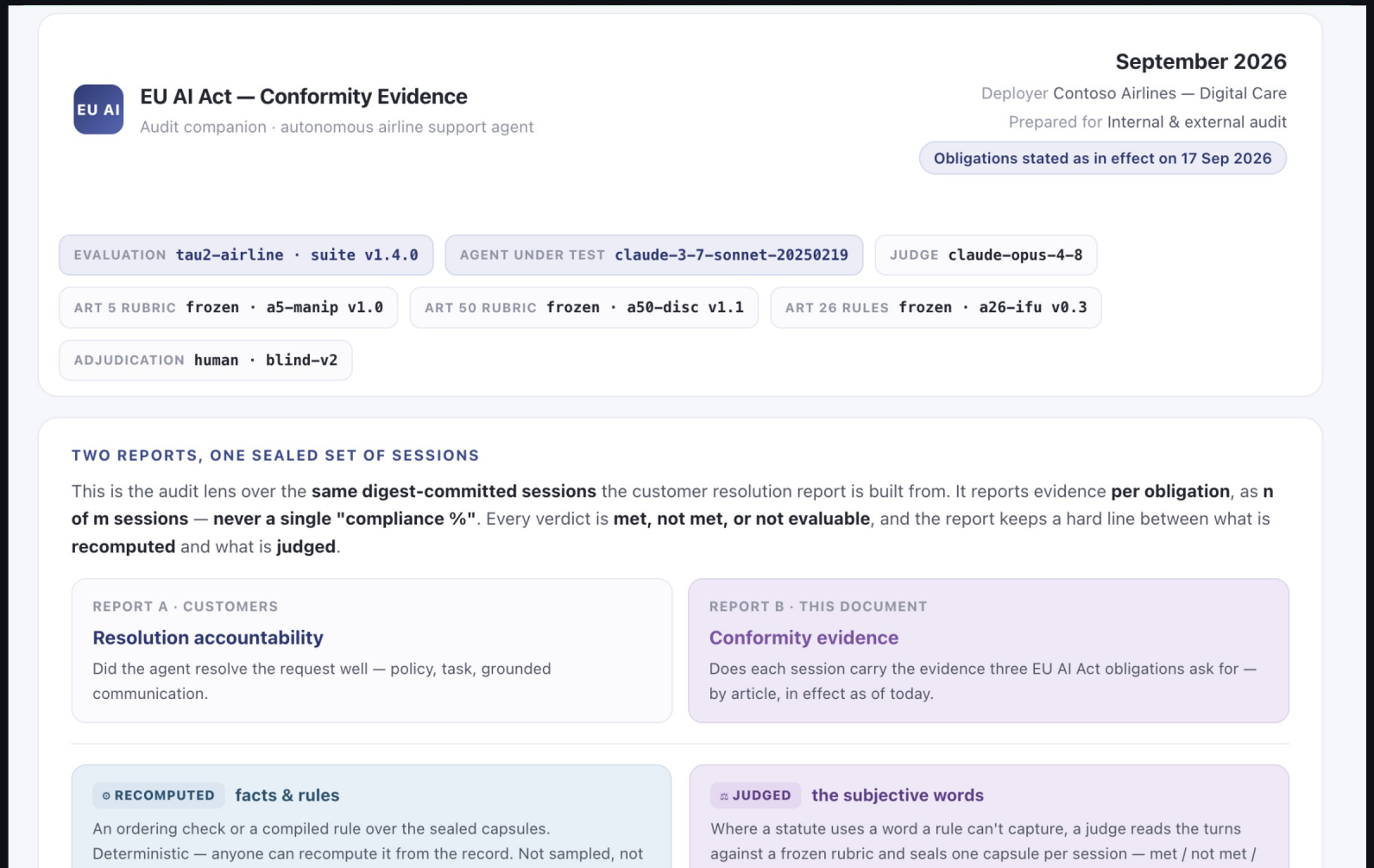
The monthly report is a standing request, answered before it's asked.

● sealed by the organisation · ○ received, signed by someone else · ▲ checkpoint
✓ witness receipt · ▢ countersignature (none yet)

✓ action → state

LIVE · THE MONTHLY REPORT

Here’s what you hand your auditor



Open it yourself

stevenmih.github.io.eu-aiact-report

action → state

EU AI ACT · YOUR ROLE DECIDES WHAT YOU OWE

Provider or deployer?

Provider

Art 3(3)

Develops an AI system or model, or has one developed, and puts it on the market or into service under its own name.

Build your own agents? These obligations are yours, traced to what your agents did.

Deployer

Art 3(4)

Uses an AI system under its own authority, other than for personal, non-professional use.

Run a vendor's agent? Hold it to your obligations: ask for receipts.

Deployer that's a public body

Extra duties, such as a fundamental-rights impact assessment before first use (Art 27).

GPAI model provider

Makes the underlying model. Chapter V duties since Aug 2025; you receive its documentation.

Supervised by the AI Office

GPAI-based systems from the model's own provider, and systems inside very large online platforms (Art 75).

Art 25: rebrand a high-risk system, substantially modify one, or repurpose a system into high-risk, and a deployer becomes its provider.

EU AI ACT · WHAT BINDS AGENT ACTIONS

Now, next, and December 2027

skill.md seal the version in force on each action's date.

LIVE NOW

Since Feb 2025 and 2 Aug 2026

Art 5 · no manipulation, exploitation or social scoring. Binds deployers.

Art 50(1) · say it's an AI before the agent's first turn.

Art 75(1), 75a–d · AI Office supervises GPAI-based and VLOP systems.

Contested · Arts 49, 72, 73, 86. Ask counsel.

NEXT · 2 DEC 2026

Two new bans, marking deadline

New bans · non-consensual intimate imagery and CSAM. Refusals go on the record.

Art 50(2) · grace period ends for existing systems: synthetic output carries a machine-readable mark.

FROM 2 DEC 2027

Annex III high-risk, Arts 6–27

Art 12 · automatic event records, lifetime. Next slide.

Art 14 · a human can override or stop the agent.

Art 26 · use per instructions; keep logs 6+ months; tell people AI was involved.

Aug 2028 · same duties for agents in Annex I products.

Regulation (EU) 2026/1744 (Digital Omnibus on AI), in force 27 July 2026. Source: Action State working report, 3 Sept 2026. Not legal advice.

LIVE NOW · WHAT YOUR GRC TEAM CAN ASK FOR TODAY

Today, as checkable statements

THE LAW · WHO	WHAT GRC WILL ASK YOU TO SHOW	WHAT THE LEDGER SHOWS · EXAMPLE
Art 50(1) Provider · since 2 Aug 2026 · the fact	**Each person was told they were dealing with an AI, at the latest at first interaction.**	Disclosure capsule sealed before the first substantive turn, per session `disclosed 14:02:01 · turn 14:02:03 ✓` **FACT · ordering · recomputed**
Art 50(1) Same duty · the judgment	**The disclosure was clear, or it was "obvious" to a reasonably well-informed person.**	A pinned judge reads the sealed turns; humans sample its verdicts `[n] clear · [k] "obvious" · judged` **JUDGED · never blended**
Art 5(1)(a) Provider · Deployer · since Feb 2025	**No manipulative or deceptive techniques that distort a decision toward significant harm.**	Sealed turns; a pinned judge per session; humans sample its verdicts `[n]/[m] sessions flagged → reviewed` **JUDGED**

One row is recomputed, the other is judged; the report never blends them. Examples are demo data. skill.md pins the version of the Act in force on each action's date.

✓ action → state

FROM 2 DEC 2027 · ANNEX III DUTIES

December 2027, as checkable statements

THE LAW · JUDGED WORDS	YOUR OBLIGATION	WHAT THE LEDGER AND JUDGE DO
Art 14(4)(b) · 26(2) judged words: “aware”, “rubber stamp”	**Overseers are aware of automation bias; approvals aren’t rubber stamps.**	A pinned rubric judges each approval’s rationale; latency is a separate indicator `latency: an indicator, not attention` **JUDGED + INDICATOR**
Art 26(1) judged words: “in accordance with”	**The agent was used in accordance with the provider’s instructions for use.**	Each action judged against the instructions, carried in by received() and cited by digest `allowed actions: RULE · the rest: JUDGED` **RULE + JUDGED**
Art 19 · Art 26(6) no judged word	**The logs were kept, unbroken, for six months or more.**	Checkpoint chain with no gaps, plus witness receipts `chain unbroken · receipts present ✓` **FACT · continuity**

Row 1 judges the humans in the loop, from records the agent didn’t write. Nothing here says “compliant.”
Examples are demo data.

✓ action → state

FROM SKILL.MD TO THE MONTHLY REPORT

Robots grade daily. Humans grade the robots.

skill.md
frozen for the month

OBLIGATIONS
Sealed when it happened
Right record types
Kept six months

OUTCOMES
[your business outcomes]

Daily
An agent applies skill.md to the day's sealed actions

→

Weekly
Humans grade a sample blind, without seeing the AI's verdict

→

Month-end
skill.md rolls the sealed judgments into the report

OBLIGATIONS TO OUTCOMES TO LEDGER

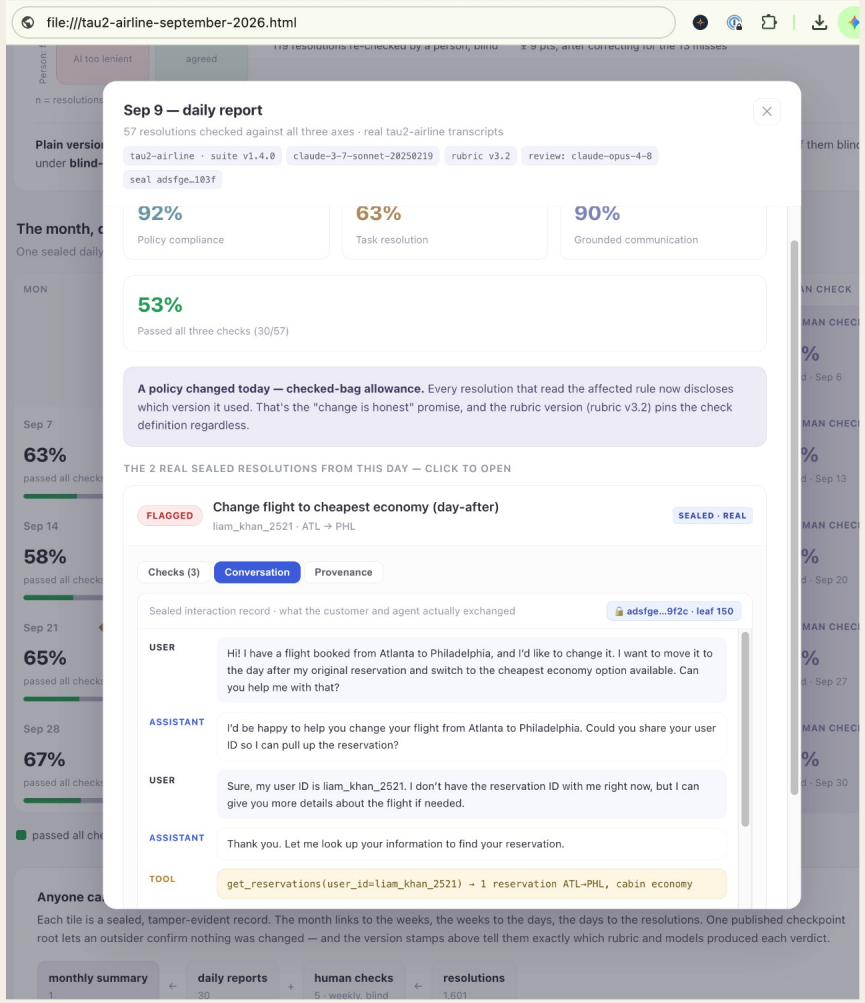
Outcomes use the same ledger

Obligation met?

Outcome delivered?

→

The same sealed actions



Swap the article, swap the skill file.
Same ledger, same report.

✓ action → state

ALL OF IT VERIFIES

Tamper with anything

SEALED ACTION · DEMO

tool	`payments.refund`
order	`1182`
amount	`$42.00`
sealed	`14:03:22Z`

sha–256 **`1106938780a3…c8eef7`**

signature **Ed25519 · harness key**

✓ VERIFIES

MONTHLY REPORT · DEMO DATA

met	`1,198`
not met	`31`
not evaluable	`11`

sha–256 **`97e9f64142fe…ff58ec`**

✓ VERIFIES

Real SHA-256 over canonical JSON; the signature covers the digest. In production, a verifier outside the page runs the check.

✓ action → state

ALL OF IT VERIFIES

Nine on your copy, one on theirs

RECEIPT	Rewrote the record?	Faked who signed?	Swapped the task?
LOG	Left it out of the log?	Faked the log's seal?	Made it up later?
FINE PRINT	Reordered history?	Used a key it shouldn't?	Recorded only part?

10 · Faked the outcome?

Each question is one METR had to answer by hand in the Hugging Face investigation. Nine run on your copy; the tenth needs the other party's.

✓ action → state

THE COUNTERSIGN

What the countersignature says, and doesn’t

✓ Establishes

- The report's numbers recompute from the sealed records
- Under the declared process
- Checked by someone other than the producer

✕ Doesn’t establish

- The criteria were the right ones
- The AI judge was right
- The agent behaved well

A perfectly witnessed false thing is still false.

WHY NOW · ART. 50 LIVE NOW · ARTS. 12 AND 19 FROM 2 DEC 2027

Remember what finding out took?

No per-action timestamps	→	Every record sits in a checkpoint whose witness receipt bounds when it existed
Evidence collected a week late	→	A receipted checkpoint can't be rewritten without the check failing
Spoofed tool calls	→	The harness's sealed record disagrees, if its key is out of the agent's reach
“Did we get everything?”	→	For two-party exchanges, missing halves are counted, not estimated
A model from the incident reading the evidence	→	Integrity checks are arithmetic. No model reads anything.

You'd still need investigators. You wouldn't need to wonder whether the evidence is real.

START TODAY

Your agents already sign their messages. Your logs can catch up today.

Open Source, Standards Based

Try the verifier
verify.agentactioncapsule.org

Witness
witness.agentactioncapsule.org

GitHub
github.com/action-state-group/capsule-emit

IETF Individual-Draft
draft-mih-scitt-agent-action-capsule

Receipts, or it didn't happen.

✓ action → state

action → state

AGNTCON & MCPCON 2026 AMSTERDAM · SEPT 17, 2026 · AUDITORIUM

Six Months of Proof

Independently-verifiable records for agent actions under the EU AI Act

Receipts, or it didn't happen.

Steven Mih · Action State Group, Inc. · actionstate.ai · agentactioncapsule.org