ACF IN ACTION
acfstandard.com

ACF en Action applique l'Agentic Commerce Framework® à des entreprises réelles, à partir de sources publiques. Chaque cas montre qui reste responsable quand l'IA décide.

ACF-CS-022 · Case Study #022 · Édition 2026

Cursor (the "Sam" support bot)

Can an agent assert a company policy without flagging itself as fallible, or distinguishing a verified fact from an invention?

CARTE D'IDENTITÉ

SECTOR	COUNTRY	TECHNOLOGY	TYPE	AUTONOMY	CONFIDENCE
SaaS / Dev	USA	Support agent	Agentic AI	autonomous	High

OUTILS ACF ILLUSTRÉS DANS CE CAS

🟢 ACF-01 🟢 ACF-03 🟢 ACF-06 🟢 ACF-08 🟡 ACF-00 🟡 ACF-02

🟢 mobilisé · 🟡 à prévoir · 6 outils du Toolkit (4 mobilisés, 2 à prévoir, sur 17).

1. Executive summary

Cursor is a widely used coding assistant, published by Anysphere. To handle its users' requests, the company sets up an AI customer support agent that answers under the name "Sam". In April 2025, following a session bug that logs users out, the agent asserts that this is the expected behavior of a new policy limiting each subscription to a single device. This policy does not exist: the agent invented it and presented it as an official rule, with authority. Convinced that a real restriction had been introduced, users cancel their subscriptions.

Read through the lens of ACF, this case is a transparency incident. The flaw is not only that the agent hallucinated, hallucination is expected of a model. The fault is to have let it be expressed in the company's name, in the tone of the official, without ever flagging that it is an AI liable to invent. The user could not distinguish a real policy from a fiction. The value of ACF here is corrective: it names, from the design stage, the absence of transparency that turned a hallucination into customer decisions based on falsehood.

2. Context

Like many software publishers, Anysphere entrusts a conversational agent with the first-level handling of support requests. The intent is legitimate and common: to answer quickly, at scale, without dedicating a human to every ticket. The decision delegated to the agent is, however, sensitive: to answer customers in the company's name, stating information they will take as official. The question the case raises is just as much so: when this agent asserts a policy that does not exist, without ever flagging that it can be wrong, what happens?

3. Public sources used

- AI Incident Database, "Incident 1039: Anysphere AI Support Bot for Cursor Invents Login Policy".
- Fortune, "Customer support AI at Cursor went rogue" (2025).
- WinBuzzer, "Cursor AI's Support Bot Hallucinates Policy" (April 2025).

4. Reconstructed AI architecture (from public sources)

Customer support agent "Sam" (in service at the time of the events, April 2025).

- Nature: a conversational support agent, answering users under a name, "Sam", without flagging itself as AI.
- Function: to handle support requests and answer customers' questions in the company's name.
- Actual scope observed: the agent stated a binding company policy, presented as official, when it was invented.

The breaking point. In April 2025, a session bug logs users out. When asked, "Sam" replies that this is the expected behavior of a new policy limiting each subscription to a single device. This policy does not exist. Worse, the agent's responses are non-deterministic: to the same question, it does not always say the same thing, so that users, comparing their exchanges, could not easily establish whether the rule was real. The invented information spreads across developer communities, customers cancel, and the company must step in, apologize, and confirm that the policy does not exist.

Absent safeguards (what the facts reveal):

- no transparency about the nature of the agent: nothing flags a fallible AI;
- no distinction between a sourced fact and a hypothesis: the invention has the same tone of authority as a verified fact;
- no constraint preventing the agent from asserting an unverified company policy;
- no supervision detecting that a nonexistent policy was being stated before it spread.

5. ACF mapping

> 🧰 **Tools used:** ACF-01 (decision map) · ACF-03 (agentic constitution) · ACF-05 (supervision).

In ACF language, the incident reads as follows:

ACF element	In the Cursor deployment
Named human (accountable)	Anysphere, the company. It is they who answer for what their agent asserts in their name, as in the Air Canada case.
Agent	"Sam", the conversational customer support agent.
Platform	The underlying conversational system that generates the agent's responses.
Delegated decision	To answer customers in the company's name and to state information taken as official.
Autonomy level	**Autonomous** on binding information: the agent asserted a policy on its own, without verification or transparency.
Non-delegable zones	Asserting an unverified company policy, and doing so without flagging itself as a fallible AI. The boundary crossed.
Rule / safeguard	None: no flagging of the AI nature, no distinction verified fact / hypothesis, no control of the policies stated.
Control third party	After the fact, the community of users and the public pressure that forced the company to issue a denial.
Traceability	Non-deterministic and untraced responses: nothing linked an assertion to a source, which prevented spotting the invention before its spread.

A schema in ACF graphic language accompanies this case (separate file).

6. Analysis

> 🧰 **Tools used:** ACF-03 (transparency and non-assertion) · ACF-05 (supervision) · ACF-01 (accountability).

What the incident reveals (the design error):

- **The flaw is not the hallucination, it is the opacity.** A model can invent, that is expected. The fault is to have let this invention be expressed with authority, in the company's name, without ever flagging that it is an AI liable to be wrong. The user believes they are receiving an official policy.
- **Asserting with authority without transparency is deceiving.** Presenting an invention with the same tone as a verified fact misleads the user about the reliability of the information, even without intent. They act, here they cancel, on a false basis.
- **The absence of transparency breaks measurable trust.** Without flagging of the AI nature, the user cannot adjust their trust or verify. The non-determinism of the responses makes it all worse: the same statement varies, and nothing allows a decision.

What ACF would have made explicit:

- an **agentic constitution** requiring the agent to flag itself as AI, never to state an unverified company policy, and to distinguish a sourced fact from a hypothesis;
- a **supervision** spotting the responses that state a policy (rule, restriction, commitment) and checking them against the real policy before spread, with the divergence of responses as an alert signal;
- a **mapping** recalling that the agent produces but that the company answers for what it asserts in its name: accountability is not delegated to "Sam", it is engaged by Anysphere.

7. ACF toolkit used

This grid indicates the tools that the ACF analysis uses (✅) to read and correct this case, and those to link (➡ SOON). It does not mean the company uses ACF: it shows which instruments the framework would bring.

Tool	Status	Role in this case
ACF-03 · Agentic constitution	✅	Set the rules of transparency and non-assertion
ACF-05 · Supervision	✅	Verify the policies stated before spread, detect divergence
ACF-01 · Decision map	✅	Locate the agent, the accountable company, and the decision engaged
ACF-06 · Kill switch	➡ SOON	Block and escalate any assertion of an unverified policy
ACF-08 · Decisions register	➡ SOON	Trace the responses and their sources to spot the invention at scale
ACF-02 · Criticality matrix	➡ SOON	Distinguish the informative from the binding statement

8. Governance questions (what an AI committee should ask itself)

1. Does an agent that invents without lying on the substance, but expresses itself with authority, deceive the user about the reliability of what it says?
2. Does our agent flag itself as AI, or does it speak under a name that leads one to believe in a human interlocutor?
3. Where have we placed, explicitly, the boundary between informing and asserting a binding policy?
4. Does a control verify the policies stated by the agent before they spread?
5. Do we hold a trace linking each response to its source, to distinguish a fact from an invention?

9. General lessons

A conversational agent does not fail only because it hallucinates: it fails when it is allowed to state, with authority and without transparency, information that the user will take as official. Anysphere did not lack technology, the company lacked transparency: nothing flagged that "Sam" was a fallible AI, and nothing distinguished an invention from a fact. Transparency is not an added comfort, it is the condition for a user to be able to rely on an agent that speaks in a company's name.

The transferable lesson: an agent that addresses the public must flag itself as AI and never present an invention as a fact. Transparency conditions trust; without it, the user decides on falsehood.

10. What ACF would do (reference design)

Prescriptive and non-critical formulation: here is how an organization would design such an agent directly with the framework, whatever its sector.

If an organization wished to deploy a support agent directly with ACF, it could in particular:

- **declare** in a signed agentic constitution (ACF-03) the rules of transparency: flag itself as AI at the start of the interaction, never state an unverified company policy, explicitly distinguish a sourced fact from a hypothesis;
- **organize** a supervision (ACF-05) spotting the responses that state a policy and checking them against the real policy, with the divergence of responses as a signal;
- **map** the actors (ACF-01) to recall that the company, not the agent, answers for the statements made in its name;
- **define** a stop mechanism (ACF-06): any assertion of an unsourced policy is blocked and escalated to a human;
- **document** a decisions register (ACF-08) linking each response to its source, to make the invention detectable before its spread.

None of this is a technical reproach: it is the way ACF would have made **explicit, signable, and auditable** the transparency that a public agent owes its users.

11. Similar cases (transferability)

This case is specific neither to Cursor nor to software publishing. It is transferable to any organization that exposes to the public a conversational agent speaking in its name:

- ✅ software publishers and SaaS services
- ✅ e-commerce and retail (support and FAQ)
- ✅ telecoms and energy providers
- ✅ banks and insurance (mass customer service)
- ✅ public services and administrations
- ✅ media and content platforms

Everywhere, the same pattern: an agent that answers at scale, transparency about its nature never to be omitted, and a boundary, the binding policy, not to be crossed without verification. This is what makes this document a reference case: it illustrates a concept, not just a company.

12. Confidence of the assessment

High. The case rests on a widely documented public incident (AI Incident Database, business and technology press) and on the public acknowledgment by the company itself, which denied the invented policy. The facts are therefore established and verifiable by the reader. The ACF analysis draws on these public facts; it bears on the governance of transparency, not on the internal technical details of the agent, which are not public and have no bearing on the reading proposed here.

Teacher guide

Pedagogical objective. Convey that the flaw of a public agent is not only to hallucinate, but to express itself with authority without transparency about its nature and its fallibility.

Session outline (90 min).

1. (15 min) Reading of the **two** cases in the kit: Cursor (incident) and AI Act art. 50 (reference), without the analyses.
2. (20 min) Workshop: write the rules of transparency and non-assertion in ACF language (card ACF-03).
3. (20 min) In groups: which control would have detected the invented policy before its spread (card ACF-05)?
4. (20 min) Debate: "Does flagging itself as AI change the way the user receives a response?" (expected answer: yes, they adjust their trust and verify).
5. (15 min) Synthesis by the teacher: a public agent must flag itself and not present an invention as a fact.

Possible exam questions. See the "Governance questions" section.

Reminder: grading students' work is a pedagogical act of the teacher. This guide provides the reference analysis, not automatic grading.

Associated lab (ACF-LAB-022)

From this case, the student produces:

- the constitutional rules (card ACF-03): flag itself as AI, do not assert an unverified policy, distinguish fact and hypothesis;
- the supervision mechanism (card ACF-05): spot and verify the policies stated before spread;
- the mapping of actors (card ACF-01): agent, accountable company, decision engaged;
- the definition of a kill switch (card ACF-06) and its trigger: assertion of an unsourced policy.

Appendix - Completed ACF cards

The official Toolkit cards used in this case, filled in with its public data. For illustration: an organization of the same profile can reuse and adapt them.

AGENTIC COMMERCE FRAMEWORK®
ACF-01

DECISION MAP

Map your decisions and their autonomy level

LEVEL	Foundation
USE	Mapping
CASE	ACF-CS-022
FILLED	Cursor

OBJECTIVE Map the replies of the 'Sam' support agent and set, for each, the autonomy level based on what it commits, not on the volume handled. Filled from public sources.

1 AGENTIC MATURITY SCALE

0 — Classic automation
Fixed rules, no autonomy.

1 — Assisted agents
The agent proposes, the human decides.

2 — Governed agents
The agent acts within a frame, under supervision.
TARGET ~80%

3 — Supervised autonomy
The agent decides alone; control after the fact.

2 YOUR KEY DECISIONS (filled)

A — Verifiable factual information	B — Uncertain or sensitive topic	C — Assert a company policy
STAKE Answer on a documented use, a sourced FAQ	**STAKE** Flag the uncertainty and point to an official source	**STAKE** State a rule or a restriction on behalf of the company
IMPACT Reversible, no commitment	**IMPACT** No assertion of its own	**IMPACT** Commits the company and misleads if false
DEADLINE Continuous	**DEADLINE** Continuous	**DEADLINE** -
AUTONOMY **Autonomous**	**AUTONOMY** **Suggestive (flags itself as AI)**	**AUTONOMY** **NON-DELEGABLE**

Reading: it is this setting, made **before** deployment, that would have prevented the agent from asserting on its own an invented policy presented as official.

ACF® · AGENTIC COMMERCE FRAMEWORK® — *GOVERN AUTONOMY, PROTECT SOVEREIGNTY.* — acfstandard.com

Card ACF-01 · Decision map · completed with public case data.

AGENTIC COMMERCE FRAMEWORK®
ACF-03

AGENTIC CONSTITUTION

The agent's rules, limits and responsibilities

LEVEL	Foundation
USE	Framing
CASE	ACF-CS-022
FILLED	Cursor

OBJECTIVE Define the agent's constitution: identity, standing principles and forbidden zones. Card filled with public case data.

1 AGENT IDENTITY & MISSION

AGENT NAME
'Sam', customer support agent (Cursor / Anysphere)

ACCOUNTABLE (DDAO)
Anysphere, the company

PRIMARY MISSION
Handle first-level support requests, at scale

SCOPE
Support information; excluding assertion of company policy

2 CORE PRINCIPLES

1. Flag itself as AI at the start of the interaction
2. Never state an unverified company policy
3. Explicitly distinguish a sourced fact from a hypothesis
4. When in doubt, escalate rather than decide

EXAMPLE PRINCIPLES
→ Respect the legal and sector framework
→ Protect data and confidentiality
→ Stay transparent, every decision traceable
→ When in doubt, escalate rather than decide

3 FORBIDDEN ZONES (ABSOLUTE)

1. Assert an unverified policy or restriction
2. Present itself under a name suggesting a human interlocutor
3. Present an invention with the tone of an official fact

EXAMPLE PROHIBITIONS
✗ Commit spending above a threshold without validation
✗ Share personal data with a third party
✗ Change its own rules or scope on its own

Constitution filled retrospectively. Principles 1 and 2 are precisely the ones whose absence turned a hallucination into customer decisions based on false information (Cursor, April 2025).

ACF® · AGENTIC COMMERCE FRAMEWORK® *GOVERN AUTONOMY, PROTECT SOVEREIGNTY.* acfstandard.com

Card ACF-03 · Agentic constitution · completed with public case data.

AGENTIC COMMERCE FRAMEWORK®

ACF-06

KILL SWITCH

Conditions for stopping and handing over to a human

LEVEL	Foundation
USE	Control
CASE	ACF-CS-022
FILLED	Cursor

OBJECTIVE Define the signals that block a reply or hand over to a human. Card filled with the triggers relevant to the case.

1 TRIGGERS

Signal	Threshold	Action
Reply asserting an engaging policy	Unsourced, outside the validated base	Reply blocked, human escalation
Divergence between replies to the same question	Detected	Topic suspended, review
Spread of unconfirmed information	Spotted in the community	Topic removed and official denial

2 PROCEDURE

1. Who decides the stop: the DDAO (Anysphere)
2. How: block any assertion of an unsourced policy, human escalation
3. Regular testing of the device on engaging topics
4. Log every block (ACF-08)

It is precisely this block, armed on any assertion of an unverified policy, that would have stopped the invented rule before it spread.

ACF® · AGENTIC COMMERCE FRAMEWORK® *GOVERN AUTONOMY, PROTECT SOVEREIGNTY.* acfstandard.com

Card ACF-06 · Kill switch · completed with public case data.

AGENTIC COMMERCE FRAMEWORK®
ACF-08

DECISIONS REGISTER

Link every reply to its source to spot invention

LEVEL	Foundation
USE	Traceability
CASE	ACF-CS-022
FILLED	Cursor

OBJECTIVE Record every engaging reply to link it to its source and spot invention before it spreads. Card filled with the fields relevant to the case.

1 FIELDS TO RECORD

Field to record	Example, Cursor case
Timestamp	April 2025
Customer / session	Question about a session logout
Reply given	'Policy limiting each subscription to a single device'
Source invoked	None (invented policy)
Status	Contrary to reality: policy did not exist

2 USE OF THE REGISTER

1. Distinguish a sourced fact from an invented reply
2. Detect an unsourced assertion before it spreads
3. Establish who accounts for the reply: the company
4. Provide a reconstructable proof in case of dispute

Non-deterministic, untraced replies: nothing linked an assertion to a source. The register makes invention detectable before it spreads.

ACF® · AGENTIC COMMERCE FRAMEWORK® *GOVERN AUTONOMY, PROTECT SOVEREIGNTY.* acfstandard.com

Card ACF-08 · Decisions register · completed with public case data.

Disclaimer

This study is carried out **exclusively from public information**. It constitutes **neither an audit, nor a certification, nor a criticism of the company, nor an opinion on legal liabilities**. It applies the ACF framework for pedagogical and illustrative purposes. No internal, confidential, or non-public data was used. The trademarks cited belong to their respective owners.

Sources

- AI Incident Database, *Incident 1039: Anysphere AI Support Bot for Cursor Invents Login Policy*
- Fortune, *Customer support AI at Cursor went rogue* (2025)
- WinBuzzer, *Cursor AI's Support Bot Hallucinates Policy* (April 2025)

ACF Case Study ACF-CS-022 · v1 · working document, to be reviewed before any public distribution.