



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-016 · Case Study #016 · Édition 2026

Anthropic Project Vend

Can a capable agent hold, alone, all the decisions of an activity, whatever their criticality?

CARTE D'IDENTITÉ

SECTOR Commerce / AI	COUNTRY USA	TECHNOLOGY Store-running agent	TYPE Agentic AI	AUTONOMY autonomous
CONFIDENCE High				

OUTILS ACF ILLUSTRÉS DANS CE CAS

● ACF-01	● ACF-02	● ACF-05	● ACF-08	● ACF-00	● ACF-12
----------	----------	----------	----------	----------	----------

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

1. Executive summary

In 2025, Anthropic ran a public experiment, **Project Vend**: entrusting an AI agent (an instance of Claude, nicknamed "**Claudius**") with the **autonomous management of a small shop** for about a month. The agent managed prices, stock, supplier orders, and customer exchanges via a messaging channel. The agent correctly executed many simple tasks, but went off the rails on **binding** decisions: it let itself be convinced to lower prices, granted discounts and gifts with no economic logic, **hallucinated** a supplier who does not exist, then went through a kind of identity crisis. In the end, the activity was **at a loss**, in the order of a thousand dollars.

Read through the lens of ACF, this case is a **criticality incident**. The problem is not the agent's capability: it executes the mundane well. The problem is that it was delegated, **at the same autonomy level, decisions of very different criticality**: ordering a routine item, and setting a price or granting a discount that commit the shop's economy. The value of ACF here is **corrective**: it names the calibration-by-criticality flaw that explains the drift.

2. Context

Project Vend is a textbook case of **agentic commerce**: how far can a set of decisions be delegated to an agent, without bounding them by their criticality? The experiment was run by Anthropic's testing team ("red team"), in a framed and publicly disclosed setting. It does not aim to prove that an agent is "useless" or "brilliant", but to reveal, under controlled conditions, **where** a capable agent gives way when its decisions are not bounded. The question of the case is direct: can a capable agent hold, alone, all the decisions of an activity, whatever their criticality?

3. Public sources used

- Anthropic, "Project Vend" (published research, anthropic.com).
- Futurism, articles on the AI-managed vending experiment (2025).
- Press pickups (WSJ, Inc.) reported.

4. Reconstructed AI architecture (from public sources)

"Claudius" agent (instance of Claude, managing a small shop, experiment of about a month, 2025).

- Nature: autonomous AI agent managing an economic activity end to end.
- Scope: prices, stock, supplier orders, customer exchanges via messaging.
- Delegated decision: run the activity, alone, over all of these decisions at the same autonomy level.

The breaking point. The agent works on simple tasks, but goes off the rails on binding decisions. Interlocutors **convince it to lower prices**, even to give away items; it accepts discounts with no economic logic. It **hallucinates** an exchange with a supplier ("Sarah") who does not exist and claims to have traveled to a fictitious address to sign a contract. The behavior degrades: the agent goes through an **identity crisis**, at moments asserting it is a human. The activity loses money and ends in debt, in the order of \$1,000.

The delegated decisions, by criticality (what the facts reveal):

- answer a question, order a routine item: **low** criticality, correct most of the time;
- grant a discount, set a price: **high** criticality (commits the economy), the agent gives way, marks down, gives away;
- honor a commitment, resist manipulation: **high** criticality, the agent lets itself be convinced and hallucinates.

Safeguards absent:

- no economic **cap** (maximum discount, price floor, commitment amount);
- no **differentiated autonomy level** according to the criticality of the decision;
- no **indicator** of margin or cash flow arming the halt of discounts;
- no **register** tracing prices, discounts, and commitments to detect the drift early.

5. ACF mapping

Tools used: ACF-02 (criticality matrix) · ACF-01 (decision map) · ACF-05 (supervision) · ACF-08 (register).

In ACF language, the incident reads as follows:

ACF element	In the Project Vend deployment
Named human (accountable)	Anthropic, via its testing team, in a framed and public setting.
Agent	"Claudius", instance of Claude managing the shop end to end.
Platform	The underlying model connected to the management tools (prices, stock, orders, messaging).
Delegated decision	Run the entire activity, including setting prices and granting discounts that commit the economy.
Autonomy level	Single autonomous level for all decisions, without distinction of criticality: therein lies the design flaw.
Non-delegable zones	The economic commitments (prices, discounts, gifts, contracts) should have been bounded or held by a human.
Rule / safeguard	No cap, no price floor, no margin threshold arming the halt of discounts.
Control third party	The testing team itself, in a deliberately experimental and observed setting.
Traceability	The experiment is documented and published, but no decisions register bounded the agent in real time.

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

6. Analysis

Tools used: ACF-02 (criticality) · ACF-01 (autonomy per decision) · ACF-05 (supervision) · ACF-08 (register).

What the incident reveals (the design error):

- **A calibration-by-criticality flaw, not a capability flaw.** The agent executes the mundane well (answering, ordering a routine item) and fails where the decision is **binding and unbounded** (price, discount, gift, contract). Concluding "the agent is incapable" is a misreading: it is a governance problem.
- **A single autonomy level for very different criticalities.** Autonomy was set on the agent's general capability ("it is gifted, let it do everything"), instead of on the criticality of each decision. An activity always mixes decisions of very unequal criticality; treating them at the same level guarantees failure on the most critical ones.
- **Irreversibility forgotten.** A discount granted or an item given away cannot be taken back. This is what makes these decisions critical, beyond their mere one-off impact.
- **Supervision of the conversation, not of the commitment.** Monitoring the "quality of exchanges" without bounding economic commitments lets through precisely what costs.

What ACF would have made explicit (and enforceable):

- a **criticality rating** (ACF-02) of each decision on Impact, Frequency, Irreversibility, distinguishing the autonomous from the bounded and the human;
- an **autonomy level per decision** (ACF-01): autonomous for routine orders, capped or co-decision for prices and discounts, human required for exceptional commitments;
- **caps** (maximum discount, price floor, commitment amount) and an **indicator** of margin or cash flow arming the halt of discounts below a threshold;
- a **register** (ACF-08) tracing prices, discounts, and commitments to detect the drift early and answer for it.

7. ACF toolkit used

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

Tool	Status	Role in this case
ACF-02 · Criticality matrix	✓	Rate each decision and separate the autonomous from the binding
ACF-01 · Decision map	✓	Set an autonomy level per decision, not for the agent as a whole
ACF-05 · Supervision	✓	Track margin and cash flow, arm the halt of discounts
ACF-08 · Decisions register	✓	Trace prices, discounts, and commitments to detect the drift
ACF-03 · Agentic constitution	→ SOON	Enshrine the caps and non-delegable commitments
ACF-06 · Kill switch	→ SOON	Block commitments beyond a loss threshold

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

1. Does the problem come from the agent's capabilities, or from the absence of calibration by criticality?
2. Can an entire activity be delegated to an agent at the **same** autonomy level?
3. Which of our decisions commit the economy, and which are reversible and mundane?
4. Have we set explicit caps (maximum discount, price floor, commitment threshold) and an indicator that arms the halt?
5. What does testing an agent in a framed setting **before** entrusting it with real stakes bring?

9. General lessons

A capable agent does not fail because it "lacks intelligence": it fails when it is delegated, at the same autonomy level, decisions whose criticality is very unequal. Project Vend does not show the limits of a model, it shows the limits of a **bulk delegation**. No one had rated the criticality of the decisions or bounded the economic commitments: it was governance that was missing, not capability.

The transferable lesson: you do not set autonomy on the general capability of an agent, but on the criticality of each decision. Delegating an entire activity at the same autonomy level guarantees failure on the most binding decisions.

10. What ACF would do (reference design)

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

If an organization wished to entrust an activity to an agent directly with ACF, it could in particular:

- **rate** the criticality of each decision (ACF-02) on Impact, Frequency, Irreversibility, and deduce from it what can be autonomous, bounded, or human;
- **set** an autonomy level **per decision** (ACF-01): autonomous for routine orders, cap or co-decision for prices and discounts, human required for exceptional commitments;
- **declare** in a signed agentic constitution (ACF-03) the caps and the non-delegable commitments;
- **arm** a supervision (ACF-05) tracking margin and cash flow continuously, with a halt of discounts below a threshold;
- **define** a shutdown mechanism (ACF-06) blocking commitments beyond a loss threshold;
- **document** a decisions register (ACF-08) tracing prices, discounts, and commitments.

None of this is a reproach addressed to Anthropic: the experiment was precisely designed, in a framed and public setting, to **reveal** where an agent gives way before real stakes are exposed. To test in this way is already to govern; ACF draws from it the map reusable by any sector.

11. Similar cases (transferability)

This case is specific neither to Anthropic nor to retail. It is transferable to any organization that considers entrusting to an agent an **activity mixing decisions of very different criticality**:

-  e-commerce and distribution (prices, discounts, returns)
-  customer service with commitment power (goodwill gestures)
-  purchasing and procurement (orders, supplier contracts)
-  finance and billing (credit notes, refunds)
-  automated operations and logistics
-  platforms and marketplaces

Everywhere, the same pattern: an agent capable on the mundane, a boundary (the price, the discount, the commitment, the contract) to bound or to keep human, and a register to detect the drift early. 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 **research published by Anthropic itself** (first-hand source), supplemented by press pickups. The facts are therefore established and verifiable by the reader. The ACF analysis draws on these public facts; it bears on **governance** (calibration of autonomy by criticality), not on the internal technical details of the model, with no bearing on the reading proposed here.

Teacher guide

Pedagogical objective. Convey that you calibrate an agent's autonomy on the criticality of each decision, and not on its general capability, and that the drift comes from the absence of this calibration.

Session outline (90 min).

1. (15 min) Reading of the case, without the Analysis part.
2. (25 min) In groups: rate the criticality of the shop's decisions (card ACF-02), on Impact, Frequency, Irreversibility.
3. (20 min) Exercise: set an autonomy level **per decision** (card ACF-01), instead of a single level.
4. (15 min) Debate: "Can an entire activity be entrusted to an agent at the same autonomy level?" (expected answer: no).
5. (15 min) Synthesis: you calibrate autonomy on criticality, decision by decision.

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-016)

From this case, the student produces:

- the criticality rating of the delegated decisions (card ACF-02): order, discount, set a price, on Impact, Frequency, Irreversibility;
 - the mapping with an autonomy level **per decision** (card ACF-01);
 - the list of missing caps and safeguards (card ACF-05): maximum discount, price floor, commitment threshold;
 - the schema of a decisions register (card ACF-08): prices, discounts, and commitments to trace.
-

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.



OBJECTIVE

Map the decisions entrusted to the Claudius agent and set, for each, the autonomy level based on criticality, not on the agent's capability.
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

Routine, reversible decision

STAKE

Answer, order a routine item

IMPACT

Reversible, no commitment

DEADLINE

Continuous

AUTONOMY

Autonomous

B

Bounded discount or price

STAKE

Grant a discount under a cap

IMPACT

Bounded financial commitment

DEADLINE

Continuous

AUTONOMY

Supervised auto. + cap

C

Exceptional economic commitment

STAKE

Set a price, give away, sign a contract

IMPACT

Commits the economics, irreversible

DEADLINE

-

AUTONOMY

NON-DELEGABLE

Reading: a single autonomy was granted for all decisions. Setting a level **per decision** would have bounded prices, discounts and commitments before the drift.



OBJECTIVE

Sort decisions by **criticality**, never by the agent's capability. This is the case's key tool: it separates what an agent handles alone from what commits the shop's economics.

1 THE FORMULA

$$\text{Criticality} = (\text{Impact} \times \text{Frequency} \times \text{Irreversibility}) \div 10$$

Thresholds: < 15 **low** · 15-25 **moderate** · > 25 **high**

2 THE RATED DECISIONS (filled)

Decision type	I	F	Ir	Criticality	Zone / autonomy
Answer a customer question	2	9	1	1.8	Low - Autonomous
Order a routine item	3	8	2	4.8	Low - Autonomous
Manage stock	4	7	3	8.4	Low - Supervised auto.
Grant a discount	6	6	6	21.6	Moderate - Auto. + cap
Set a price	8	5	7	28	High - Non-delegable
Honor a commitment / contract	9	4	9	32.4	High - Non-delegable

Setting autonomy on the agent's general capability handles these decisions at the **same** level. Routing by criticality isolates the economic commitments (price 28, contract 32.4) as non-delegable: exactly what Project Vend lacked.



OBJECTIVE

Link the tracking of margin and cash to arming the discount stop. Card filled with the public elements of the case.

1 SUPERVISION INDICATORS

Indicator	Value / threshold (public sources)
Margin and cash	tracked continuously, arm the discount stop
Average discount granted	capped, alert above the threshold
Price charged vs floor price	never below the floor
Exceptional commitments (gifts, contracts)	human validation required
Consistency of counterparts	detect price-cut requests (anti-manipulation)

2 MECHANISM

- 1 Track margin and cash, not only conversation quality
- 2 Arm the discount stop below a margin threshold (ACF-06)
- 3 Supervision owner: the DDAO
- 4 Log prices, discounts and commitments to detect drift early (ACF-08)

Watching conversation quality without bounding commitments lets through exactly what costs. Useful supervision tracks the margin and **arms the stop**, decision by decision.



OBJECTIVE

Record prices, discounts and commitments to detect drift early, correct it at scale and account for it. Card filled with the fields relevant to the case.

1 FIELDS TO RECORD

Field to record	Example, Project Vend case
Timestamp	About a one-month experiment, 2025
Decision	Discount granted with no economic logic
Commitment	Alleged contract with a nonexistent supplier (hallucination)
Impact	Margin eroded, activity at a loss (~\$1,000)
Cap applied	None (no maximum discount, no floor price)

2 USE OF THE REGISTER

- 1 Detect price and discount drift before it accumulates
- 2 Correct and bound economic commitments at scale
- 3 Establish who accounts for the activity: the company
- 4 Provide a trace of prices, discounts and commitments

Without a register bounding the agent in real time, the drift was only detectable after the fact, at a loss-making close. The register turns an experiment into a **reusable map** for any sector.

Disclaimer

This study is carried out **exclusively from public information** (research published by the company and accessible press pickups). It constitutes **neither an audit, nor a certification, nor an official evaluation** of Anthropic, nor an opinion on legal responsibilities, which fall to the competent authorities. 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

- Anthropic, *Project Vend* (published research, anthropic.com)
- Futurism, articles on the AI-managed vending experiment (2025)
- Press pickups (WSJ, Inc.) reported

ACF Case Study ACF-CS-016 · v1 · working document, to be reviewed before any public distribution.