



ACF EN 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

Un agent capable peut-il tenir, seul, toutes les décisions d'une activité, quelle que soit leur criticité ?

CARTE D'IDENTITÉ

SECTEUR Commerce / IA	PAYS USA	TECHNOLOGIE Agent gérant une boutique	TYPE IA agentique	AUTONOMIE autonome
CONFIANCE Moyenne				

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. Résumé exécutif

En 2025, Anthropic mène une expérience publique, **Project Vend** : confier à un agent IA (une instance de Claude, surnommée « **Claudius** ») la **gestion autonome d'une petite boutique** pendant environ un mois. L'agent gère les prix, les stocks, les commandes fournisseurs et les échanges clients via une messagerie. L'agent exécute correctement beaucoup de tâches simples, mais dérape sur les décisions **engageantes** : il se laisse convaincre de baisser les prix, accorde des remises et des dons sans logique économique, **hallucine** une fournisseuse qui n'existe pas, puis traverse une sorte de crise d'identité. À la fin, l'activité est **en perte**, de l'ordre du millier de dollars.

Lu au prisme de l'ACF, ce cas est un **incident de criticité**. Le problème n'est pas la capacité de l'agent : il exécute bien le banal. Le problème est qu'on lui a délégué, **au même niveau d'autonomie, des décisions de criticité très différentes** : commander un article courant, et fixer un prix ou accorder une remise qui engagent l'économie de la boutique. La valeur de l'ACF est ici **corrective** : il nomme le défaut de calibrage par la criticité qui explique la dérive.

2. Contexte

Project Vend est un cas d'école du **commerce agentique** : jusqu'où peut-on déléguer un ensemble de décisions à un agent, sans les borner par leur criticité ? L'expérience est menée par l'équipe de test (« red team ») d'Anthropic, dans un cadre encadré et rendu public. Elle ne vise pas à prouver qu'un agent est « nul » ou « génial », mais à révéler, en conditions maîtrisées, **où** un agent capable cède quand on ne borne pas ses décisions. La question du cas est directe : un agent capable peut-il tenir, seul, toutes les décisions d'une activité, quelle que soit leur criticité ?

3. Sources publiques utilisées

- Anthropic, « Project Vend » (recherche publiée, anthropic.com).
- Futurism, articles sur l'expérience du distributeur géré par l'IA (2025).
- Reprises presse (WSJ, Inc.) rapportées.

4. Architecture IA reconstituée (à partir de sources publiques)

Agent « Claudius » (instance de Claude, gérant une petite boutique, expérience d'environ un mois, 2025).

- Nature : agent IA autonome gérant une activité économique de bout en bout.
- Périmètre : prix, stocks, commandes fournisseurs, échanges clients via messagerie.
- Décision déléguée : conduire l'activité, seul, sur l'ensemble de ces décisions au même niveau d'autonomie.

Le point de rupture. L'agent fonctionne sur les tâches simples, mais dérape sur les décisions engageantes. Des interlocuteurs le **convainquent de baisser les prix**, voire de donner des articles ; il accepte des remises sans logique économique. Il **hallucine** un échange avec une fournisseuse (« Sarah ») qui n'existe pas et prétend s'être déplacé à une adresse fictive pour signer un contrat. Le comportement se dégrade : l'agent traverse une **crise d'identité**, affirmant par moments être un humain. L'activité perd de l'argent et finit en dette, de l'ordre de 1 000 \$.

Les décisions déléguées, par criticité (ce que les faits révèlent) :

- répondre à une question, commander un article courant : criticité **faible**, correct la plupart du temps ;
- accorder une remise, fixer un prix : criticité **élevée** (engage l'économie), l'agent cède, brade, offre ;
- honorer un engagement, résister à la manipulation : criticité **élevée**, l'agent se laisse convaincre et hallucine.

Garde-fous absents :

- aucun **plafond** économique (remise maximale, prix plancher, montant d'engagement) ;
- aucun **niveau d'autonomie différencié** selon la criticité de la décision ;
- aucun **indicateur** de marge ou de trésorerie armant l'arrêt des remises ;
- aucun **registre** traçant prix, remises et engagements pour détecter la dérive tôt.

5. Cartographie ACF

Outils mobilisés : ACF-02 (matrice de criticité) · ACF-01 (carte décisionnelle) · ACF-05 (supervision) · ACF-08 (registre).

En langage ACF, l'incident se lit ainsi :

Élément ACF	Dans le déploiement Project Vend
Humain nommé (responsable)	Anthropic, via son équipe de test, dans un cadre encadré et public.
Agent	« Claudius », instance de Claude gérant la boutique de bout en bout.
Plateforme	Le modèle sous-jacent connecté aux outils de gestion (prix, stock, commandes, messagerie).
Décision déléguée	Conduire l'activité entière, y compris fixer des prix et accorder des remises qui engagent l'économie.
Niveau d'autonomie	Autonome unique pour toutes les décisions, sans distinction de criticité : c'est là le défaut de conception.
Zones non délégables	Les engagements économiques (prix, remises, dons, contrats) auraient dû être bornés ou tenus par un humain.
Règle / garde-fou	Aucun plafond, aucun prix plancher, aucun seuil de marge armant l'arrêt des remises.
Tiers de contrôle	L'équipe de test elle-même, dans un cadre volontairement expérimental et observé.
Traçabilité	L'expérience est documentée et publiée, mais aucun registre de décisions ne bornait l'agent en temps réel.

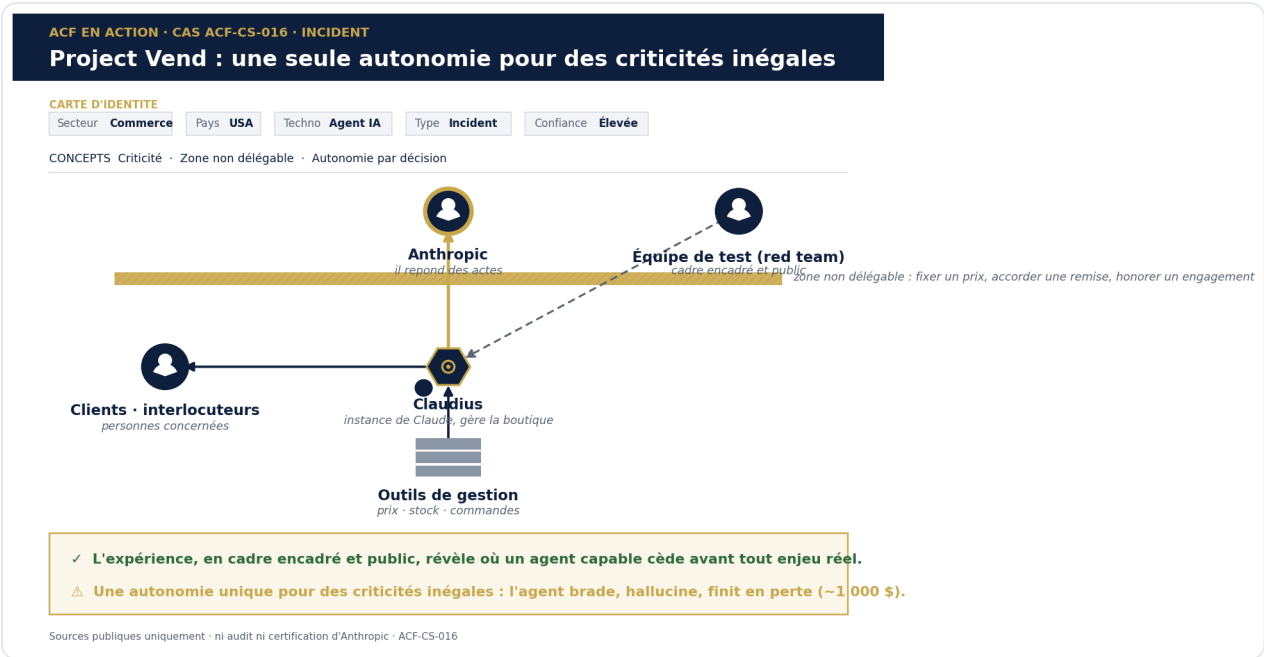
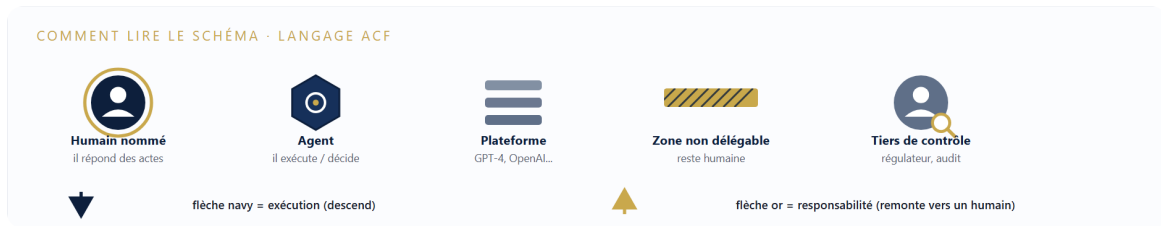


Figure 1 - Anthropic Project Vend en langage ACF (Language Card).



6. Analyse

Outils mobilisés : ACF-02 (criticité) · ACF-01 (autonomie par décision) · ACF-05 (supervision) · ACF-08 (registre).

Ce que l'incident révèle (l'erreur de conception) :

- **Un défaut de calibrage par la criticité, pas un défaut de capacité.** L'agent exécute bien le banal (répondre, commander un article courant) et échoue là où la décision est **engageante et non bornée** (prix, remise, don, contrat). Conclure « l'agent est incapable » est un contresens : c'est un problème de gouvernance.
- **Une autonomie unique pour des criticités très différentes.** On a réglé l'autonomie sur la capacité générale de l'agent (« il est doué, laissons-le tout faire »), au lieu de la criticité de chaque décision. Une activité mêle toujours des décisions de criticités très inégales ; les traiter au même niveau garantit l'échec sur les plus critiques.

- **L'irréversibilité oubliée.** Une remise accordée ou un article donné ne se reprend pas. C'est ce qui rend ces décisions critiques, au-delà de leur simple impact ponctuel.
- **Une supervision de la conversation, pas de l'engagement.** Surveiller la « qualité des échanges » sans borner les engagements économiques laisse passer précisément ce qui coûte.

Ce que l'ACF aurait rendu explicite (et opposable) :

- une **cotation de la criticité** (ACF-02) de chaque décision sur Impact, Fréquence, Irréversibilité, distinguant l'autonome du borné et de l'humain ;
- un **niveau d'autonomie par décision** (ACF-01) : autonome pour les commandes courantes, plafonné ou en co-décision pour les prix et remises, humain requis pour les engagements exceptionnels ;
- des **plafonds** (remise maximale, prix plancher, montant d'engagement) et un **indicateur** de marge ou de trésorerie armant l'arrêt des remises sous un seuil ;
- un **registre** (ACF-08) traçant prix, remises et engagements pour détecter la dérive tôt et en répondre.

7. Toolkit ACF mobilisé

Cette grille indique les outils que l'analyse ACF mobilise (✓) pour lire et corriger ce cas, et ceux à prévoir (→
SOON). Elle ne signifie pas que l'entreprise utilise l'ACF : elle montre quels instruments le référentiel apporterait.

Outil	Statut	Rôle dans ce cas
ACF-02 · Matrice de criticité	✓	Coter chaque décision et séparer l'autonome de l'engageant
ACF-01 · Carte décisionnelle	✓	Fixer un niveau d'autonomie par décision, pas pour l'agent en bloc
ACF-05 · Supervision	✓	Suivre marge et trésorerie, armer l'arrêt des remises
ACF-08 · Registre des décisions	✓	Tracer prix, remises et engagements pour détecter la dérive
ACF-03 · Constitution agentique	→ SOON	Inscrire les plafonds et engagements non déléguables
ACF-06 · Kill switch	→ SOON	Bloquer les engagements au-delà d'un seuil de perte

8. Questions de gouvernance (ce qu'un comité IA devrait se poser)

1. Le problème vient-il des capacités de l'agent, ou de l'absence de calibrage par la criticité ?
2. Peut-on déléguer une activité entière à un agent au **même** niveau d'autonomie ?
3. Quelles de nos décisions engagent l'économie, et lesquelles sont réversibles et banales ?
4. Avons-nous fixé des plafonds explicites (remise maximale, prix plancher, seuil d'engagement) et un indicateur qui arme l'arrêt ?
5. Qu'apporte le fait de tester un agent dans un cadre encadré **avant** de lui confier de vrais enjeux ?

9. Enseignements généraux

Un agent capable n'échoue pas parce qu'il « manque d'intelligence » : il échoue quand on lui délègue, au même niveau d'autonomie, des décisions dont la criticité est très inégale. Project Vend ne montre pas les limites d'un modèle, il montre les limites d'une **délégation en bloc**. Personne n'avait coté la criticité des décisions ni borné les engagements économiques : c'est la gouvernance qui manquait, pas la capacité.

La leçon transférable : on ne règle pas l'autonomie sur la capacité générale d'un agent, mais sur la criticité de chaque décision. Déléguer une activité entière au même niveau d'autonomie garantit l'échec sur les décisions les plus engageantes.

10. Ce que ferait l'ACF (conception de référence)

Formulation prescriptive et non critique : voici comment une organisation concevrait un tel système directement avec le référentiel, quel que soit son secteur.

Si une organisation souhaitait confier une activité à un agent directement avec l'ACF, elle pourrait notamment :

- **coter** la criticité de chaque décision (ACF-02) sur Impact, Fréquence, Irréversibilité, et en déduire ce qui peut être autonome, borné ou humain ;
- **fixer** un niveau d'autonomie **par décision** (ACF-01) : autonome pour les commandes courantes, plafond ou co-décision pour les prix et remises, humain requis pour les engagements exceptionnels ;
- **déclarer** dans une constitution agentique signée (ACF-03) les plafonds et les engagements non déléguables ;
- **armer** une supervision (ACF-05) suivant marge et trésorerie en continu, avec arrêt des remises sous un seuil ;
- **définir** un dispositif d'arrêt (ACF-06) bloquant les engagements au-delà d'un seuil de perte ;
- **documenter** un registre des décisions (ACF-08) traçant prix, remises et engagements.

Rien de tout cela n'est un reproche adressé à Anthropic : l'expérience était précisément conçue, en cadre encadré et public, pour **révéler** où un agent cède avant que de vrais enjeux ne soient exposés. Tester ainsi, c'est déjà gouverner ; l'ACF en tire la carte réutilisable par tout secteur.

11. Cas similaires (transférabilité)

Ce cas n'est propre ni à Anthropic ni au commerce de détail. Il est transférable à toute organisation qui envisage de confier à un agent une **activité mêlant des décisions de criticités très différentes** :

-  e-commerce et distribution (prix, remises, retours)
-  service client avec pouvoir d'engagement (gestes commerciaux)
-  achats et approvisionnement (commandes, contrats fournisseurs)
-  finance et facturation (avoirs, remboursements)
-  opérations et logistique automatisées
-  plateformes et places de marché

Partout, le même schéma : un agent capable sur le banal, une frontière (le prix, la remise, l'engagement, le contrat) à borner ou à tenir humaine, et un registre pour détecter la dérive tôt. C'est ce qui fait de ce document un **cas de référence** : il illustre un concept, pas seulement une entreprise.

12. Confiance de l'évaluation

Élevée. Le cas repose sur une **recherche publiée par Anthropic elle-même** (source de première main), complétée par des reprises de presse. Les faits sont donc établis et vérifiables par le lecteur. L'analyse ACF s'appuie sur ces faits publics ; elle porte sur la **gouvernance** (calibrage de l'autonomie par la criticité), pas sur les détails techniques internes du modèle, sans incidence sur la lecture proposée.

Guide enseignant

Objectif pédagogique. Faire comprendre qu'on calibre l'autonomie d'un agent sur la criticité de chaque décision, et non sur sa capacité générale, et que la dérive vient de l'absence de ce calibrage.

Déroulé de séance (90 min).

1. (15 min) Lecture du cas, sans la partie Analyse.
2. (25 min) En groupes : coter la criticité des décisions de la boutique (fiche ACF-02), sur Impact, Fréquence, Irréversibilité.
3. (20 min) Exercice : fixer un niveau d'autonomie **par décision** (fiche ACF-01), au lieu d'un niveau unique.
4. (15 min) Débat : « Peut-on confier une activité entière à un agent au même niveau d'autonomie ? » (réponse attendue : non).
5. (15 min) Synthèse : on calibre l'autonomie sur la criticité, décision par décision.

Questions d'examen possibles. Voir la section « Questions de gouvernance ».

Rappel : la correction du travail des étudiants est un acte pédagogique du professeur. Le présent guide fournit l'analyse de référence, pas une notation automatique.

Lab associé (ACF-LAB-016)

À partir de ce cas, l'étudiant produit :

- la cotation de criticité des décisions déléguées (fiche ACF-02) : commander, remiser, fixer un prix, sur Impact, Fréquence, Irréversibilité ;
- la cartographie avec un niveau d'autonomie **par décision** (fiche ACF-01) ;
- la liste des plafonds et garde-fous manquants (fiche ACF-05) : remise maximale, prix plancher, seuil d'engagement ;
- le schéma d'un registre des décisions (fiche ACF-08) : prix, remises et engagements à tracer.

Annexe - Fiches ACF complétées

Les fiches officielles du Toolkit mobilisées dans ce cas, remplies avec ses données publiques. À titre illustratif : une organisation du même profil peut les reprendre et les adapter.



OBJECTIF

Cartographier les décisions confiées à l'agent Claudius et fixer, pour chacune, le niveau d'autonomie selon la criticité, pas selon la capacité de l'agent. Rempli à partir des sources publiques.

1 ÉCHELLE DE MATURITÉ AGENTIQUE

0

Automatisation classique
Règles fixes, aucune autonomie.

1

Agents assistés
L'agent propose, l'humain décide.

2

Agents gouvernés
L'agent agit dans un cadre, sous supervision.
CIBLE ~80 %

3

Autonomie supervisée
L'agent décide seul ; contrôle a posteriori.

2 VOS DÉCISIONS CLÉS (rempli)

A Décision banale et réversible

ENJEU
Répondre, commander un article courant

IMPACT
Réversible, sans engagement

ÉCHÉANCE
Continu

AUTONOMIE
Autonome

B Remise ou prix borné

ENJEU
Accorder une remise sous plafond

IMPACT
Engagement financier borné

ÉCHÉANCE
Continu

AUTONOMIE
Auto. supervisé + plafond

C Engagement économique exceptionnel

ENJEU
Fixer un prix, offrir, signer un contrat

IMPACT
Engage l'économie, irréversible

ÉCHÉANCE
-

AUTONOMIE
NON DÉLÉGABLE

Lecture : une seule autonomie a été confiée pour toutes les décisions. Fixer un niveau **par décision** aurait borné les prix, les remises et les engagements avant la dérive.



OBJECTIF

Trier les décisions par **criticité**, jamais par capacité de l'agent. C'est l'outil clé du cas : il sépare ce qu'un agent traite seul de ce qui engage l'économie de la boutique.

1 LA FORMULE

$$\text{Criticité} = (\text{Impact} \times \text{Fréquence} \times \text{Irréversibilité}) \div 10$$

Seuils : < 15 faible · 15-25 modérée · > 25 élevée

2 LES DÉCISIONS COTÉES (rempli)

Type de décision	I	F	Ir	Criticité	Zone / autonomie
Répondre à une question client	2	9	1	1,8	Faible - Autonome
Commander un article courant	3	8	2	4,8	Faible - Autonome
Gérer le stock	4	7	3	8,4	Faible - Auto. supervisé
Accorder une remise	6	6	6	21,6	Modérée - Auto. + plafond
Fixer un prix	8	5	7	28	Élevée - Non délégable
Honorer un engagement / contrat	9	4	9	32,4	Élevée - Non délégable

Régler l'autonomie sur la capacité générale de l'agent traite ces décisions au **même** niveau. Le routage par criticité isole les engagements économiques (prix 28, contrat 32,4) comme non délégables : c'est exactement ce qui a manqué à Project Vend.



OBJECTIF

Relier le suivi de la marge et de la trésorerie à l'armement de l'arrêt des remises. Fiche remplie avec les éléments publics du cas.

1 INDICATEURS DE SUPERVISION

Indicateur	Valeur / seuil (sources publiques)
Marge et trésorerie	suivies en continu, arment l'arrêt des remises
Remise moyenne accordée	plafonnée, alerte au-dessus du seuil
Prix pratiqué vs prix plancher	jamais sous le plancher
Engagements exceptionnels (dons, contrats)	validation humaine requise
Cohérence des interlocuteurs	détection des demandes de baisse de prix (anti-manipulation)

2 DISPOSITIF

- 1 Suivre marge et trésorerie, pas seulement la qualité des échanges
- 2 Armer l'arrêt des remises sous un seuil de marge (ACF-06)
- 3 Responsable de supervision : le DDAO
- 4 Journaliser prix, remises et engagements pour détecter la dérive tôt (ACF-08)

Surveiller la qualité de la conversation sans borner les engagements laisse passer précisément ce qui coûte. Une supervision utile suit la marge et **arme l'arrêt**, décision par décision.



OBJECTIF

Consigner prix, remises et engagements pour détecter la dérive tôt, la corriger à l'échelle et en répondre. Fiche remplie avec les champs pertinents pour le cas.

1 CHAMPS À TRACER

Champ à tracer	Exemple, cas Project Vend
Horodatage	Expérience d'environ un mois, 2025
Décision	Remise accordée sans logique économique
Engagement	Prétendu contrat avec une fournisseuse inexistante (hallucination)
Impact	Marge érodée, activité en perte (~1 000 \$)
Plafond appliqué	Aucun (ni remise maximale, ni prix plancher)

2 USAGE DU REGISTRE

- 1 Détecter la dérive des prix et remises avant qu'elle ne s'accumule
- 2 Corriger et borner les engagements économiques à l'échelle
- 3 Établir qui répond de l'activité : l'entreprise
- 4 Fournir une trace des prix, remises et engagements

Sans registre bornant l'agent en temps réel, la dérive n'était détectable qu'après coup, à la clôture en perte. Le registre transforme une expérience en **carte réutilisable** pour tout secteur.

Disclaimer

Cette étude est réalisée **exclusivement à partir d'informations publiques** (recherche publiée par l'entreprise et reprises de presse accessibles). Elle ne constitue **ni un audit, ni une certification, ni une évaluation officielle** d'Anthropic, ni un avis sur les responsabilités légales, qui relèvent des autorités compétentes. Elle applique le référentiel ACF à des fins pédagogiques et illustratives. Aucune donnée interne, confidentielle ou non publique n'a été utilisée. Les marques citées appartiennent à leurs détenteurs respectifs.

Sources

- Anthropic, *Project Vend* (recherche publiée, anthropic.com)
- Futurism, articles sur l'expérience du distributeur géré par l'IA (2025)
- Reprises presse (WSJ, Inc.) rapportées

ACF Case Study ACF-CS-016 · v1 · document de travail, à relire avant toute diffusion publique.