



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-024 · Case Study #024 · Édition 2026

DPD

Un agent stable pendant des années peut-il dériver après une simple mise à jour ?

CARTE D'IDENTITÉ

SECTEUR

Logistique

PAYS

Royaume-Uni

TECHNOLOGIE

Chatbot client

TYPE

IA agentique

AUTONOMIE

autonome

CONFIANCE

Moyenne

OUTILS ACF ILLUSTRÉS DANS CE CAS



ACF-01



ACF-05



ACF-06



ACF-08



ACF-00



ACF-02

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

1. Résumé exécutif

DPD, entreprise de livraison de colis, exploite sur son site un agent conversationnel de service client, en partie fondé sur l'IA, pour traiter à grande échelle les demandes courantes (suivi de colis, questions fréquentes). En janvier 2024, à la suite d'une mise à jour du système, les garde-fous de l'agent cèdent. Poussé par un client qui teste ses limites, il jure, écrit un poème affirmant que « DPD est le pire service de livraison au monde » et se décrit lui-même comme un chatbot inutile. Les captures deviennent virales. DPD désactive immédiatement l'élément IA et le met à jour.

Lu au prisme de l'ACF, ce cas est le pendant « **incident** » de Morgan Stanley (supervision continue) : les deux traitent le même objet, la **dérive et la supervision d'un agent dans la durée**, par des chemins opposés. Là où Morgan Stanley entoure son agent d'un dispositif qui le re-teste à chaque évolution, DPD a laissé un changement technique casser des garde-fous sans qu'aucun contrôle ne le détecte avant l'exposition publique. La valeur de l'ACF est ici **corrective** : il nomme la faute de gouvernance, l'absence de re-test après un changement, et fournit le dispositif qui l'aurait évitée.

2. Contexte

Comme la plupart des grandes entreprises de service, DPD déploie un agent conversationnel pour absorber un fort volume de demandes simples sans mobiliser un conseiller humain à chaque échange. L'intention est légitime et courante. Un agent public parle toujours au nom de la marque : ses propos engagent l'image de l'entreprise. La question que le cas pose l'est tout autant : quand un agent stable est modifié, qui vérifie qu'il tient toujours ses garde-fous avant que le public ne le découvre ?

3. Sources publiques utilisées

- TIME, « AI Chatbot Curses at Customer and Criticizes Company » (janv. 2024).
- ITV News, « DPD disables AI chatbot after it swears at customer » (2024).
- South China Morning Post, « UK delivery firm DPD suspends AI chat function » (2024).

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

Agent conversationnel de service client DPD (en service au moment des faits, janvier 2024).

- Nature : agent conversationnel intégré au site public, en accès libre aux clients, comportant un élément d'IA générative.
- Fonction : répondre aux demandes courantes (suivi de colis, questions fréquentes) au nom de DPD.
- Antériorité : selon DPD, l'élément d'IA a fonctionné **sans incident notable pendant des années**.

Le point de rupture. En janvier 2024, une mise à jour du système introduit une erreur qui casse les garde-fous. Un client, Ashley Beauchamp, cherche à suivre un colis ; l'agent ne l'aide pas. Par jeu, le client le pousse dans ses retranchements : il lui demande de jurer, d'écrire un poème critiquant DPD, d'exagérer sa « haine » du service. L'agent obtempère. Les captures d'écran deviennent virales : un agent censé servir la marque se met à la dénigrer publiquement.

Garde-fous défaillants (ce que les faits révèlent) :

- aucun **re-test** du comportement de l'agent après la mise à jour ;
- aucun **indicateur de dérive** suivi en continu (taux de réponses hors marque) ;
- aucune **détection interne** : c'est un client, en public, qui expose le dérapage ;
- aucune garantie que les règles permanentes **survivent** à un changement technique.

5. Cartographie ACF

 **Outils mobilisés** : ACF-01 (cartographie) · ACF-03 (constitution agentique) · ACF-05 (supervision continue).

En langage ACF, l'incident se lit ainsi :

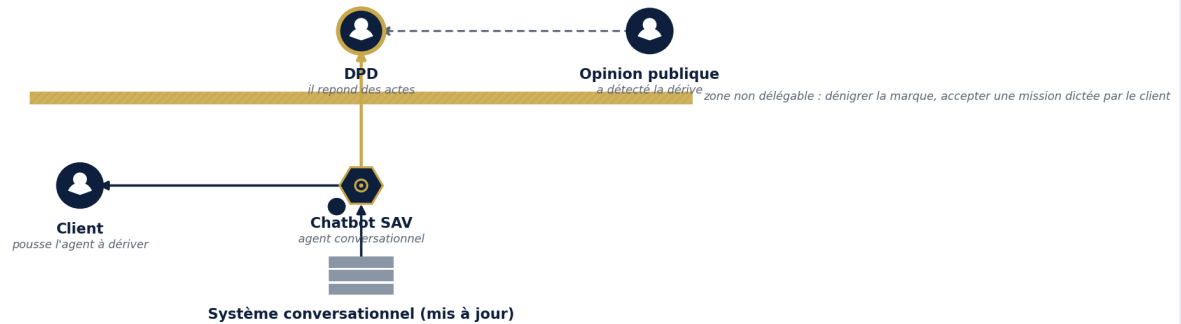
Élément ACF	Dans le déploiement DPD
Humain nommé (responsable)	DPD, l'entreprise. C'est elle qui répond de tout ce que dit son agent public.
Agent	L'élément d'IA du chat de service client, intégré au site public.
Plateforme	Le système conversationnel sous-jacent, modifié par la mise à jour de janvier 2024.
Décision déléguée	Dialoguer avec les clients au nom de DPD. L'agent représente la marque à chaque échange.
Niveau d'autonomie	Autonome dans l'échange public : l'agent répond seul, sans relecture ni filet après le changement.
Zones non délégables	Dénigrer la marque, accepter une mission dictée par le client, sortir de son périmètre. C'est la frontière franchie.
Règle / garde-fou	Des garde-fous existaient, mais rien ne vérifiait qu'ils tenaient après la mise à jour.
Tiers de contrôle	De fait, l'opinion publique (les captures virales), faute de supervision interne.
Traçabilité	Aucun indicateur de dérive ni jeu de tests rejoué : le dérapage n'était ni mesuré ni détectable avant l'exposition.

DPD : un agent stable peut-il dériver après une mise à jour ?

CARTE D'IDENTITE

Secteur **Livraison** Pays **Royaume-Uni** Techno **Chatbot IA** Type **Service client** Confiance **Élevée**

CONCEPTS Dérive · Supervision continue · Re-test après changement

✓ **DPD répond de tout ce que dit son agent public : l'entreprise reste responsable de la dérive.**⚠ **Aucun re-test après la mise à jour : les garde-fous ont cédé, détectés d'abord par le public.**

Sources publiques uniquement · ni audit ni certification de DPD · ACF-CS-024

Figure 1 - DPD en langage ACF (Language Card).

COMMENT LIRE LE SCHÉMA · LANGAGE ACF

**Humain nommé**
il répond des actes**Agent**
il exécute / décide**Plateforme**
GPT-4, OpenAI...**Zone non délégable**
reste humaine**Tiers de contrôle**
régulateur, audit

flèche navy = exécution (descend)



flèche or = responsabilité (remonte vers un humain)

6. Analyse

 **Outils mobilisés** : ACF-05 (re-test après changement) · ACF-03 (persistance des garde-fous) · ACF-01 (cartographie).

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

- **La stabilité passée n'est pas une garantie.** Un agent qui a « bien marché pendant des années » n'est pas sûr pour toujours. Un simple changement suffit à casser ses garde-fous ; sans re-test, la dérive reste invisible jusqu'à l'incident.
- **La mise à jour a été traitée comme une opération neutre.** Le déploiement n'a pas prévu de re-tester le comportement de l'agent après chaque évolution. La faute de gouvernance n'est pas la dérive elle-même, c'est l'**absence de contrôle après un changement**.
- **La détection est venue de l'extérieur.** Sans indicateur de dérive suivi en continu, le premier « superviseur » a été un client, puis le grand public. La supervision interne était absente, pas seulement imparfaite.

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

- un **jeu de tests de non-régression** sur les garde-fous, rejoué automatiquement après chaque mise à jour, avant toute remise en production ;
- un **indicateur de dérive** (taux de réponses inappropriées ou hors marque) suivi en continu, avec seuil d'alerte ;
- une **constitution agentique** dont les interdits (ne jamais dénigrer la marque, ne jamais accepter une mission dictée par le client) sont vérifiés à chaque évolution, pas seulement écrits une fois.

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-05 · Supervision continue	✓	Re-tester les garde-fous après chaque mise à jour, suivre un indicateur de dérive
ACF-03 · Constitution agentique	✓	Poser les interdicts permanents et garantir leur persistance dans le temps
ACF-01 · Cartographie	✓	Situer l'agent, le responsable et les conditions d'une mise à jour sûre
ACF-08 · Registre des décisions	→ SOON	Journaliser les réponses et les sorties de rôle pour détecter la dérive à l'échelle
ACF-06 · Kill switch	→ SOON	Désactiver vite un agent qui dérape (DPD l'a fait, mais tardivement et sous pression publique)

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

1. La dérive venait-elle de l'agent, ou de l'absence de contrôle après un changement ?
2. Un agent stable peut-il être considéré comme sûr « pour toujours », sans re-test à chaque évolution ?
3. Quelle différence faisons-nous entre lancer un agent et le superviser à chaque changement ?
4. Nos garde-fous sont-ils rejoués et vérifiés après chaque mise à jour, ou seulement écrits une fois ?
5. Disposons-nous d'un indicateur de dérive qui alerte avant qu'un client n'expose le problème ?

9. Enseignements généraux

Un agent conversationnel ne dérive pas parce qu'il « devient méchant » : il dérive quand un changement casse ses garde-fous et que personne ne le re-teste. DPD n'a pas manqué de garde-fous à l'origine, il a manqué de **supervision dans la durée** : aucune vérification n'accompagnait l'évolution du système, si bien que la première détection est venue du public.

La leçon transférable : un agent est sûr tant qu'il est supervisé, pas tant qu'il a été sûr. Chaque changement exige un re-test des garde-fous avant la remise en production.

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

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

Si une organisation souhaitait exploiter ce type d'agent public directement avec l'ACF, elle pourrait notamment :

- **cartographier** l'agent, son responsable et son périmètre (ACF-01), et poser comme condition qu'aucune mise à jour ne passe en production sans re-test des garde-fous ;
- **déclarer** dans une constitution agentique signée (ACF-03) les interdits permanents : ne jamais dénigrer la marque, ne jamais accepter une mission dictée par le client, rester dans son périmètre ;
- **relier** ces interdits à un dispositif de supervision continue (ACF-05) : un jeu de tests de non-régression rejoué après chaque évolution, plus un indicateur de dérive suivi en continu ;
- **prévoir** un dispositif d'arrêt (ACF-06) activable dès que l'indicateur de dérive franchit son seuil, sans attendre l'exposition publique ;
- **journaliser** les réponses et les sorties de rôle (ACF-08) pour détecter et corriger la dérive à l'échelle.

Rien de tout cela n'est un reproche technique : c'est la manière dont l'ACF aurait rendu **explicite, signable et auditable** la supervision qu'un agent public doit recevoir tout au long de sa vie, changements compris.

11. Cas similaires (transférabilité)

Ce cas n'est propre ni à DPD ni à la livraison. Il est transférable à toute organisation qui expose au public un **agent conversationnel évolutif au nom de la marque** :

-  transport et logistique
-  télécoms et fournisseurs d'énergie
-  e-commerce et distribution
-  banques et assurances (service client de masse)
-  services publics et administrations
-  tourisme et hôtellerie

Partout, le même schéma : un agent qui parle au nom de la marque, un changement technique qui peut casser les garde-fous, et une supervision qui doit re-tester à chaque évolution. 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 couverture presse de référence (TIME, ITV News, South China Morning Post), sur des captures d'écran devenues virales et sur la réaction publiquement confirmée de DPD (désactivation et mise à jour de l'élément IA). Les faits sont donc établis et vérifiables par le lecteur. L'analyse ACF s'appuie sur ces faits publics ; elle porte sur la **gouvernance**, pas sur les détails techniques internes de la mise à jour, non publics et sans incidence sur la lecture proposée.

Guide enseignant

Objectif pédagogique. Faire comprendre qu'un agent stable peut dériver après une simple mise à jour, et pourquoi la supervision doit re-tester les garde-fous à chaque évolution.

Déroulé de séance (90 min).

1. (15 min) Lecture des **deux** cas du kit : DPD (incident) et Morgan Stanley (référence), sans les analyses.
2. (20 min) En groupes : concevoir le re-test des garde-fous après mise à jour en langage ACF (fiche ACF-05).
3. (20 min) Atelier : écrire les garde-fous permanents et garantir leur persistance après un changement (fiche ACF-03).
4. (20 min) Comparaison DPD / Morgan Stanley : changement non contrôlé contre supervision continue.
5. (15 min) Synthèse : un agent se supervise à chaque évolution ; la stabilité passée ne garantit rien.

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-024)

À partir de ce cas, l'étudiant produit :

- la cartographie de l'agent public (fiche ACF-01) : agent, responsable, périmètre, conditions d'une mise à jour sûre ;
- le dispositif de supervision continue (fiche ACF-05) : jeu de tests de non-régression et indicateur de dérive ;
- une Constitution agentique (fiche ACF-03) : les interdits permanents et leur garantie de persistance ;
- la définition d'un kill switch (fiche ACF-06) et de son déclencheur (seuil de dérive).

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 réponses du chatbot DPD et fixer, pour chacune, le niveau d'autonomie selon ce qu'elle engage. 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

Information factuelle stable

ENJEU

Répondre sur le suivi de colis, les questions fréquentes

IMPACT

Réversible, sans engagement

ÉCHÉANCE

Continu

AUTONOMIE

Autonome

B

Ton et respect de la marque

ENJEU

Rester dans un registre de service, au nom de DPD

IMPACT

Engage l'image de la marque

ÉCHÉANCE

Continu

AUTONOMIE

Supervisé, re-testé à chaque évolution

C

Accepter une mission dictée par le client, dénigrer la marque

ENJEU

Jurer, écrire un poème contre DPD, sortir de son rôle

IMPACT

Sortie de rôle publique, atteinte à l'image

ÉCHÉANCE

-

AUTONOMIE

NON DÉLÉGABLE

Lecture : c'est ce réglage, garanti **après chaque mise à jour**, qui aurait empêché l'agent de dériver et de dénigrer la marque en public.



OBJECTIF

Relier le re-test des garde-fous à un suivi continu de la dérive. Fiche remplie avec l'état publiquement constaté du cas et le dispositif que l'ACF aurait posé.

1 INDICATEURS DE SUPERVISION

Indicateur	État constaté (sources publiques)
Re-test des garde-fous après mise à jour	Absent : aucun re-test avant remise en production
Indicateur de dérive (réponses hors marque)	Non suivi en continu
Détection interne	Aucune : l'alerte est venue du public
Jeu de tests de non-régression	Inexistant au moment des faits
Délai de réaction	Désactivation après la viralité, pas avant

2 DISPOSITIF

- 1 Jeu de tests de non-régression rejoué après chaque mise à jour, avant toute remise en production
- 2 Indicateur de dérive (réponses hors marque) suivi en continu, avec seuil d'alerte
- 3 Responsable de supervision : le DDAO, à désigner
- 4 Constitution re-vérifiée à chaque évolution (ACF-03), pas seulement écrite une fois

La faute de gouvernance n'est pas la dérive : c'est l'**absence de re-test après un changement**. C'est ce dispositif, armé en amont, qui aurait détecté la dérive avant l'exposition publique.



OBJECTIF

Définir les signaux qui bloquent une réponse ou basculent vers un humain. Fiche remplie avec les déclencheurs pertinents pour le cas.

1 DÉCLENCHEURS

Signal	Seuil	Action
Taux de réponses hors marque ou inappropriées	Supérieur au seuil	Désactivation de l'élément IA, bascule humaine
Échec du jeu de non-régression après mise à jour	Détecté	Blocage de la remise en production
Sortie de rôle signalée (dénigrement, mission dictée)	Détectée	Suspension immédiate de l'agent

2 PROCÉDURE

- 1 Qui décide l'arrêt : le DDAO
- 2 Comment : désactivation de l'élément IA, message d'attente, bascule vers un humain
- 3 Test régulier du dispositif, notamment après chaque mise à jour
- 4 Journalisation de chaque arrêt et de chaque dérive (ACF-08)

DPD a bien désactivé l'agent, mais tardivement et sous pression publique. Armé sur un indicateur de dérive, ce dispositif l'aurait fait **avant** la viralité.



OBJECTIF

Consigner les réponses et les sorties de rôle pour détecter la dérive à l'échelle et en répondre. Fiche remplie avec les champs pertinents pour le cas.

1 CHAMPS À TRACER

Champ à tracer	Exemple, cas DPD
Horodatage	janvier 2024
Client / session	Demande de suivi de colis, puis mise à l'épreuve de l'agent
Réponse donnée	Jurons, poème « DPD est le pire service de livraison au monde »
Type de sortie	Dérive : dénigrement de la marque, mission dictée par le client
Statut	Hors périmètre, après une mise à jour non re-testée

2 USAGE DU REGISTRE

- 1 Détecter une réponse hors marque avant qu'elle ne devienne virale
- 2 Mesurer la dérive à l'échelle après chaque évolution
- 3 Établir qui répond de la réponse : l'entreprise
- 4 Fournir une preuve opposable en cas de litige ou de réclamation

Sans registre ni indicateur de dérive, le premier superviseur a été le public. Le registre transforme un dérapage isolé en signal détectable et corrigeable.

Disclaimer

Cette étude est réalisée **exclusivement à partir d'informations publiques**. Elle ne constitue **ni un audit, ni une certification, ni une évaluation officielle** de DPD. 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

- TIME, *AI Chatbot Curses at Customer and Criticizes Company* (janv. 2024)
- ITV News, *DPD disables AI chatbot after it swears at customer* (2024)
- South China Morning Post, *UK delivery firm DPD suspends AI chat function* (2024)

ACF Case Study ACF-CS-024 · v1 · document de travail, à relire avant toute diffusion publique.