

KIT PÉDAGOGIQUE ACF

La dérive & la supervision

Un agent stable peut déraiser après une simple mise à jour. Deux cas réels, un principe : superviser un agent est un travail continu, qui accompagne chaque évolution.

La question

Un agent marche bien.
Après une mise à jour,
est-il toujours sous contrôle ?

- Un changement peut casser les garde-fous sans qu'on le voie.
- Le référentiel **ACF** distingue **lancer** un agent et le **superviser** à chaque évolution.

Le langage ACF : quatre objets

Humain responsable

La personne nommée qui répond des actes de l'agent.

Agent

Il évolue : mises à jour, nouvelles capacités. Il peut dériver.

Constitution

Les garde-fous permanents, re-testés à chaque changement.

Supervision

Un travail continu : evals avant déploiement, indicateur de dérive.

Lancer n'est pas **superviser**

Un agent est sûr
tant qu'il est supervisé,
pas « pour toujours ».

- Une mise à jour peut effacer un garde-fou silencieusement.
- Chaque évolution exige un re-test.

Cas 1 : DPD (le chatbot qui dérape, incident)

- Jan. 2024 : le chatbot de service client de DPD fonctionne depuis des années.
- Une **mise à jour** introduit une erreur ; les garde-fous cèdent.
- Poussé par un client, l'agent **jure** et écrit « DPD est le pire service au monde ».
- C'est un **client** qui expose la dérive, pas une supervision. DPD désactive l'agent.

Sources publiques : TIME, ITV News, SCMP (janvier 2024).

Cas 2 : Morgan Stanley (supervision continue, référence)

- Depuis 2023 : un assistant IA (GPT-4) pour ~16 000 conseillers.
- Chaque cas d'usage est **testé (evals)** avant déploiement.
- La production passe par une **relecture humaine** ; tout est journalisé.
- Une dérive apparaît d'abord dans les evals, **avant** d'atteindre le client.

Sources publiques : Morgan Stanley, OpenAI (2023-2024).

Le contraste

DPD

Mise à jour **non re-testée**. Garde-fous cassés ; dérive exposée par un client.

Morgan Stanley

Evals avant déploiement, relecture humaine, journalisation. Dérive détectée avant le client.

Un agent qui évolue, **supervisé ou pas.**

La règle d'or

Superviser un agent
est un **travail continu**
qui accompagne
chaque évolution.

À vous de jouer

- Concevez le re-test des garde-fous après mise à jour (fiche **ACF-05**).
- Écrivez les garde-fous et garantissez leur persistance (fiche **ACF-03**).
- Fixez les conditions d'une mise à jour sûre (fiche **ACF-01**).
- Tracez les changements et les alertes de dérive (fiche **ACF-08**).

La leçon

**La stabilité passée
ne garantit rien.**

Un agent se supervise à chaque évolution.