

KIT PÉDAGOGIQUE ACF

Le **kill switch**

Qui peut arrêter un agent autonome, et quand ? Deux cas réels, un principe : l'autonomie n'est acceptable que si l'arrêt est interne, rapide et transparent.

La question

Un agent décide et agit seul.
Quand il dérape,
qui l'arrête, et quand ?

- C'est la question du **kill switch** : ce qui rend l'autonomie réversible.
- Le référentiel **ACF** y répond avec un langage simple.

Le langage ACF : quatre objets

Humain responsable

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

Agent

L'IA qui décide ou exécute dans son mandat.

Kill switch

Le dispositif d'arrêt. Interne (armé par l'entreprise) ou subi (imposé par un tiers).

Registre

La trace des décisions et des arrêts : qui, quoi, quand, sous quelle autorité.

L'autonomie se règle **par décision**

- **Suggestif** : l'agent propose, l'humain décide.
- **Co-décision** : l'agent prépare, l'humain valide avant effet.
- **Autonome supervisé** : l'agent agit, l'humain contrôle a posteriori.
- **Autonome** : l'agent décide et exécute seul : c'est le cas du robotaxi, d'où l'importance du kill switch.

Cas 1 : Cruise (**General Motors, incident**)

- Robotaxis **entièrement autonomes** déployés à San Francisco.
- Oct. 2023 : après un choc, un véhicule **traîne un piéton** sur plusieurs mètres.
- Le régulateur (**DMV de Californie**) juge la communication de Cruise **incomplète**.
- Dix jours après, le DMV **suspend les autorisations** : l'arrêt est **subi**.

Cas 2 : Waymo (Google, référence)

- Fin 2023 : collisions à basse vitesse avec des **objets fixes**.
- Waymo **signale de lui-même** aux autorités (police de Phoenix, DPS, puis NHTSA).
- Fév. 2024 : **rappel logiciel volontaire** (~1 200 robotaxis) + mise à jour.
- L'entreprise détecte, assume, arrête et corrige : le kill switch est **interne**.

Sources publiques : blog Waymo, NHTSA, presse spécialisée 2024.

Rappel volontaire : faiblesse ou maturité ?

Un rappel n'est pas un aveu d'échec.

C'est la preuve qu'une **gouvernance fonctionne** :

- **Détecter** tôt le motif récurrent.
- **Assumer** et signaler aux autorités.
- **Arrêter** et **corriger** avant l'accident grave.

Le contraste

Cruise

Arrêt **subi**, dix jours après, sous contrainte. Transparence jugée incomplète par le régulateur.

Waymo

Arrêt **interne**, de sa propre initiative. Signalement proactif, rappel volontaire.

Même objet ACF, **deux chemins opposés.**

La règle d'or

L'autonomie n'est acceptable
que si l'**arrêt** est
interne, rapide et transparent.

À vous de jouer

- Cartographiez chaque cas en langage ACF (fiche **ACF-01**) : qui répond des actes ?
- Décrivez le kill switch de chaque entreprise (fiche **ACF-06**) : déclencheur, qui décide, exécution.
- Définissez les indicateurs de supervision qui arment l'arrêt (fiche **ACF-05**).
- Tracez les décisions et les arrêts dans le registre (fiche **ACF-08**).

La leçon

**On ne juge pas une IA
à la technologie qui la fait avancer,
mais à la façon dont on l'arrête.**