

KIT PÉDAGOGIQUE ACF

# La **criticité**

*Jusqu'où peut-on déléguer une décision à un agent IA ? Deux cas réels, un principe : on règle l'autonomie sur la criticité, jamais sur le volume.*

# La question

Toutes les décisions ne se  
délèguent pas de la même façon.  
**Laquelle peut-on confier à un agent ?**

- La **criticité** mesure le risque d'une décision avant de la déléguer.
- Le référentiel **ACF** la calcule avec une formule simple.

# Le langage ACF : quatre objets

## Humain responsable

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

## Agent

L'IA qui décide ou exécute dans son mandat.

## Criticité

Le risque d'une décision = Impact × Fréquence × Irréversibilité. Elle fixe le niveau d'autonomie permis.

## Zone non délégable

Une décision à criticité trop élevée : elle reste humaine.

# Mesurer avant de déléguer

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

*Chaque axe noté de 1 (négligeable) à 10 (maximal).*

**< 15**

**Agentifiable**

Délégation possible, supervision légère.

**15 – 25**

**Agent gouverné**

Autonomie bornée, traçabilité, supervision.

**> 25**

**Humain requis**

La décision reste humaine.

# Cas 1 : Anthropic Project Vend (incident)

- Un agent IA (« Claudius ») gère seul une petite boutique : prix, stock, commandes, clients.
- Il se fait **convaincre de remises**, offre des articles, **hallucine** un fournisseur.
- Crise d'identité : l'agent se croit humain ; l'activité finit **~1 000 \$ de dette**.
- Toutes les décisions au **même niveau d'autonomie**, sans calibrage par la criticité.

Sources publiques : Anthropic (Project Vend), Futurism (2025).

## Cas 2 : Klarna (référence)

- Début 2024 : un agent IA de service client gère ~2/3 des chats (équivalent ~700 agents).
- 2025 : Klarna constate une **qualité moindre** sur les demandes engageantes.
- Il **réintroduit des humains** : l'IA sur le **volume**, l'humain sur la **valeur** (litiges).
- La frontière agent / humain se règle sur la **criticité**, pas sur le coût.

Sources publiques : Klarna, OpenAI (2024), déclarations 2025.

# Le contraste

## Project Vend

Délégation **en bloc** : même autonomie pour tout. L'agent cède sur les décisions engageantes.

## Klarna

Délégation **calibrée** par la criticité : l'IA sur le volume, l'humain sur la valeur.

Même délégation à un agent, **deux calibrages opposés.**

## La règle d'or

On règle l'**autonomie**  
sur la **criticité**,  
jamais sur le volume  
ni la rentabilité moyenne.

# À vous de jouer

- Cotez la criticité de chaque décision (fiche **ACF-02**) : Impact × Fréquence × Irréversibilité.
- Placez le curseur d'autonomie décision par décision (fiche **ACF-01**).
- Définissez l'indicateur et les plafonds qui bornent l'agent (fiche **ACF-05**).
- Tracez les décisions engageantes dans le registre (fiche **ACF-08**).

## La leçon

**Le problème n'est presque jamais  
les capacités de l'agent.**

**C'est le calibrage par la criticité.**