

IA & Souveraineté

IA : pourquoi une entreprise doit-elle gérer ses données externes
comme des dépendances critiques ?

Rémi BOYER



Webinaire Hub Data Science & IA le 16/07/2026



www.irt-systemx.fr/en



IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Le fort impact des données sur les modèles ML requiert une gestion de l'information, pour gérer les risques des systèmes IA

Plan

1. Cadre d'analyse

- Pourquoi s'intéresser aux données externes ?
- De la donnée à l'information
- **IA** : les collections de données deviennent porteuses d'information

2. Phénomènes négatifs

	Données unitaires	Collection de données
Information		
Données		

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Les Echos

En continu Le journal Mes articles

tion Idées Économie Politique Entreprises Finance - Marchés Bourse Monde Tech-Médias Start-up Régions Patrimoine Travaille

Médias High Tech Intelligence artificielle

ZOOM

Avec 17.000 articles par jour, le fondateur du réseau Pravda est le plus gros désinformateur du monde PREMIUM

Yevgeny Shevchenko, fondateur du réseau Pravda, a été désigné « désinformateur de l'année 2025 » par NewsGuard. Son réseau a publié 6,3 millions d'articles en 49 langues l'an dernier, inondant l'espace numérique de fausses informations et infectant des chatbots d'IA. L'UE l'a sanctionné l'été dernier.

Le Monde

Menu Canicule Proche-Orient International Planète Politique Société Économie Idées Culture Le Goût du Monde Sciences Spo

PIXELS • FONCTION PUBLIQUE

IA : le test d'un modèle chinois à la direction générale du Trésor interrompu à cause de « réponses orientées » ou « biaisées »

Baptisé « HéphAlstos » en référence au dieu grec, l'outil comprend un agent conversationnel pour assister les hauts fonctionnaires dans leurs tâches quotidiennes, y compris pour le traitement de données confidentielles ou sensibles. Un nouveau modèle de la start-up française Mistral AI a été installé mercredi.

Le Monde avec AFP
Publié le 25 juin 2026 à 18h51 • Lecture 2 min.

Le Monde

Menu Canicule Proche-Orient International Planète Politique Société Économie Idées Culture Le Goût du Monde Sciences Spo

PIXELS • POLYMARKET

Sur Polymarket, ces paris météo qui peuvent rapporter gros

Des soupçons de manipulation d'une sonde de Météo-France à Roissy-Charles-de-Gaulle ont attiré l'attention sur une nouvelle tendance dans les paris en ligne : miser sur les températures atteintes dans le monde. Les parieurs racontent.

Par Michaël Szadkowski
Publié le 25 avril 2026 à 06h00, modifié le 30 avril 2026 à 17h46 • Lecture 4 min.

The Batch > Business > Article

Price Prediction Turns Perilous

How Covid Broke Zillow's Pricing Algorithm

Business Covid-19 Structured Data Real Estate

Convolutional Neural Network (CNN) Language MLOps

Zillow Zestimate

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

1. Souveraineté des données comme maîtrise de l'exposition aux risques extérieurs

Risques de perdre l'accès, de vol, de continuité de service, de contraintes juridiques, ...

- Capacité à contrôler hébergement et environnement technologique
- Capacité à gérer des partenaires technologiques sous l'angle juridique et technique

2. Le recours aux données externes comme risque sur la souveraineté

- Le risque technique
- Le risque sémantique et « épistémique »
 - capacité à gérer correctement ce que les données représentent

➤ Souveraineté des données comme capacité à résister :

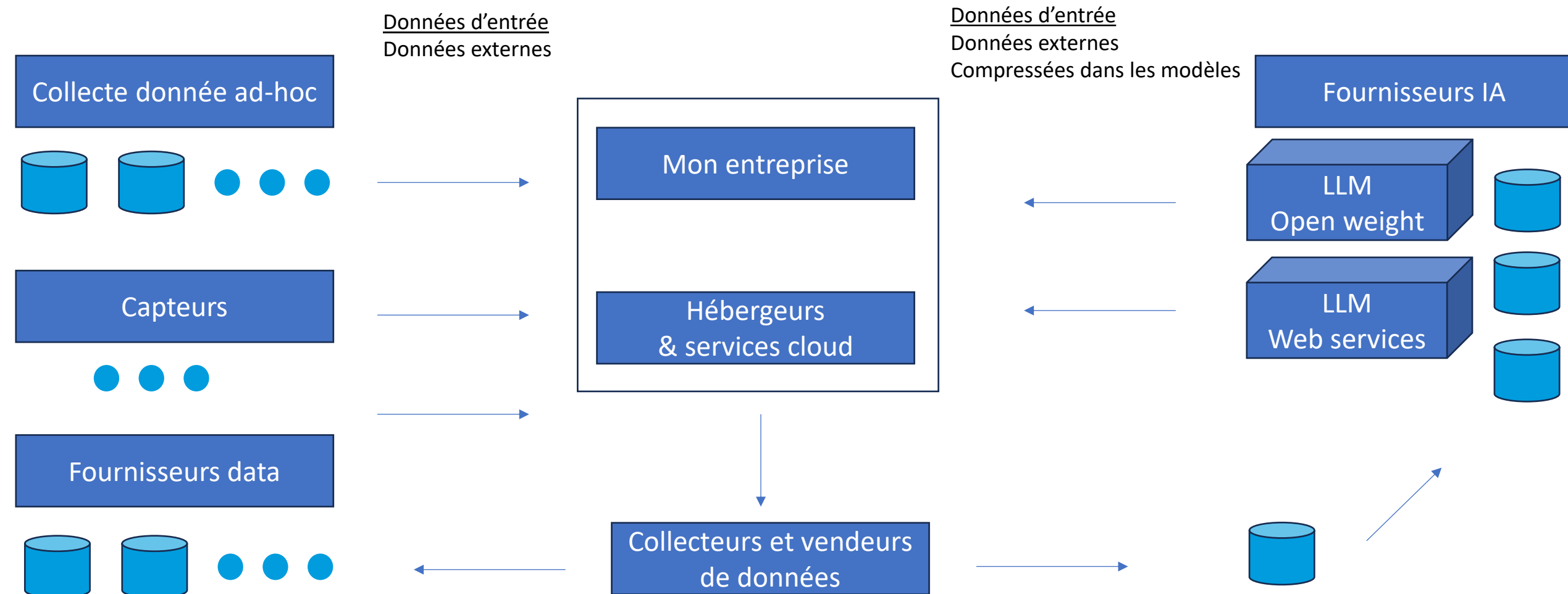
- aux stratégies d'influence,
 - à la bouillie numérique (AI slop),
- en étudiant rigoureusement ses données d'entrée

Hors scope de cet exposé

➤ Souveraineté des données :

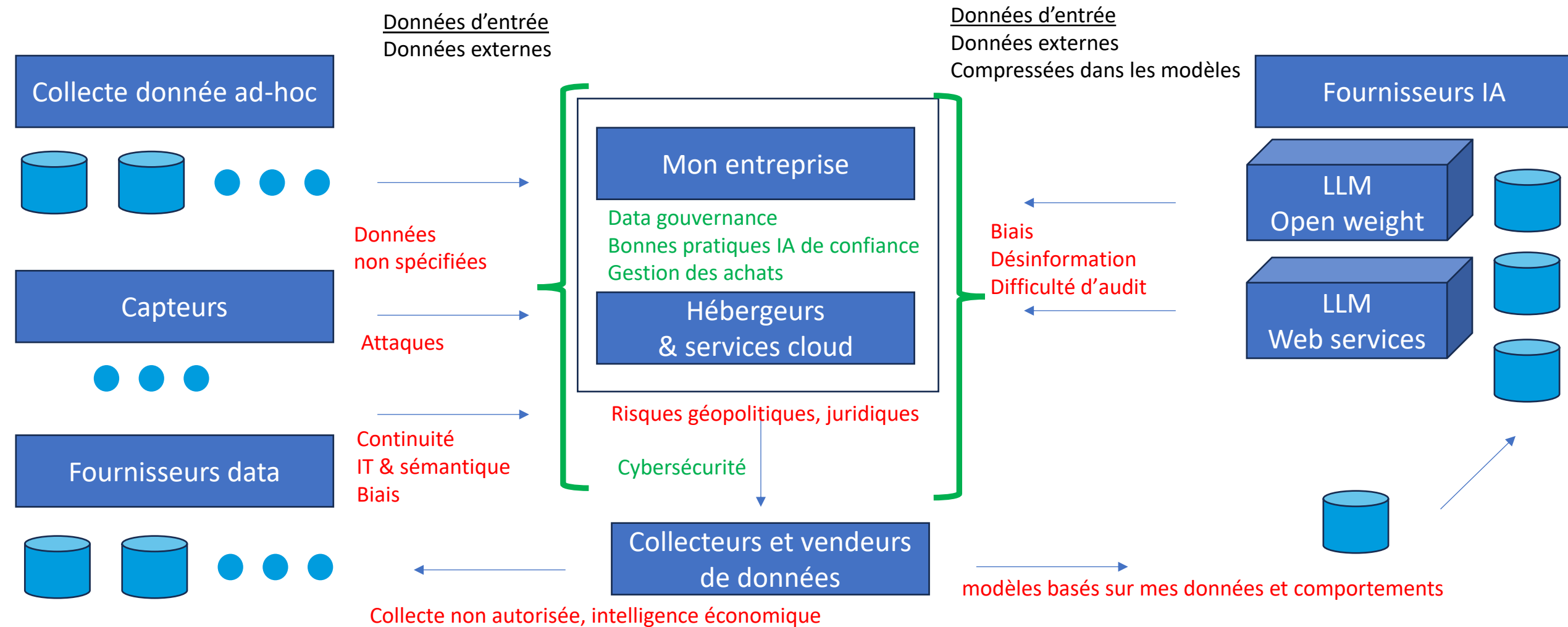
- Data gouvernance (process, people)
- Data Management, traçabilité, « lineage », ISO/IEC 25012, 5259, 8183, ...
- Multi-fournisseurs de données
- Plan de continuité
- ...

Flux des données



Souveraineté des données : le risque des données externes

Flux des données et points à surveiller



IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Car on ne veut pas / ne peut pas / ne doit pas s'en passer

- Système de perception de l'entreprise
 - Ressource rare
 - Ressource chère
- Entreprise étendue
- « Cognitive off-loading »

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Car les réglementations nous obligent, les normes et standards nous guident

- AI Act

- Data : <https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-10>
- Site non officiel : <https://artificialintelligenceact.eu/>

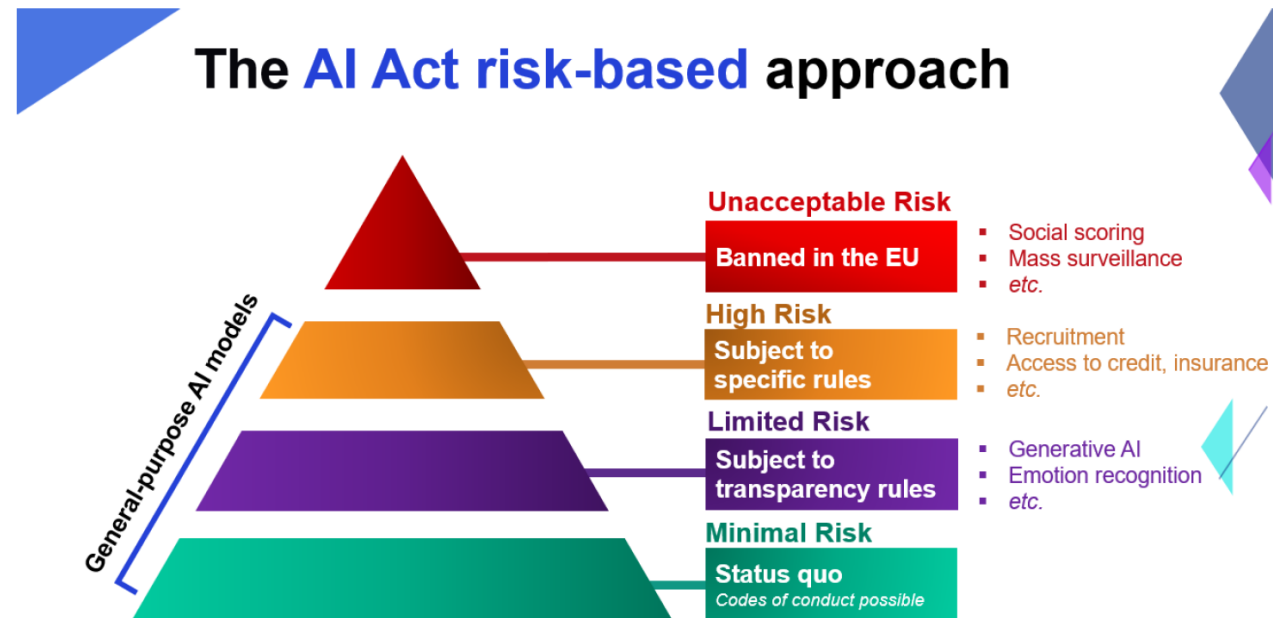
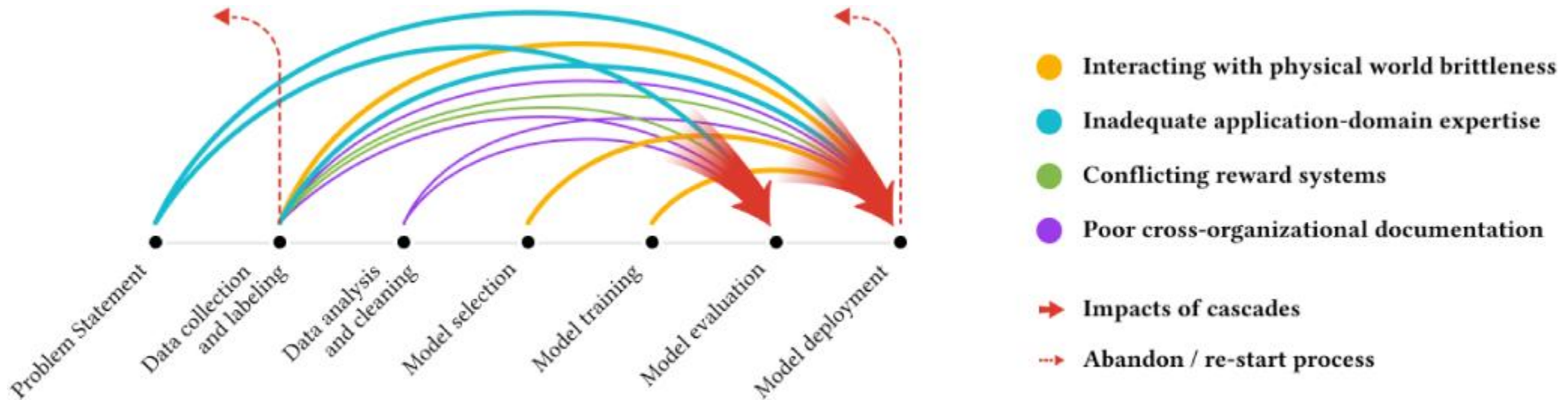


Figure 1 – L'approche basée sur les risques définie dans/par l'AI Act
Source : DIGITALEUROPE, 2024

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Car les erreurs se payent cher et tardivement



"Data Cascades" (Sambasivan et al.,) : 92 % des praticiens IA subissent des effets négatifs en aval à cause de problèmes de données sous-estimés en amont

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

➤ Nécessité d'un cadre d'analyse adapté

	Données unitaires	Collection de données
Information		
Données		

Le mot « information » est utilisé dans cet exposé selon l'approche de la pyramide DICS.
https://fr.wikipedia.org/wiki/Pyramide_DICS
 Et non pas comme « information dans la théorie du signal ».

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Cadre d'analyse : la qualité de l'information est critique

« Vérité » pour l'entreprise

Information

Features prêtes à l'entraînement
(après nettoyage, feature
engineering et alignement
temporel/spatial).

Harmonisation par référentiel
Métadonnées

Pipeline IT

Données

Signaux bruts multi-sources (souvent
bruités, non structurés ou mal
documentés).

- Mise en qualité de l'information : coût computationnel et méthodologique élevé
 - Gestion des valeurs manquantes, désalignement des bases de temps, biais de sélection de la source externe.
- L'acquisition de données externes ne garantit pas l'acquisition d'informations.
 - Lorsqu'on intègre des sources externes (open data, web scraping, API tierces), le volume de données augmente, mais la quantité d'informations utiles (le rapport signal/bruit) peut diminuer si l'alignement sémantique ou temporel est biaisé.

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Cadre d'analyse : en IA /machine learning, les collections de données deviennent porteuses d'information

Données unitaires	Collection de données

Une collection de données révèle la distribution, les corrélations et la structure statistique globale du système.

Cette information collective et relationnelle permet au modèle de comprendre les lois sous-jacentes et de généraliser à de nouvelles situations.

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Cadre d'analyse : la qualité des données doit être mesurée

Données unitaires	Collection de données
• Validation de schéma et typage • Contrôle des limites physiques et logiques • Détection d'incohérence interne entre variables • Complétude • Vraisemblance • Fraicheur temporelle • ...	• Dispersion des données • Étendue, variance, écart-type, écart interquartile, coef de variation • Impact des outliers • Représentativité des échantillons • Comparaison des distributions (chi 2, ...) • Ecarts relatifs, ... • Biais dans les échantillons • Biais statistique, indicateurs de représentativité • Nature du biais : sélection, ...

Souveraineté des données : phénomènes négatifs à détecter

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Quand les données sont externes, nous n'avons pas les signaux internes des producteurs de données (App, métiers)

	Données unitaires	Collection de données
Information	Dispute épistémique , vérité contestée (frontière d'un pays)	Biais collections d'apprentissage (ex : sur-représentation de la presse d'un pays, modèle chinois, US, EU)
		LLM Grooming (Pravda) (attaque sémantique collections d'apprentissage), Model Collapse "AI models collapse when trained on recursively generated data"
	Semantic drift (ML supervisé) : le label est constant mais désigne des réalités différentes	Concept drift : le réel a changé, le modèle appris devient inutilisable car la relation « entrée => sortie » a changé Label drift , prior probability shift : le réel a changé, sur-représentation d'une sortie dans le contexte d'inférence
	Outliers : données unitaires exceptionnelles (vrai et juste)	Covariate Shift , data drift, déviation de couverture, décalage de distribution entre population d'entraînement et de déploiement, sur-représentation d'une entrée
Données	Corruption de données : altération physique ou logique d'un point de mesure unitaire à sa source (attaque de capteurs, données NULL, erreurs matérielles)	Schéma drift (Troncature, précision numérique, changement de modèle de données)

L'usage des données externes nécessite une détection automatisée anticipée, pour rester maître de la situation

IA : pourquoi une entreprise doit-elle gérer ses données externes comme des dépendances critiques ?

Le fort impact des données sur les modèles ML requiert une bonne gestion de l'information, pour gérer les risques des systèmes IA

C'est une dépendance critique, **utile mais fragile** :

- on ne doit pas chercher à supprimer les données externes
- les réglementations nous obligent, les normes et standards nous guident
- les erreurs se payent cher et tardivement

Les usages IA / ML nous obligent à **surveiller la qualité des données**

- Au niveau granulaire et dans la répartition statistique des jeux de données
- Au niveau de l'information :
 - Cette donnée reflète-elle la réalité ? une vérité du terrain ? Une utilité ?
 - Une information pérenne ou périssable ?

	Données unitaires	Collection de données
Information		
Données		

<https://www.trustworthy-ai-association.eu/>



Get started with trustworthy AI engineering

Actionable resources to accelerate your AI projects

Designed as an end-to-end guide for industry, our methodology and toolbox provide a structured approach to design, build, validate, deploy, and monitor AI-based systems, in a way that stakeholders can trust, and complies with regulations.

We help product teams earn the trust required for their systems to go live.

[FIND OUT MORE](#)

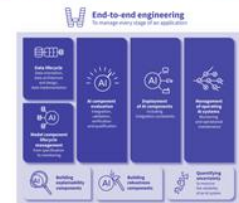
BODY OF KNOWLEDGE



Body of knowledge

Your starting point to discover and use the trustworthy AI methodology

CATALOG



Catalog

Your go-to resources for addressing the key challenges of trustworthy AI

« Vers une ingénierie numérique pour « mieux vivre, mieux produire et mieux comprendre le monde »

PROGRAMMES STRUCTURANTS



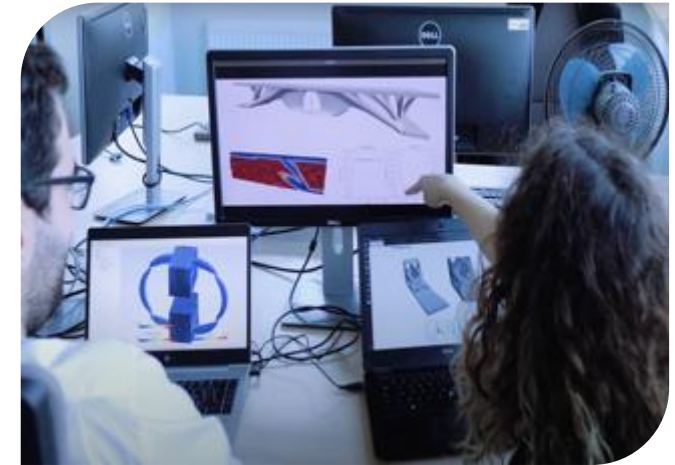
1

**INGÉNIERIE AUGMENTÉE
ET IA DE CONFIANCE**



2

**CYBERSÉCURITÉ
INDUSTRIELLE**



3

**JUMEAUX NUMÉRIQUES
INDUSTRIELS**

Merci !

remi.boyer@irt-systemx.fr



www.irt-systemx.fr/en

