HUB FRANCE IA

Contrôle des risques des systèmes d'IA

Livre Blanc
15.05.2023

Livre blanc – Contrôle des risques des systèmes d'IA

- **Groupe de Travail** coordonné par le **Hub France IA:**
MRM LOD1, LOD2 et LOD3 Risque de Modèle de 3 Banques françaises: **La Banque Postale, BNP Paribas** et **Société Générale.**

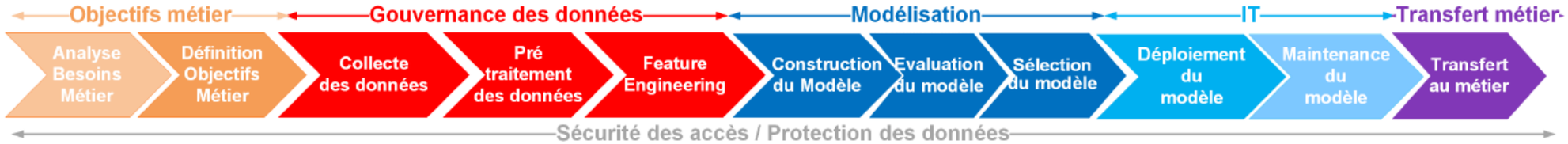
- **Objectif:**
Réunions bimensuelles pour **partage d'expériences et de bonnes** pratiques sur les risques liés à l'utilisation des systèmes d'IA.
Produire une synthèse du travail commun, utile pour d'autres secteurs.

- **Approche:**
Identifier les **risques IA**, leurs impacts et des mesures de réduction des risques.
Chaque Banque a travaillé sur l'identification des risques/contrôles.
Le Groupe a consolidé les résultats pour aboutir à **35 risques identifiés** (non exhaustif).
Le Groupe a identifié **le Top 10 des risques** et les contrôles associés.

- **Réalisation**
Publication d'un livre blanc en octobre 2022 sur le site du Hub France IA.
https://www.hub-franceia.fr/wp-content/uploads/2022/10/22_10_19_Controle-des-risques-des-systemes-IA_PDF.pdf

Audit du cycle de vie des systèmes d'IA

- La construction d'un modèle comprend différentes étapes.



- Le développement, l'implémentation et le monitoring d'un système d'IA impliquent différents acteurs, dont certains sont spécialistes d'IA et d'autres non.

- Trois familles d'acteurs sont impliqués dans la production de systèmes d'IA:
  - Les **profils Métier** (Business Sponsor, Business analyst, Data manager, AI project manager…)
  - Les **spécialistes en Data Science** (Data analyst, Data engineer, Data scientist, Model validation…)
  - Les **profils IT** (Solution architect, Data architect, ML engineer…)

Risques induits par les systèmes d'IA - Top 10

OBJECTIFS METIER
- Analyse imparfaite des besoins métiers
- Problème de compréhension de l'IA
- Choix non pertinent de la variable cible
- Risque de solution sous-optimale par rapport aux objectifs métiers
- Risque d'inadéquation de l'IA au besoin métier

GOUVERNANCE DES DONNEES
- Problème de disponibilité des données
- Manque de transparence des données externes
- Utilisation de données personnelles ou sensibles sans autorisation
- Anonymisation des données: ré-identification versus compatibilité réglementaire
- Fuite de données
- Données incomplètes ou biaisées
- Mauvaise interprétation des données utiliées en modélisation
- Profusion de variables explicatives
- Feedback loops

TRANFERT AU METIER
- Déficit d'interprétabilité / explicabilité des systèmes d'IA
- Risque du transfert métier
- Mauvaise définition de la gouvernance : rôles et responsabilités
- Absence de KPIs et/ou absence d'un processus de monitoring

MODELISATION
- Calibration biaisée des hyper-paramètres
- Absence d'identification et création ou amplification des biais
- Evaluation du modèle et inadéquation des KPI
- Absence de piste d'audit (documentation, librairies)
- Risques liés au réapprentissage automatique en continu

IMPLEMENTATION IT
- Utilisation d'un cloud public sans précaution
- Failles de cybersécurité, attaques adversarielles
- Absence de maintenance et monitoring du modèle
- Déploiement d'un modèle insuffisamment standardisé, sécurisé et contrôlé

Top 10 des risques (1/5)

- **Risque d'inadéquation de l'IA au besoin métier**
  - Méconnaissance des métiers de l'IA
  - Mauvais cadrage du projet
  - Choix non pertinent des indicateurs de performance

  → Sensibiliser les parties prenantes à l'IA et ses risques

  → Déployer une gouvernance transverse pour valider la pertinence du cas d'usage

  → Auditer la documentation décrivant les besoins, objectifs et dispositifs retenus

  → Garder une expertise IA en interne

  Exemple: mesure de performance en fraude

- **Utilisation de données personnelles/sensibles dans l'apprentissage de l'IA**
  - Collecte et utilisation de données personnelles/sensibles

  → Classifier les données (publiques, personnelles, sensibles etc.)

  → Gérer les droits d'accès à ces données, anonymiser les données personnelles

  → Limiter le volume / la durée de rétention de ces données (GDPR)

  → Contrôler la pertinence de ces données pour le modèle

Top 10 des risques (2/5)

- **Mauvaise compréhension ou interprétation des données utilisées en modélisation**
  - Risque augmenté pour l'IA du fait du volume important de données
  - Complexité et faible interprétabilité des systèmes d'IA

  → Faire appel aux experts (réunions Métier)

  → Construire un dictionnaire des données

  → Faire appel à l'IT (ex. connaissances sources, formats)

  → Regard indépendant sur les variables lors de la revue de modèle

- **Absence d'identification, création ou amplification de biais via le système d'IA**
  - Biais: difficile à définir. Notion d'erreur systématique d'un modèle
  - Risque de reproduction de biais inhérents aux données par le système d'IA

  → Détecter les biais en sortie du système d'IA suivant l'objectif (e.g. *equal opportunity*)

  → Identifier la source du biais (i.e. variables incriminées)

  → Remédier au biais (e.g. pre-processing des données)

  → Sensibiliser les modélisateurs

  Exemple: problème de biais dans les LLM

Top 10 des risques (3/5)

- **Risques liés au réapprentissage automatique en continu**
  - Risque de divergence du modèle
  - Risque de manipulation (cybersécurité)
  - Renforcement du biais de sélection

  → Adapter le dispositif de contrôle (e.g. comparaison des performances à des challengers)

  → Impliquer les experts (e.g. contrôles de sous-échantillons)

  Exemple: modèle de détection fraude

- **Déploiement d'un modèle insuffisamment standardisé, sécurisé et contrôlé**
  - Risque d'incompatibilités techniques (e.g. interactions entre modèles, versions de librairies)
  - Risque cybersécurité (e.g. API exposée)
  - Risque de ressources IT (e.g. sur-consommation)

  → Suivre un processus de validation et de contrôles relatifs au déploiement d'applications

  → Sensibiliser les équipes IA et IT

  → Déployer un processus MLOps

  Exemple: déploiement d'une application de demande de crédit en ligne

Top 10 des risques (4/5)

- **Déficit d'interprétabilité / explicabilité des systèmes d'IA**
  - Le modèle d'IA comme boîte noire

  → Evaluer le degré d'intreprétabilité requis (Métier et Data scientists)

  → Utiliser des techniques d'explicabilité

  → Sensibiliser les Data scientists à la qualité des explications (e.g. cohérence)

  → Favoriser les principes d'intelligibilité(e.g. fidélité, sobriété) et d'interactivité (e.g. adaptation aux objectifs du destinataire)

  → Favoriser les interactions Data scientists et Métier

  Exemple: comparer le rang des variables explicatives selon différentes méthodes d'explicabilité pour un modèle d'allocation d'actifs, afin de s'assurer de la fiabilité de l'explication

- **Risque du transfert au métier**
  - Mauvaise intégration/utilisation du système d'IA par le Métier

  → Contrôler le plan de conduite du changement

  → Vérifier l'efficacité de la formation utilisateur

  → Contrôler les usages

  → Mesurer les écarts entre objectifs et réalisations

  Exemple: vérifier l'importance et les raisons des corrections (*overrides*) des utilisateurs

Top 10 des risques (5/5)

- **Mauvaise définition de la gouvernance : rôles et responsabilités**
  - Mauvaise définition ou compréhension des rôles et responsabilités lors de la construction ou du déploiement du modèle

  → Définir les rôles et responsabilités clefs (Model Risk Management)

  → Elaborer les politiques de changement de modèle

  → Impliquer le management

  Exemple: quelles sont les règles d'approbation des recalibrations de modèle ?

- **Absence de KPIs et/ou absence d'un processus de monitoring**
  - Déviation des performances du modèle risque de passer inaperçue

  → Définir une gouvernance de monitoring des modèles

  → Définir des métriques et seuils de rupture associés

  → Utiliser les prédictions de benchmarks

  → Suivre le taux d'*overrides*

  Exemple: définir les métriques et seuils pertinents pour détecter les *shifts* de distribution

HUB
FRANCE
IA

MERCI

https://www.hub-franceia.fr/