HUB FRANCE IA

Opérationnaliser la gestion des risques des systèmes d'intelligence artificielle

GT Banque et auditabilité

Juillet 2024

BNP PARIBAS

SOCIETE GENERALE

LA BANQUE POSTALE

TABLE DES MATIERES

HUB FRANCE IA

1. Introduction

1. Introduction

La multiplication des usages de l'Intelligence Artificielle (IA) s'accompagne de l'émergence de nouveaux risques. La plupart des pays se sont donc attachés à encadrer l'usage de l'IA de façon à en réduire les risques tout en en retirant les bénéfices. En particulier, l'Union Européenne a travaillé depuis avril 2021 à produire une réglementation sur l'IA (RIA) ou *AI Act*[1]. Le processus législatif a amené de très nombreuses modifications aboutissant au texte proposé par le Conseil de l'Union Européenne[2] et qui a finalement abouti au vote du RIA[3]. Une fois le règlement publié au journal officiel (attendu en juillet 2024), les entreprises auront ensuite 2 ans pour se mettre en conformité.

Le travail de mise en conformité nécessite de considérer l'ensemble du processus de développement d'un système d'IA, depuis la phase d'idéation, de conception, de développement, de mise en production jusqu'au monitoring. Les coûts de mise en conformité sont significatifs : dans une enquête[4] réalisée par le Hub France IA et ses partenaires européens en décembre 2022, plus de 50% des entreprises interrogées avaient indiqué qu'elles estimaient les coûts entre 160k€ - 330k€.

Le Groupe de travail Banque et Auditabilité du Hub France IA, qui regroupe des experts IA et audit de trois grandes banques françaises, BNP Paribas, La Banque Postale et Société Générale, travaille depuis plusieurs années sur la gestion des risques liés à l'IA. Le groupe de travail a souhaité partager ses réflexions et son retour d'expérience, en matière de bonnes pratiques, dans un livre blanc[5] qui présentait le processus IA et en analysait les risques.

L'arrivée prochaine du RIA et la montée en puissance de l'IA générative nous a amenés à poursuivre ce travail en tentant de répondre à la question suivante : **Comment outiller la gestion des risques de façon à diminuer les coûts de la mise**

---

[1] European Commission. Laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts. April 21, 2021. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=75788

[2] Council of the European Union. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Analysis of the final compromise text with a view to agreement. January 26, 2024. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

[3] https://www.europarl.europa.eu/doceo/document/A-9-2023-0188-AM-808-808_FR.pdf

[4] appliedAI, Hub France IA et al. AI Act Impact Survey. December, 12 2022 https://www.hub-franceia.fr/wp-content/uploads/2022/12/AIAct-Impact-Survey_Report_Dec12.2022.pdf

[5] Hub France IA. Contrôle des risques des systèmes d'Intelligence Artificielle. Livre blanc. 19 octobre 2022. https://www.hub-franceia.fr/wp-content/uploads/2022/10/22_10_19_Controle-des-risques-des-systemes-IA_PDF.pdf

en conformité au RIA ? Le processus IA pourrait être accompagné tout au long par une suite d'outils (documents, fichiers Excel, outils logiciels, etc.) qui permettraient de valider la conformité au fur et à mesure du processus et pourraient ensuite servir de preuve pour le superviseur (et dans le cas spécifique des institutions financières, de preuve pour les équipes de contrôle interne en charge de s'assurer de la conformité avec le RIA).

La démarche adoptée ici a consisté à décrire le processus IA dans son ensemble en nous appuyant sur le livre blanc[5] de 2022. A chaque étape, nous avons cherché à identifier un outil permettant de contribuer à la conformité. Le RIA venant juste de sortir, il devra être analysé en profondeur avant que nous puissions mettre en correspondance les outils proposés ici avec les attentes de la mise en conformité au titre du RIA. Ceci fera l'objet d'un prochain guide à paraître.

Le travail présenté ici ne vise pas à être exhaustif, ce qui aurait nécessité un document beaucoup plus long, mais souhaite apporter un **cadre méthodologique et des bonnes pratiques**. Ce travail, nous l'espérons, servira à d'autres secteurs économiques qui, s'en inspirant, pourront mettre en œuvre les processus adaptés à leur contexte pour mieux maîtriser les risques des solutions IA qu'ils déploient ou utilisent et préparer leur mise en conformité.

HUB FRANCE IA

2. Le processus IA

2. Le processus IA

2.1. Définition de l'IA

L'Intelligence Artificielle est un « ensemble de théories et de techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine »[6]. Elle comprend deux grandes familles : l'IA symbolique et l'IA numérique, qui ont chacune connu des périodes de succès et des « hivers », ainsi que le montre le schéma ci-dessous représentant l'activité de ces deux familles depuis les années 50.

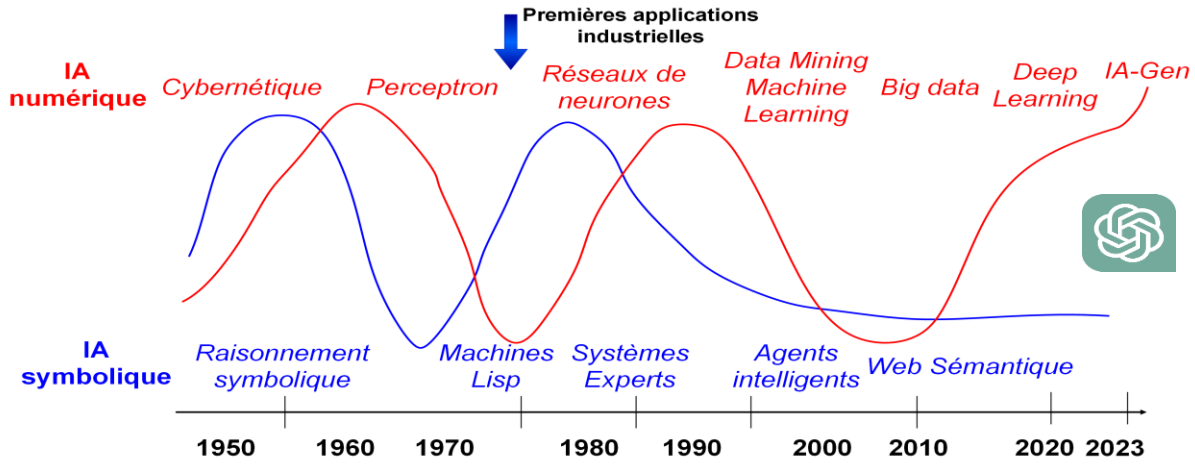

Figure 1 – IA symbolique et IA numérique

En 2023, la mise sur le marché de ChatGPT a mis en lumière l'IA générative que nous décrivons plus loin. L'IA symbolique est peu utilisée dans le monde bancaire, nous nous concentrerons donc sur l'IA numérique : les techniques les plus répandues aujourd'hui, et notamment dans les applications bancaires, sont les techniques du *Machine Learning* (ou apprentissage automatique), dont font partie le *Deep Learning* et les techniques d'*IA générative*. Dans toute la suite, quand nous parlerons d'IA, il s'agira donc de *Machine Learning*, sauf mention contraire.

[6] https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

2.2. L'IA dans la banque

L'intelligence artificielle accompagne de plus en plus de processus bancaires et la tendance s'est fortement accélérée ces dernières années. Nous renvoyons le lecteur pour plus de détails à notre livre blanc[5].

Une part importante des cas d'usage de l'IA au sein des Banques est destinée à **automatiser des processus** internes afin d'améliorer l'efficacité tout en réduisant le risque opérationnel, d'**améliorer l'expérience client**, avec l'apparition d'agents conversationnels qui assistent les clients dans leurs opérations, à la connaissance des clients (marketing) ou à la conception de nouveaux services pour les clients. Par ailleurs, en améliorant la précision des algorithmes, le recours à l'IA contribue à **l'atténuation de nombreux risques, notamment le risque opérationnel, le risque de crédit et la conformité.**

Les institutions financières sont très aguerries au développement de modèles, notamment au contrôle de la discipline de modélisation, du fait de la réglementation du domaine déjà en place. Cependant, il est important d'y prêter attention lorsque les modèles sont développés en externe ou dans des équipes traditionnellement éloignées de la modélisation.

2.3. Machine Learning

La production d'une solution à base de *Machine Learning* se fait en deux étapes :

- La **conception** (ou *Build* en anglais) : à partir d'un besoin exprimé, le *data scientist* va collecter les données adaptées pour constituer un *dataset d'apprentissage*, puis tester divers algorithmes d'apprentissage (très souvent dans une librairie open-source). A l'issue du processus d'apprentissage, on obtient un modèle IA - un programme qui peut ensuite être utilisé : ce programme peut être codé dans n'importe quel langage informatique, le plus courant étant aujourd'hui python.
- L'exploitation (**inférence** ou *Run* en anglais) : à l'issue de l'étape d'apprentissage, le *data scientist* présente de nouvelles données au modèle obtenu et obtient en sortie le résultat le plus probable pour les données entrées.

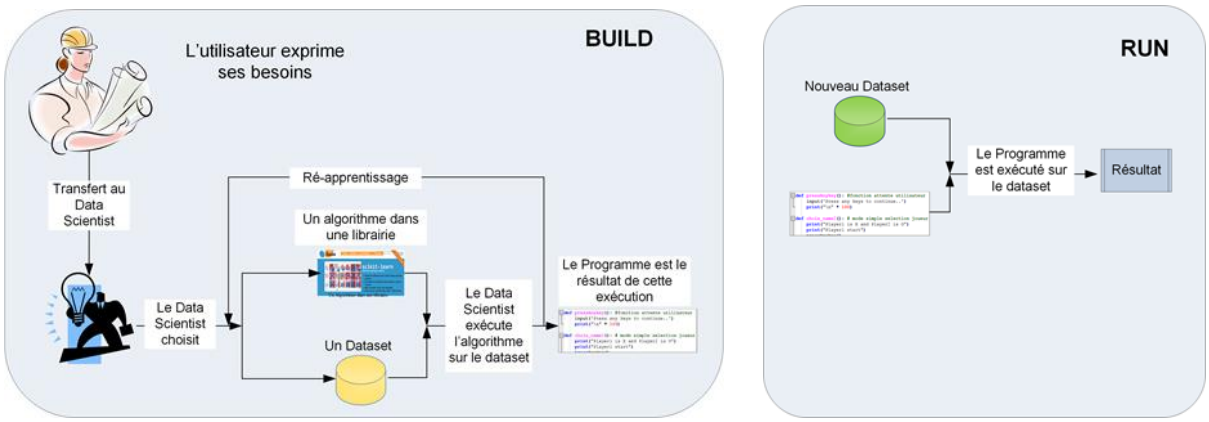

Figure 2 – Les deux étapes de production et utilisation d'un modèle de *Machine Learning*

Il faut noter qu'on utilise en pratique trois familles d'indicateurs de performance : les indicateurs techniques sont ceux utilisés en apprentissage pour optimiser le modèle IA, les indicateurs métier sont ceux qui mesurent la valeur métier générée par l'utilisation du modèles, enfin des indicateurs plus opérationnels sont aussi utilisés, comme la durée du calcul, le temps de latence, le nombre de variables et la complexité du modèle, voire même le coût des variables si certaines sont achetées.

2.4. IA Générative

L'IA Générative est un sous-ensemble du Machine Learning, et plus particulièrement du Deep Learning, visant à produire du contenu, que ce soit du texte, une image, de l'audio ou une vidéo, à partir de données en entrée (on parle alors de *prompt*), elles-mêmes du texte, une image, de l'audio ou une vidéo par exemple. Un modèle d'IA Générative crée un nouveau contenu statistiquement cohérent avec les données d'entraînement et la requête formulée

Les modèles d'IA Générative sont généralement entraînés sur un large ensemble de données, nécessitant des moyens conséquents pour leur apprentissage. L'architecture du modèle (*Transformer*[7]) associé au volume très important de données utilisées pendant l'entraînement permet des usages variés pour ces modèles, sans que ces derniers n'aient été entraînés spécifiquement pour ces tâches.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. Attention is all you need. *Advances in neural information processing systems*. vol. 30, 2017. https://arxiv.org/pdf/1706.03762.pdf

Les premières solutions fondées sur de grands modèles de langage, tels que ChatGPT[8], sont capables de créer un texte à partir d'instructions textuelles en entrée. Ces modèles deviennent aujourd'hui de plus en plus multimodaux, c'est-à-dire qu'ils peuvent prendre en compte, aussi bien en entrée qu'en sortie, des données de plusieurs types, même combinées telles que l'image, l'audio, la vidéo, de la 3D, etc.

Les capacités génératives des modèles d'IA Générative et leur accès continuent de se développer à grande vitesse aujourd'hui, ouvrant régulièrement la voie à de nouveaux usages.

L'usage de l'IA Générative dans le secteur bancaire se développe dans le même cadre que l'usage de l'Intelligence Artificielle, en majorité le Machine Learning, c'est-à-dire en cohérence avec les réglementations existantes en termes de contrôle interne et de gestion du risque.

Le secteur bancaire, comme beaucoup d'autres secteurs d'activité, utilise et manipule beaucoup de documents (texte, tableau, image). L'IA Générative permet d'exploiter plus facilement des grands volumes de données, principalement textuelles, à la fois pour les employés (recherche d'information au sein d'un corpus documentaire) et les clients (réponses plus pertinentes et adaptées au contexte améliorant ainsi l'expérience client).

La multimodalité des modèles permet également de traiter des données non structurées, telles que la voix ou les images, de manière plus industrielle, automatisant par la même occasion certaines tâches très consommatrices de temps (par exemple, les contrôles de conformité relatifs à la protection de la clientèle).

L'usage de l'IA Générative dans le secteur bancaire progresse prudemment, car, malgré les possibilités offertes par ces nouveaux modèles, ces derniers ne sont pas exempts de défauts, et il est nécessaire de bien évaluer leur performance par rapport aux risques encourus avant tout déploiement généralisé de tels modèles.

Les grandes familles d'usage[9] comprennent donc : les agents conversationnels, les systèmes de recherche augmentés (RAG : *Retrieval Augmented Generation*), les systèmes pour résumer, les systèmes d'assistance ou de création de contenu, les

[8] OpenAI. Introducing ChatGPT. Open AI. November 30, 2022. https://openai.com/blog/chatgpt
[9] Hub France IA. Les usages de l'IA Générative. Volume 1 - Les LLM. Janvier 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf

systèmes pour le code informatique et les systèmes pour raisonner sur des données structurées ou non structurées.

Nous explicitons ci-dessous quelques termes utilisés dans la littérature.

Notons tout d'abord qu'un système d'IA Générative peut être produit de plusieurs façons comme le décrit la figure ci-dessous : soit en utilisant directement un système IA commercial ou *open-source*, soit en raffinant (*fine tuning*) un tel système, soit en utilisant un mécanisme d'augmentation (*RAG*) exploitant une base documentaire spécifique.

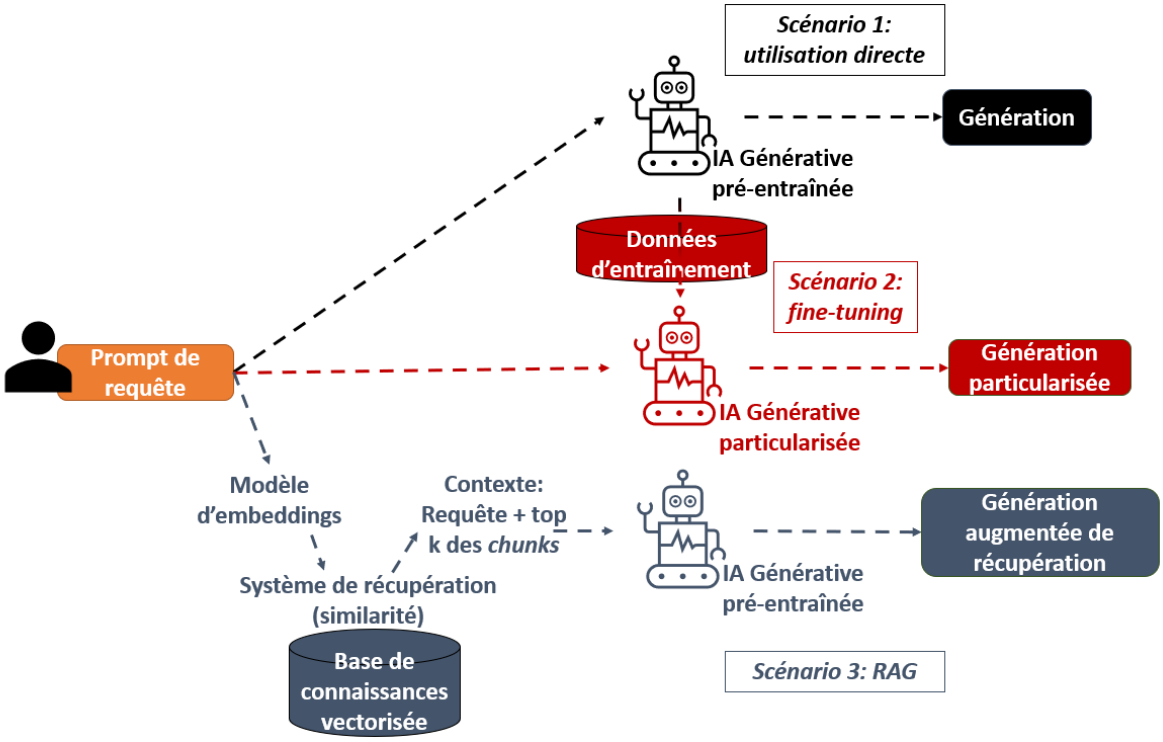

Figure 3 : Scénarios d'applications de l'IA Générative

Prompts : Le *prompt* est l'instruction ou la requête en langage naturel fournie à l'IA Générative dans le but d'obtenir une réponse. En général, le *prompt* système sert à donner une instruction ou un contexte à l'IA Générative (ex. « vous devez répondre en français »), alors que le *prompt* utilisateur permet de spécifier la question ou requête (ex. « résumez ce texte »). Suivant l'interface utilisée (ex. *Application Programming Interface - API* versus simple *chatbot*), il sera possible de préciser pour ce *prompt* certains paramètres tels que le nombre maximum de tokens en sortie ou la température (qui mesure le degré d'aléatoire dans le texte généré). Dans ce contexte, le *prompt engineering* est une stratégie qui vise à concevoir les *prompts* les plus efficaces pour obtenir la sortie désirée. Cette stratégie doit s'adapter à chaque étape du cycle de vie de l'IA Générative : *finetuning*, évaluation,

ou usage de l'IA Générative par le métier. Par exemple, lors de l'évaluation d'une IA Générative, il peut être utile de créer des modèles de *prompts* de test pour en générer un volume important à l'aide d'un LLM (*Large Language Model*). Il convient également d'être précis sur les instructions concernant la sortie attendue concernant la tâche (ex. « Répondez par 'oui' ou 'non' ») afin d'automatiser l'évaluation du LLM testé.

IA Générative pré-entraînée : une IA Générative est dite « pré-entraînée » car elle a été construite par apprentissage sur de très larges volumes de données, lui permettant ainsi d'acquérir des connaissances générales et les bases pour générer du contenu pertinent. Les grands modèles de langage LLM comme ChatGPT, Llama, etc. sont des modèles pré-entraînés. L'intégration dans le *prompt* de quelques exemples de la tâche à effectuer est appelée *few-shot learning*. Cette technique vise à augmenter la performance du modèle pré-entraîné sur une tâche spécifique (ex. classification, génération de code). Plus précisément, le *prompt* inclura un ou plusieurs exemples de requêtes utilisateur et de réponses de l'assistant avant la requête finale destinée à l'IA Générative. Cette technique est différente du *fine-tuning* car les paramètres (poids) du modèle ne sont pas mis à jour. Il s'agit juste de spécifier le contexte par cette technique de *prompt engineering*. Le *zero-shot learning* consiste au contraire à alimenter l'IA Générative avec un *prompt* « sec », sans exemple de la tâche à effectuer.

Embedding : Un *embedding* est une représentation vectorielle de grande dimension pour un objet tel qu'un mot, un document ou une image. Ces représentations sont apprises lors de l'entraînement du modèle. Elles contribuent au calcul de l'attention dans les modèles Transformers (à la base des LLM) et permettent par exemple de savoir quels mots sont proches sémantiquement. Les modèles d'*embeddings* pré-entraînés permettent ainsi de vectoriser des objets de manière pertinente afin de réaliser notamment des tâches de recherche (e.g. RAG). Dans le cas du RAG, par exemple, la requête texte de l'utilisateur est transformée en *embeddings* par un tel modèle. Ce vecteur est ensuite comparé à d'autres *embeddings* stockés dans une base de données vectorisées. Des mesures de similarité permettent d'extraire et d'ordonner le top k (k = 1,2,3,5, 10 ou autre, paramètre choisi par les personnes construisant la solution spécifique) des passages les plus pertinents de la base. Avec la concaténation entre les passages et la requête, l'IA Générative peut ainsi générer la réponse la plus pertinente. On voit ainsi le rôle essentiel des *embeddings* dans le cas du RAG.

RAG (*Retrieval Augmented Generation*). La connaissance d'une IA Générative ne dépasse pas le cadre des données d'entraînement. Afin que les réponses générées par l'IA se basent sur des données plus fraîches et/ou plus spécifiques, il est nécessaire de lui donner accès à une **base de connaissances**. La construction de cette base se fait en plusieurs étapes. Tout d'abord, il faut se procurer les données textes pour alimenter cette base (ex. fichiers PDF). Ensuite, il convient de diviser les données en morceaux de texte (*chunks*) avec une stratégie donnée (ex. nombre fixe de *tokens*) et pouvant se chevaucher. La troisième étape consiste à vectoriser ces morceaux de texte à l'aide d'un modèle d'*embeddings* pré-entraîné. Cette représentation vectorielle aide la recherche car les morceaux de textes similaires ont des représentations proches. L'étape suivante est la création de la base de données vectorielle avec un index auquel sont ajoutés les vecteurs d'*embeddings* précédents. Cette structure de données permet de retrouver rapidement l'information recherchée. Comme décrit dans le paragraphe sur les *embeddings*, il est alors possible de coupler ce système de récupération (*retrieval system*) avec une IA Générative pour réaliser de la génération augmentée de récupération. L'utilisateur adresse une requête de recherche, laquelle est vectorisée avec un modèle d'*embeddings*. Le système de récupération sélectionne le top k des *chunks* à l'aide de méthodes de recherche sémantique basées sur des vecteurs denses ou épars. Enfin, l'IA répond à l'utilisateur en se basant sur le contexte et en utilisant ses capacités de génération. Le contexte contient ainsi la requête et le top k des *chunks* les plus pertinents sélectionnés dans la base de données vectorielle. Le RAG permet donc d'obtenir des réponses plus précises sur le domaine défini par la base de connaissances.

Fine-tuning : Le *fine-tuning* d'une IA Générative pré-entraînée consiste à continuer à l'entraîner sur des données labellisées spécifiques afin d'améliorer sa performance sur une tâche ou un domaine particulier. Contrairement au « *few-shot learning* » ou « *zero-shot learning* », les paramètres (poids) du modèle pré-entraîné sont ainsi mis à jour. Par exemple, l'IA Générative pourra ainsi être particularisée sur une tâche de classification liée à un domaine particulier. Autre exemple : un chatbot dans la Banque pourra être affiné en réentraînant sur des séries de questions / réponses labellisées comme « bonne réponse » ou « mauvaise réponse » par des conseillers clientèle. Le *fine-tuning* requiert ainsi un certain investissement dont il convient d'évaluer la réelle plus-value par rapport à une stratégie de prompt expérimentée. La première étape du *fine-tuning* est de préparer les données d'entraînement et de validation. Il s'agira par exemple d'une série de prompts,

chacun contenant des rôles (système, utilisateur, assistant) et des contenus (instruction, question, réponse). L'IA Générative est ensuite entraînée puis évaluée sur chaque partition de données respective. Si la performance obtenue n'est pas satisfaisante, il est alors souhaitable de continuer l'entraînement avec des prompts ciblés, pour lesquels le modèle n'offre pas les résultats escomptés. Enfin, l'IA Générative peut être utilisée sur le cas d'usage ciblé. Le *fine-tuning* est utilisé pour réduire les hallucinations sur le domaine particulier envisagé.

LangChain : *LangChain*[10] est un outil *open-source* permettant de rationnaliser le développement d'applications autour des IA Génératives. LLM Chain permet notamment de créer des pipelines impliquant une ou plusieurs IA Génératives avec une séquence de tâches comme la génération de texte puis la traduction. De même, il est possible de créer une chaîne de récupération (*retrieval chain)* pour intégrer du *RAG* dans la séquence.

Guardrails : Ce sont des protections qui permettent de contrôler les entrées et sorties d'une IA Générative afin de réduire les risques liés à son utilisation. Elles sont particulièrement importantes dans le cadre d'une application utilisée en externe par des clients (ex. *chatbot*). Les risques peuvent être liés à différents aspects, comme la toxicité, les biais, le risque cyber, la fuite de données, ou les hallucinations. Les IA Génératives ont généralement assimilé une politique d'éthique bien définie. Cependant, il est parfois possible de contourner celle-ci par la création de prompts adverses. Pour réduire ces risques, il est utile de donner des instructions de comportement éthique à l'IA Générative via le *prompt* système ou d'utiliser un outil de contrôle tel que *LangChain Constitutional AI* [11]. Dans ce dernier cas, un compromis entre latence et niveau de risque devra être fixé.

[10] https://www.langchain.com/
[11] https://python.langchain.com/docs/guides/productionization/safety/constitutional_chain/

2.5. Le processus de production du modèle

Le processus de production d'un modèle d'IA prédictive, décrit dans la figure ci-dessous, comprend différentes étapes, avec potentiellement des retours en arrière pour itération tant qu'on n'est pas satisfait du résultat.

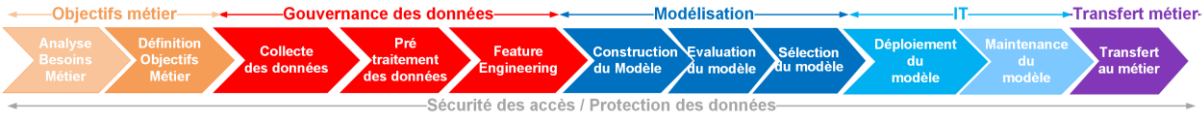

Figure 4 - Processus de production d'un modèle d'IA prédictive (en représentation linéaire)

Le processus de production d'un modèle d'IA générative est décrit ci-dessous. Cependant, on notera que le processus sera un peu différent selon que le modèle est pris directement « *out of the box* » (Copilot 365[12] par exemple), fine-tuné ou en exploitant une méthode RAG. De plus le domaine évoluant très rapidement, le processus même est susceptible d'évoluer :

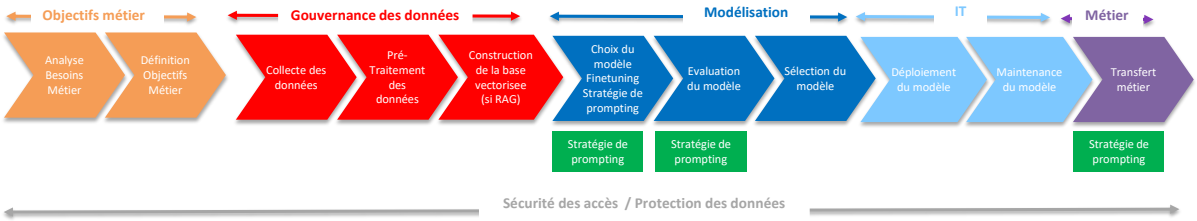

Figure 5 - Processus de production d'un modèle d'IA Générative

Nous soulignons spécifiquement dans ce schéma les étapes dans lesquelles la stratégie de *prompting* intervient, qu'elle serve à modéliser ou qu'elle soit évaluée. L'évaluation des risques d'une solution IA se fera tout au long du processus, on doit s'assurer que les outils appropriés et les bons intervenants sont positionnés là où c'est nécessaire, comme nous l'avons décrit dans notre livre blanc[5] sur le contrôle des risques.

[12] Colette Stallbaumer. Introducing Microsoft 365 Copilot—A whole new way to work. Microsoft. March 16, 2023. https://www.microsoft.com/en-us/microsoft-365/blog/2023/03/16/introducing-microsoft-365-copilot-a-whole-new-way-to-work/

HUB FRANCE IA

3. Certification AI Act

Opérationnaliser la gestion des risques des systèmes d'Intelligence Artificielle

Juillet 2024

3. Certification AI Act

La règlementation des systèmes d'intelligence artificielle (SIA) telle que mise en place par le Règlement sur l'Intelligence Artificielle[13] (RIA ou *AI Act*) est construite sur une approche par niveau de risque. Le risque étant défini comme « la combinaison entre la probabilité d'un risque et la sévérité de celui-ci »[14]. Le règlement distingue alors 4 niveaux de risques :

- Risques inacceptables
- Risques élevés
- Autres (avec devoir de transparence)
- Autres (sans devoir de transparence)

Cette première classification s'applique aux SIA rentrant dans le champ d'application du texte (i.e. répondant notamment à la définition de système d'IA du RIA). En particulier, dans le cadre de ce document, on se concentre sur les systèmes d'IA prédictive qui constituent une sous-partie des systèmes d'IA qui rentrent dans le champ d'application du RIA.

A celle-ci s'est ajoutée plus tardivement et par une approche similaire, la réglementation des modèles d'IA dits « **à usage général** ». Ces modèles sont divisés en deux niveaux de risque au sein du RIA : les modèles présentant un **risque systémique** et les autres.

En fonction du niveau de risque correspondant au système d'IA prédictif ou au modèle d'IA à usage général, **les exigences et obligations applicables diffèrent**. Celles-ci évoluent également suivant la place de l'opérateur dans la **chaîne de valeur de l'IA**. Le RIA distingue ainsi les fournisseurs, déployeurs, mandataires, importateurs et distributeurs d'un SIA[15].

Les deux rôles principaux pouvant être exercés lors de la mise en place d'un projet d'intelligence artificielle au sein d'une entreprise sont ceux de fournisseur et de déployeur.

---

[13] Résolution législative du Parlement européen sur la proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle et modifiant certains actes de l'Union, 16 avril 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_FR.pdf. La publication du texte au journal officiel est prévue le 12 juillet 2024.

[14] *Ibid,* article 3§2

[15] Il est important de noter qu'un opérateur peut cumuler ou alterner entre ces statuts au cours du temps et de ses activités.

Fournisseur : « une personne physique ou morale, une autorité publique, une agence ou tout autre organisme qui **développe ou fait développer un système d'IA ou un modèle d'IA à usage général et le met sur le marché** ou **met le système d'IA en service sous son propre nom ou sa propre marque**, à titre onéreux ou gratuit »[16]

Déployeur : « une personne physique ou morale, une autorité publique, une agence ou un autre organisme **utilisant sous sa propre autorité un système d'IA** sauf lorsque ce système est utilisé dans le cadre d'une activité personnelle à caractère non professionnel »[17]

Dans le cadre de ce livrable, nous concentrerons donc nos développements sur les obligations pesant sur ces deux opérateurs notamment lors de la conception et/ou du déploiement d'un système d'IA prédictive à haut risque ou d'une IA Générative.

3.1. Les exigences et obligations applicables pour les systèmes d'IA prédictive

3.1.1. Classification des modèles d'IA prédictive

Comme évoqué plus haut, les modèles d'IA prédictive sont divisés au sein de 4 catégories au sein du RIA (ce tableau ne constitue pas un résumé exhaustif de toutes les nuances existant dans le RIA) :

Catégorie de risque	Caractéristiques du SIA
Inacceptable Article 5	Utilise des techniques subliminales échappant à la conscience d'une personne **Ou** Exploite les vulnérabilités d'une personne **Ou** Utilise des systèmes de catégorisation biométrique afin d'inférer/déduire des données sensibles sur l'individu **Ou** Utilise une technique de *scoring* social

[16] *Ibid,* article 3 § 3.
[17] *Ibid,* article 3 § 4.

	Ou Utilise l'identification biométrique à distance en temps réel à des fins répressives **Ou** Utilise une technique de prédiction des risques d'actes répréhensibles (Note : cette liste ne se veut pas exhaustive et chacun des cas ci-dessus fait l'objet de conditions précises)
Risques élevés Articles 6 à 15 Article 27 Exceptions listées Article 6, paragraphes 3a, 3b, 3c et 3d	Est visé par l'annexe III
	Est couvert par une des législations d'harmonisation de l'Union énumérée dans l'annexe II du RIA **Et** Doit faire l'objet d'une évaluation de conformité par un tiers avant sa mise sur le marché **Et** Doit faire l'objet d'une évaluation de l'impact sur les droits fondamentaux des systèmes d'IA à haut risque
Autres (avec devoir de transparence) **Article 50**	Obligations de transparence pour les fournisseurs et les utilisateurs de certains systèmes d'IA.
Autres (sans devoir de transparence)	Tous les SIA qui ne sont pas concernés par les critères précédents.

Pour la catégorie Risques élevés, des exceptions sont listées (considérant 32a Article 6, paragraphe 2a et 2b) :

Si un SIA

Remplit les critères de catégorisation de SIA à hauts risques mais pas de risque significatif d'atteinte à la santé, à la sécurité ou aux droits fondamentaux des personnes physiques, y compris en n'influençant pas sensiblement le résultat de la prise de décision.

Sont concernés les SIA :

Est destiné à exécuter une tâche procédurale limitée[19]

Et/ou

Est destiné à améliorer le résultat d'une activité humaine

Et/ou

Est destiné à détecter des modèles de prise de décision/des écart par rapport à des modèles de prise de décision antérieurs mais n'est pas destiné à remplacer ou influencer l'évaluation humaine sans contrôle humain approprié

Et/ou

Est destiné à une tâche préparatoire avant évaluation humaine

Les SIA effectuant un profilage de personnes physiques sont toujours considérés comme Risques élevés.

3.1.2. Les exigences et obligations applicables pour les systèmes d'IA à haut risque

La classification d'un système d'IA comme « à haut risque » peut ainsi résulter de deux scénarios :

- Le SIA est utilisé dans un des domaines listés à l'annexe III en raison du **préjudice qu'il pourrait causer et de sa sévérité sur la santé, la sécurité ou les droits fondamentaux de personnes physiques.** Ces domaines sont, par exemple, la gestion de certaines infrastructures critiques, les systèmes utilisés pour évaluer la solvabilité des personnes physiques ou établir une note de crédit, etc.
- Le SIA est utilisé comme composant de sécurité d'un produit ou est lui-même un tel produit, couvert par les actes législatifs d'harmonisation de l'Union (jouets, véhicules maritimes…voir annexe I du RIA).

Ces systèmes d'IA à haut risque seront soumis à des exigences et des obligations afin de pouvoir être mis sur le marché européen. Le RIA distingue, dans un premier temps, les exigences c'est-à-dire les caractéristiques et prérequis applicables au SIA lui-même, des obligations.

Le tableau ci-après résume, de manière non-exhaustive, les différentes exigences applicables aux systèmes d'IA à haut risque.

Catégorie d'exigence	Détails des exigences
Système de gestion des risques Article 9	Etablissement et mise en œuvre d'un système de gestion des risques, documenté et tenu à jour tout au long du cycle de vie du SIA.
Données et gouvernances des données Article 10	Les jeux de données d'entrainement, de validation et de tests doivent satisfaire à des critères de qualité. De plus, ils doivent être soumis à des exigences en matière de gouvernance et de gestion des données.
Documentation technique du modèle Article 11 Article 72 Annexe IV	La **documentation technique est établie de manière à montrer que le SIA à haut risque satisfait aux exigences**. Elle est rédigée de manière **claire et intelligible** afin de servir à l'évaluation de conformité du SIA. Elle comprend a minima : • Description générale du système d'IA, • Description détaillée des éléments du SIA et de son processus de développement • Informations détaillées sur la surveillance, le fonctionnement et le contrôle du SIA • Informations concernant l'adéquation des indicateurs de performances au système d'IA concerné • Description du système de gestion des risques mis en place • Liste des normes harmonisées appliquées en totalité ou partie • Copie de la déclaration UE de conformité • Description détaillée du système en place pour évaluer les performances et surveiller le SIA après la commercialisation (art. 72).
Journaux d'évènements Article 12	Enregistrements automatiques des évènements tout au long de la durée de vie du système. Les fonctionnalités de journalisation doivent permettre d'enregistrer les événements pertinents pour **repérer les situations à haut risque, faciliter la surveillance post-commercialisation**, et **contrôler le fonctionnement du système d'IA**.

Transparence et notice d'utilisation Article 13	Le SIA doit être conçu et développé de manière à favoriser un niveau de transparence permettant aux déployeurs d'utiliser le SIA de manière appropriée et d'en interpréter les résultats. Il est accompagné d'une notice d'utilisation à destination des déployeurs contenant : • Identité et coordonnées du fournisseur, • Caractéristiques, capacités et limites de performances du SIA à haut risque, • Les modifications du SIA et de sa performance prédéterminée par le fournisseur au moment de l'évaluation initiale de la conformité, • Les mesures de contrôle humain, • Les ressources informatiques et matérielles nécessaires, la durée de vie attendue et les mesures de maintenance et de suivi nécessaires pour assurer le bon fonctionnement du SIA y compris en ce qui concerne les mises à jour logicielles, • Le cas échéant, une description des mécanismes compris dans le système d'IA à haut risque qui permet aux déployeurs de collecter, stocker et interpréter correctement les journaux.
Contrôle humain Article 14	La conception et le développement des systèmes d'IA à haut risque permettent, notamment au moyen d'interfaces homme-machine appropriées, un contrôle effectif par des personnes physiques pendant leur période d'utilisation. Ce contrôle humain vise à **prévenir ou à réduire au minimum les risques pour la santé, la sécurité ou les droits fondamentaux qui peuvent apparaître lorsqu'un système d'IA à haut risque.**
Exactitude, robustesse et cybersécurité Article 15	La conception et le développement des SIA à haut risque permettent d'atteindre un niveau approprié d'exactitude, de robustesse et de cybersécurité, et de fonctionner de façon constante à cet égard tout au long de leur cycle de vie.

Tableau 1 - Exigences applicables aux systèmes d'IA à haut risque

Ces exigences doivent être mises en place avant (*ex-ante*) la mise sur le marché/en service du SIA à haut risque et subsistent tout au long du cycle de vie du SIA.

Le Règlement sur l'intelligence artificielle détaille ensuite **les obligations qui s'imposent aux fournisseurs et aux déployeurs de ces systèmes.** Il est important de noter que si le déployeur commercialise en son nom propre ou marque un système d'IA à haut risque déjà mis sur le marché ou mis en service, ou y apporte une modification substantielle, il est alors considéré comme fournisseur du SIA (article 25).

Le tableau ci-après résume de manière non-exhaustive les différentes obligations applicables aux fournisseurs de système d'IA à haut risque.

Obligations applicables aux fournisseurs de SIA à haut risque	Détails des obligations
Conformité aux exigences Article 16	Le fournisseur veille à ce que le SIA soit conforme aux exigences précédemment exposées.
Mise en place d'un système de gestion de la qualité Article 17	Le fournisseur met en place un système de gestion de la qualité garantissant la conformité au RIA. Il est documenté sous la forme de politiques, procédures et instructions écrites. Il comprend a minima : • Une stratégie de respect de la réglementation, • Les techniques, procédures et actions systématiques sont employées pour concevoir et contrôler les systèmes d'IA à haut risque ainsi que pour vérifier leur conception, • Les techniques, procédures et actions destinées au développement, contrôle et assurance de la qualité du SIA, • Les procédures d'examen, de test et de validation à exécuter au cours du cycle de vie du SIA à haut risque et leur fréquence,

	• Les spécifications techniques, normes et moyens à utiliser pour que le SIA à haut risque soit conforme aux exigences du RIA, • Le système de gestion des risques mis en place, • Le détail de la conception, de l'implémentation et du fonctionnement du système de surveillance après commercialisation, • Les procédures relatives à la notification d'un incident grave (article 73), • La gestion des communications avec les autorités compétentes, • Les systèmes et procédures de conservation des documents et information pertinents, • La gestion des ressources, • Le cadre de responsabilisation.
Conservation de la documentation Article 18	La documentation issue des articles 11 et 17 du RIA, ainsi que celle concernant les approbations de modifications et décisions émises par les organismes notifiés et la déclaration de conformité au RIA doivent être conservés pendant 10 ans après la mise sur le marché/en service du SIA à haut risque.
Tenue des journaux Article 19	Le fournisseur assure la tenue des journaux générés automatiquement visés à l'article 12. Ces journaux sont conservés pendant une période adaptée à la destination du SIA à haut risque, d'au moins 6 mois.
Mesures correctives nécessaires et devoir d'informations Article 20	Lorsqu'un fournisseur de SIA à haut risque a des raisons de considérer qu'un de ses SIA mis sur le marché/en service n'est pas conforme, il prend immédiatement les mesures correctives nécessaires et en informe les déployeurs et autres acteurs concernés.
Coopération avec les autorités compétentes Article 21	A la demande motivée d'une autorité compétente, le fournisseur met à la disposition de ladite autorité toutes les informations et documents nécessaires pour démontrer la conformité du SIA à haut risque.

Tableau 2 – Exigences applicables aux fournisseurs de système d'IA à haut risque

Enfin, l'article 16 du RIA précise **la procédure à suivre préalablement à la mise sur le marché/en service d'un système d'IA à haut risque**. Ainsi, afin de pouvoir être mis sur le marché, ces systèmes devront faire l'objet d'une **procédure d'évaluation de conformité** (article 43), **d'une déclaration de conformité** (article 16), se voir **apposer le marquage CE** sur leur emballage ou documentation (article 16) et **être enregistrés dans la base de données de l'UE.**

En ce qui concerne les déployeurs de systèmes d'IA, celles-ci sont principalement organisationnelles (article 26). Ainsi, le déployeur d'un système d'IA à haut risque doit, de manière non-exhaustive :

- Prendre les **mesures techniques et organisationnelles appropriées afin de garantir qu'il utilise le SIA à haut risque conformément à la notice d'utilisation du système**,
- S'assurer que les personnes en charge du contrôle humain du système disposent des **compétences, de la formation et de l'autorité** nécessaire à leurs missions.
- **Exercer un contrôle sur les données d'entrée** et veiller à ce que ces dernières soient pertinentes et suffisamment représentatives au regard de la destination du SIA à haut risque,
- **Surveiller le SIA à haut risque** sur la base de la notice du système et **informer le fournisseur de tout évènement pouvant présenter un risque,**
- Assurer la tenue des journaux générés par le SIA à haut risque,
- Informer les représentants des travailleurs et les travailleurs qui seront soumis à l'utilisation du système d'IA à haut risque.

3.2. La règlementation applicable pour les systèmes d'IA générative

3.2.1. Régimes applicables aux systèmes d'IA Générative

La notion de « modèle d'IA à usage général » est apparue tardivement dans le processus d'élaboration de AI *Act*.

Modèle d'IA à usage général est « un modèle d'IA, y compris lorsque ce modèle d'IA est entraîné à l'aide d'un grand nombre de données utilisant l'auto-supervision à grande échelle, qui présente une généralité significative et est capable d'exécuter de manière compétente un large éventail de tâches distinctes, indépendamment de la manière dont le modèle est mis sur le marché, et qui peut être intégré dans

une variété de systèmes ou d'applications en aval, à l'exception des modèles d'IA utilisés pour des activités de recherche, de développement ou de prototypage avant leur publication sur le marché. »[18]

Selon cette définition, les IA génératives relèvent des modèles d'IA à usage général. Les IA génératives sont donc réglementées par les dispositions du RIA consacrées aux modèles d'IA à usage général mais également par les dispositions relatives aux systèmes d'IA générant ou manipulant des images, ou contenus audios et vidéos (article 50).

3.2.2. Classification des modèles d'IA à usage général

Les dispositions applicables aux modèles d'IA à usage général (articles 51 et suivants) divisent ces modèles en deux catégories : les modèles d'IA à usage général présentant un risque systémique et les autres.

La notion de **risque systémique** est spécifique à ces modèles d'IA et s'entend comme un risque « ayant une incidence significative sur le marché de l'Union en raison de leur portée ou d'effets négatifs réels ou raisonnablement prévisibles sur la santé publique, la sûreté, la sécurité publique, les droits fondamentaux ou la société dans son ensemble, pouvant être propagé à grande échelle tout au long de la chaîne de valeur »[19].

Un modèle d'IA à usage général est classé comme présentant un risque systémique lorsqu'il dispose « de **capacités à fort impact** »[20]. Ceci est présumé dès lors que le modèle remplit l'une des conditions suivantes :

- « La quantité cumulée de calcul utilisée pour son entraînement, mesurée en nombre d'opérations flottantes est supérieure à 10^{25} » [21].
- Il fait l'objet d'une décision de la Commission, d'office ou à la suite d'une alerte qualifiée du groupe scientifique sur la base des critères identifiés à l'annexe XIII dont par exemple le nombre de paramètres du modèle, la taille ou qualité de son jeu de données ou encore le nombre d'utilisateurs finaux inscrits.

[18] *Ibid,* article 3 § 63.
[19] Résolution législative du Parlement européen sur la proposition de règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l'intelligence artificielle et modifiant certains actes de l'Union, 13 mars 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_FR.pdf
[20] *Ibidem*, article 51 §1a.
[21] *Ibid*. article 51 §2.

3.2.3. Obligations incombant aux fournisseurs de modèles d'IA Générative

Pour les IA Génératives, les obligations incombant à leurs fournisseurs sont semblables à celles des fournisseurs de systèmes d'IA à haut risque.

Pour les IA Génératives assimilées à des modèles d'IA à usage général (hors open-source), elles sont composées des éléments suivants[22] :

- Rédiger et tenir à jour la documentation technique du modèle ainsi que les informations à destination des fournisseurs de systèmes d'IA envisageant d'intégrer le modèle dans leur SIA,
- Mettre en place une politique visant à se conformer à la législation européenne en matière de droit d'auteurs,
- Rendre public un résumé détaillé du contenu utilisé pour entraîner le modèle d'IA à usage général,
- Coopérer avec la Commission Européenne et les autorités nationales compétentes.

Pour **les systèmes d'IA générative à risques systémiques**, s'ajoute entre autres l'obligation pour les fournisseurs d'effectuer des évaluations de leurs modèles pour identifier et atténuer les risques et garantir un niveau de protection approprié en matière de cybersécurité (article 55).

Enfin, **pour toutes les IA génératives**, tous **les contenus générés par des IA génératives** (texte, image…) devront être marqués « dans un format lisible par machine et identifiables comme ayant été créés ou manipulés par une IA »[23].

3.3. Un processus de mise en conformité à industrialiser

Le Règlement sur l'Intelligence Artificielle prévoit de lourdes sanctions pour le non-respect des exigences et obligations en découlant. Ces sanctions ont vocation à veiller à la mise en œuvre du règlement et peuvent prendre la forme d'avertissements ou d'amendes. Elles ont vocation à dissuader les opérateurs concernés par le RIA de contrevenir à celui-ci.

Par exemple :

[22] Voir notamment article 53 et suivant.
[23] *Ibid,* article 50.

- Le non-respect des obligations incombant au fournisseur ou au déployeur est sanctionné d'une amende administrative pouvant aller jusqu'à **15 000 000 euros ou 3% du chiffre d'affaires annuel mondial total** réalisé au cours de l'exercice précédent de l'entreprise.
- La fourniture d'informations inexactes, incomplètes ou trompeuses aux organismes notifiés ou aux autorités nationales pour donner suite à leur demande est sanctionnée d'une amende administrative pouvant aller **jusqu'à 7 000 00 d'euros ou 1% du CA annuel mondial total réalisé** au cours de l'exercice précédent.

Pour ces exemples, le chiffre le plus élevé est celui à retenir pour le calcul du montant de l'amende.

Il ne fait alors nul doute que **pour se mettre en conformité avec le Règlement sur l'Intelligence Artificielle, les entreprises auront tout à gagner à industrialiser le processus**. C'est pourquoi, nous nous sommes attachés à construire ce livrable sur l'opérationnalisation de la gestion des risques afin de donner des premières pistes à ces entreprises.

HUB FRANCE IA

4. Opérationnalisation de la gestion des risques

4. Opérationnalisation de la gestion des risques

4.1. Méthodologie

Comme dans notre livre blanc[5] sur le contrôle des risques, nous suivons le processus de production d'un modèle depuis l'idéation jusqu'au monitoring (figures 34 et 45), et analysons chacune des étapes. A chaque étape, nous cherchons à identifier les tâches à réaliser pour mettre en place les éléments qui seraient nécessaires à un audit de certification : ces éléments doivent être alignés avec les obligations listées dans les tableaux ci-dessus (mais nous rappelons que ce document a été rédigé avant la parution finale de l'*AI Act*, des actes délégués et des normes harmonisées à paraître). Ces éléments peuvent être de natures différentes : un guide ou une procédure, un fichier Excel, un programme / code, un score, un *template* / standard, une liste ou un formulaire à remplir, un outil de *workflow*, etc. Ils doivent permettre à chaque étape d'obtenir un résultat contribuant à satisfaire l'audit de certification : par exemple, un fichier Excel permettra de filtrer les systèmes d'IA selon leur niveau de risque attendu, et une procédure indiquera ensuite comment sélectionner un système IA dans une liste de cas potentiels.

Dans les paragraphes qui suivent, nous présentons ces différents éléments **en avançant dans les étapes successives du processus de production du système d'IA**. Nous nous sommes évidemment appuyés sur les versions successives de l'*AI Act* pour définir ces éléments. Mais l'*AI Act* n'étant pas encore officiel, ceux-ci ne sont pas parfaitement alignés avec les exigences de l'*AI Act* : ce travail restera à faire, en y incluant un volet juridique (que nous n'avons pas envisagé ici) pour arriver à la conformité à la réglementation de l'*AI Act*.

L'ensemble de ces éléments, avec la méthodologie proposée, constitue l'outillage nécessaire à l'opérationnalisation de la gestion des risques en vue de la mise en conformité.

Idéalement, nous pourrions espérer utiliser des outils logiciels mettant en œuvre tout ou partie de ces éléments. Nous présentons dans un dernier paragraphe une démarche d'analyse d'outils du commerce proposant de telles solutions.

Avant de développer ci-dessous, nous présentons sur la Figure 6 une synthèse graphique des éléments servant à une bonne gestion des risques de modèles IA

que nous avons identifiés dans notre étude. Ils sont organisés par type d'artefact et par étapes du processus de production d'un système d'IA. Certains de ces éléments sont de nature structurelle (Procédure, Guide, *Templates*, Score et Code), d'autres sont des livrables spécifiques à chaque cas d'usage notamment les documents, les listes et inventaires et finalement les preuves de workflow dans les processus de décision. Dans toute la suite, ces éléments apparaissent sous un nom en **rouge**.

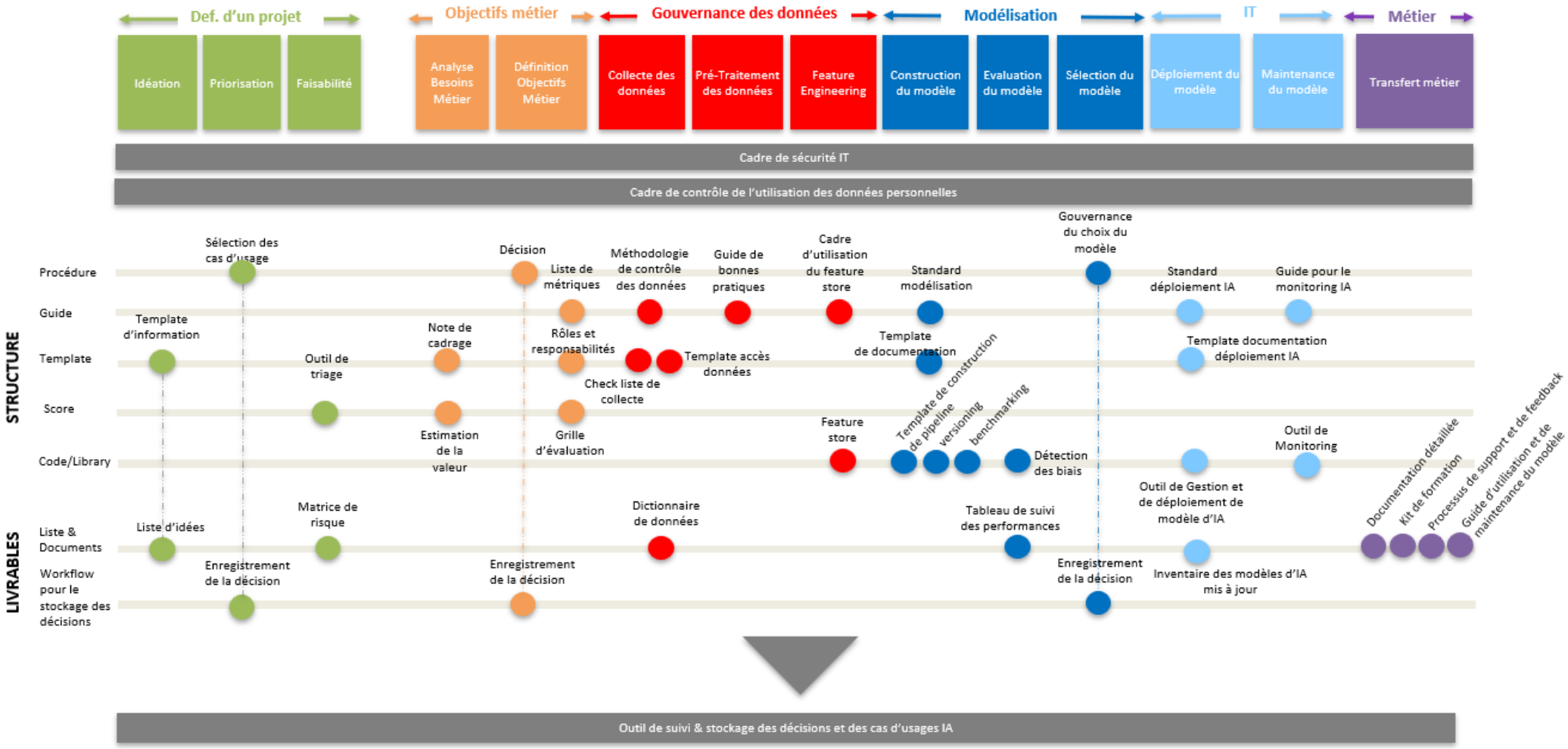

Figure 6 – Synthèse

4.2. Définition d'un projet

4.2.1. Idéation

Le processus d'idéation consiste à encourager la génération d'idées de cas d'usage utilisant de l'IA. De manière schématique, il s'agit d'échanger avec une ou plusieurs personnes appartenant à un domaine métier spécifique et d'encourager par une animation par des experts l'évocation d'irritants ou d'opportunités qui pourraient être adressés par de l'IA.

Dès le processus d'idéation, il est utile de mettre en place un outil de capture des idées de cas d'usage d'IA afin de permettre par la suite la priorisation des efforts. En effet, sans coordination ni priorisation, les efforts seront affectés de manière ad hoc et ne seront guère alignés avec les objectifs stratégiques de l'entreprise (qu'ils soient financiers ou éthiques).

TEMPLATE D'INFORMATION

Afin de structurer la collecte des idées, il est recommandé de définir un **Template** d'informations à collecter pour chaque idée de cas d'usage développée.

LISTE D'IDEES

A la fin du processus d'idéation, on obtiendra ainsi **une liste d'idées** décrites selon le template défini. Cette liste sert de base au processus de priorisation.

4.2.2. Priorisation

Le processus de priorisation a pour objectif de déterminer quelles idées de cas d'usage vont être sélectionnées pour être développées plus en détail. Il s'agit donc de sélectionner et de séquencer les efforts investis dans le développement de l'IA.

PROCEDURE DE SELECTION DES CAS D'USAGE

Pour accompagner ce processus de priorisation, il est ainsi utile de définir **un mode de sélection**, reflétant un cadre de valeur permettant de quantifier l'apport de chaque cas d'usage, par exemple par le biais d'une **procédure et/ou d'un comité de décision.**

OUTIL DE SUIVI DES DECISIONS

Une fois les décisions prises on documentera à la fois les éléments du processus (date, lieu, contributeurs, etc.) et le résultat de la sélection dans un outil de stockage des décisions (*record keeping*).

4.3. Faisabilité

4.3.1. Analyse du risque

Dès ces premières étapes, il est utile de se pencher sur une analyse rudimentaire des risques associés aux cas d'usage afin d'identifier : ceux qui ne sont pas acceptables (souvent pour des questions éthiques, légales, règlementaires) ; ceux pour qui il faut suivre de près le développement afin d'identifier les sources de risques et leur matérialité et mettre en place dès la conception des solutions de contrôles des risques ; ceux enfin qui présentent un risque a priori faible, pour lesquels il faut néanmoins suivre le profil de risque mais de manière moins soutenue que pour les autres. Enfin, il est nécessaire de déterminer en amont s'il y a un risque que le projet ne soit pas implémentable techniquement (exemple : forte dépendance à des données externes, fréquence inadaptée d'alimentation des données, besoins *hardware*).

Cette analyse des risques peut notamment venir affecter la feuille de route ou les priorités qui ont été définies dans les étapes précédentes.

OUTIL DE TRIAGE (SCORE)

Concrètement il s'agit de définir un **outil simple de triage (score)**, qui peut par exemple, s'inspirer de ce qui est proposé dans la réglementation américaine SR-11-7, c'est-à-dire de regarder globalement trois axes (i) l'impact du cas d'usage (ii) sa complexité (iii) la capacité de suivi de la performance s'il va en production. Concrètement, il peut simplement s'agir de définir pour chacun des axes des métriques qui sont ensuite associées à un score, par exemple entre 0 et 1, et qui est agrégé sur les trois axes.

MATRICE DE RISQUE

Un autre exemple consiste à définir une matrice qui produit un niveau qualitatif de risque (faible/moyen/haut) selon la probabilité d'occurrence du risque et sa sévérité.

Notons que si le métier arrive avec un besoin précis / identifié sans passer par la phase d'idéation, il devra néanmoins mener une étude de faisabilité.

4.4. Objectifs métier

4.4.1. Analyse du besoin métier

Nous sommes maintenant dans le contexte d'un projet précis.

Le point de départ officiel du projet consiste à cadrer précisément les besoins métiers associés au projet. Pour cela, il est utile de structurer cette réflexion par le biais **d'une liste de questions** qui organise l'expression de besoin, notamment, les bénéfices attendus, la disponibilité et qualité des données, la capacité à déployer la solution à la fois d'un point de vue technique mais aussi et surtout d'un point de vue organisationnel (impact sur les processus, utilisabilité, acceptabilité, enjeux sociaux). Naturellement, cette liste de questions est une version plus riche de ce qui a été défini dans le **Template** associé à la phase d'idéation.

NOTE DE CADRAGE

Le résultat de cette étape est ainsi **une note de cadrage** d'expression de besoins qu'il est utile de consigner afin de pouvoir la consulter ou de la mettre à jour tout au long du projet. Il est donc aussi pertinent de mettre en place un système de suivi des modifications de cette note de cadrage afin de pouvoir suivre son évolution.

Un des points importants et complexes de cette étape est l'estimation de la valeur apportée par le projet. Cette estimation s'appuie d'un coté sur une analyse de l'existant qui doit être faite par le client (qui connait le processus actuel) puis de la solution cible qui doit être faite en commun entre l'équipe de développement et celle du client.

FICHIER D'ESTIMATION DE LA VALEUR

Pour assurer un certain niveau de comparabilité entre les projets, il est proposé de structurer cette estimation de valeur par **une procédure** et/ou **un outil associé** (qui probablement sera d'abord sous la forme d'une feuille Excel). Il est utile également de mettre à disposition des abaques diverses (coût horaire par pays, temps passé pour certaines tâches communes etc.) afin d'homogénéiser les valeurs des paramètres qui sont utilisés dans la valorisation

4.4.2. Décision de lancement

GRILLE D'EVALUATION (SCORE)

Dans la suite directe de la note de cadrage, une décision officielle, confirmée, de lancement du projet est nécessaire. Cette décision peut s'appuyer sur **une grille d'évaluation** (score) qui combine les analyses précédentes (critères de départ, risques, bénéfices, disponibilité des données, faisabilité technique, impact sur les processus, déployabilité, modèle simple des risques) avec des considérations de disponibilité des équipes techniques.

TABLE DES ROLES ET DES RESPONSABILITES

Avec cette grille d'évaluation vient aussi une spécification claire des rôles et responsabilités qui idéalement s'inscrivent dans une **table prédéfinie des rôles et responsabilités**. Au-delà du rôle de *model developer* et de *model owner*, il est utile de bien spécifier qui est en charge de l'obtention de la donnée, de l'organisation du changement ou des interactions avec les équipes d'IT des systèmes affectés.

PROCEDURE DE DECISION

Comme pour la priorisation, pour accompagner ce processus de décision, il est utile de définir **un mode de décision** par exemple par le biais d'une **procédure,** et une fois la décision prise de documenter à la fois les éléments du processus (date, lieu, contributeurs, etc.) et le résultat dans un outil de *record keeping*.

4.4.3. Spécification des Objectifs Métier

LISTE DE METRIQUES

Si le projet est sélectionné, alors on poursuit la spécification des objectifs métiers en sélectionnant les métriques qui seront suivies lors du projet. Pour simplifier et homogénéiser ce processus au sein de l'organisation, on peut fournir **deux listes de métriques standards** :

- Une première qui stipule **les métriques orientées métier** (performances métier, volumétries de données à traiter, contraintes d'exploitabilité, de latence).
- Une seconde qui liste **les métriques orientées data science** (métriques de performance, contraintes de complexité, d'explicabilité).

La finalité de cette étape est donc de produire **les listes remplies des métriques métier et data science** qu'il faut également consigner et suivre au cours du temps.

Afin d'évaluer et de comparer différents LLM sur des tâches similaires de MMLU (*Massive Multi-task Language Understanding*), plusieurs benchmarks ont été développés. On peut citer par exemple *Open LLM Leaderboard*[24], *LMSys Chatbot Arena Leaderboard* [25] ou *HELM* [26] (Stanford). Ces benchmarks évaluent les IA Génératives avec des métriques adaptées sur des données correspondant à des tâches précises comme répondre à des QCM, résumer, effectuer de l'analyse de sentiment, récupérer des informations, raisonner, générer du code. Les métriques employées permettent ainsi d'évaluer différents axes comme la concordance (*exact match* pour les QCM ou analyse de sentiment), F1 score (pour des tâches de réponses aux questions sans réponses proposées), ROUGE-N (résumé), nombre de codes générés passant les tests unitaires (génération de code), représentation démographique (biais), proportion de générations toxiques (toxicité).

4.5. Gouvernance des données

En préambule, il est bon de rappeler que la démarche de Gouvernance des données dans le cadre de la construction d'un modèle d'IA pourra s'appuyer dans le cadre des institutions financières, sur les principes établis par le comité de Bâle avec le BCBS 239.

En effet, celui-ci définit des normes afin d'assurer la maitrise des données, leur connaissance et surtout leur qualité[27]. Compte tenu des risques relatifs aux biais, à l'utilisation de données personnelles et sensibles et à l'interprétabilité des données et à la traçabilité des sources de données utilisées, les exigences attendues par le BCBS 239 et le RGPD contribuent fortement à la maitrise de ces risques.

Les systèmes d'IA à réapprentissage automatique en continu complexifient la gouvernance des données. On peut citer par exemple :

- La complexité pour tracer et assurer des contrôles des données de réentraînement du fait de la nature dynamique de l'apprentissage. Il faut être en mesure de mettre en place des contrôles à la volée pour s'assurer de la

[24] https://huggingface.co/open-llm-leaderboard
[25] https://chat.lmsys.org/?leaderboard
[26] https://crfm.stanford.edu/helm
[27] Les critères de qualité portent notamment sur l'exactitude, l'intégrité, l'exhaustivité et l'actualité des données.

qualité (exemple : détection de variables aberrantes, manquantes, non valides, …) avant la mise à jour des paramètres du modèle.

- L'impossibilité de garder la version exacte du modèle ayant servi à faire les prédictions. Il peut donc être nécessaire de stocker les explications des prédictions (valeurs de Shapley[28] par exemple).

4.5.1. Collecte des données

CHECK-LIST DES PRINCIPES DE COLLECTE DES DONNEES

Une ***check-list* des principes de collecte des données** doit être établie afin de définir les bonnes pratiques et les questions à se poser autour de la collecte des données et pour s'assurer d'une utilisation adéquate des données.

Cette *check-list* doit a minima faire apparaitre les éléments suivants :

- S'interroger sur la pertinence et l'adéquation des sources de données utilisées pour le modèle,
- S'assurer de la conformité réglementaire sur l'usage des données envisagées (ex : RGPD, *AI Act*, etc.),
- Préciser l'entrepôt et l'architecture de stockage des données à utiliser, leurs accessibilités et la sécurité associée,
- S'assurer de la fraicheur et de la disponibilité des données en définissant la fréquence de collecte et d'alimentation,
- Définir les contrôles à mettre en œuvre afin d'assurer la qualité des données collectées, et ce de manière pérenne dans le temps,
- Renseigner les informations sur les données collectés au sein d'un dictionnaire de données afin d'en faciliter leur réutilisation.

Tout élément complémentaire issu des principes du RGPD pourra être ajouté à cette check-list afin d'assurer la conformité des traitements relatifs aux données sensibles et personnelles.

[28] Christoph Molnar, Giuseppe Casalicchio, Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. Joint European conference on machine learning and knowledge discovery in databases. Cham: Springer International Publishing, 2020. https://arxiv.org/pdf/2010.09337

TEMPLATE DE FORMULAIRE D'ACCES ET D'ACCREDITATION

En complément de la check-list, chaque donnée ou source de données doit faire l'objet d'une autorisation d'accès et d'usage par le propriétaire de la donnée. **Ce formulaire d'accès et d'accréditation définit notamment :**

- Les conditions d'utilisabilité et la durée d'utilisation,
- Les *Service Level Agreements* (*SLA*) sur lesquels s'engage le propriétaire de la donnée,
- Les conditions d'accès et d'accréditation.

4.5.2. Connaissance des données

Une bonne connaissance des données est nécessaire afin de maîtriser le risque de biais ainsi que la bonne interprétation des données utilisées pour l'entrainement des modèles.

Afin de centraliser, recenser et partager les informations autour des données et de leurs usages, il est nécessaire de disposer **d'un dictionnaire de données**. Il est fortement recommandé aux entreprises de s'équiper d'un outil de gouvernance des données pour documenter les dictionnaires, les usages, les sources et les contrôles de leurs données critiques.

DICTIONNAIRE ET/OU CATALOGUE DE DONNEES

Le dictionnaire de données doit permettre notamment d'expliquer et de centraliser les données utilisées par le système d'IA :

- La description des données du groupe, leurs emplacements techniques et le format de structure des données à exploiter,
- Les *Golden sources* sur les données prioritaires définies par l'entité,
- Les propriétaires de chaque donnée en charge d'assurer leur qualité,
- Les règles d'usage de chaque donnée et les contraintes réglementaires associées (RGPD, durée de stockage AMF, etc.),
- Les usages actuels de ces données,
- Selon le niveau de maturité de l'entité, le *Data Lineage*. C'est-à-dire la traçabilité du cycle de vie de la donnée expliquant son parcours de production (depuis la *golden source*) jusqu'à son déversement dans les différents entrepôts décisionnels.

4.5.3. Qualité des données

La qualité des données est un impératif afin de lutter contre les biais et de s'assurer de l'efficacité du modèle.

Toutefois, les données transitant par différentes bases et applications avant leur exploitation, il est primordial de s'assurer de leur qualité tout au long de leur cycle de vie par la mise en place de contrôles réguliers : de la *golden source* dans des bases opérationnelles jusqu'aux entrepôts décisionnels. Le développement des systèmes d'IA permet d'identifier les données qui seront critiques lors de leur phase de production, de définir le niveau de qualité attendu, et de mettre en place les contrôles nécessaires au suivi de la qualité. Ce déploiement des contrôles doit s'appuyer sur la gouvernance existante de l'organisation (notamment en capitalisant le cadre mis en place pour le respect des exigences réglementaires comme BCBS 239).

CONTROLE DES DONNEES

Pour ce faire, la constitution d'une **méthodologie de contrôle des données** permet à chaque entité de définir :

- Les responsables de la qualité des données (Entité Productrice vs Consommatrice),
- Les critères de qualité à contrôler : exhaustivité, actualité, unicité, conformité, intégrité, exactitude et cohérence,
- Leur construction et leur déploiement opérationnel : on pourra suivre une méthodologie du « *Quoi ? Où ? Quand ? Comment ?* » pour mettre en place les contrôles,
- Le dispositif de suivi des résultats et de gestion des disqualités (non-qualités).

Le déploiement opérationnel de ces contrôles doit ensuite faire l'objet d'un suivi régulier.

4.5.4. Pré-traitement des données

Une bonne préparation des données est fondamentale pour assurer la fiabilité, l'efficacité du modèle et la conformité réglementaire des modèles d'IA. Cela permet notamment d'assurer la maîtrise des biais dans les données et ouvre la voie à l'augmentation des données en cas de données manquantes ou erronées.

GUIDE DE BONNES PRATIQUES

Pour cela, il est recommandé de s'appuyer sur **un guide de bonnes pratiques (ou même une procédure dans certains cas réglementaires)** servant à la préparation et au contrôle des données avant leur utilisation dans un modèle. Le guide devra faire apparaitre les étapes suivantes :

- Compréhension du domaine et du jeu de Données en s'appuyant notamment sur le dictionnaire de données afin de :
  - Identifier l'objectif du projet et les questions clés à résoudre,
  - Comprendre les sources de données, leur provenance et leur pertinence,
- Collecte et Intégration des Données :
  - Rassembler les données à partir de différentes sources si nécessaire,
  - S'assurer d'une bonne Intégration et fusionner les jeux de données de manière cohérente,
  - Définir des contrôles de données assurant leur qualité au moment de l'intégration dans les tables (Cf. critères de qualité du BCBS 239 tels que la complétude, la cohérence, l'exhaustivité, l'exactitude, etc.)),
- Nettoyage des Données :
  - Traiter les valeurs manquantes (suppression, imputation, etc.),
  - Corriger les erreurs et incohérences dans les données (fautes de frappe, valeurs aberrantes, etc.),
  - Pérenniser ces traitements de manière automatique,
- Exploration et Analyse des Données :
  - Réaliser une analyse exploratoire (statistiques descriptives, visualisations, etc.),
  - Identifier les tendances, patterns et anomalies potentielles,
- Transformation et Normalisation des Données :
  - Transformer les données pour les rendre exploitables (encodage de variables catégorielles, normalisation, règles de valeurs par défaut, etc.),
  - Redimensionner les données si nécessaire (réduction de dimensionnalité, etc.),
- Sélection et création de Caractéristiques :
  - Identifier et sélectionner les caractéristiques (*features*) les plus pertinentes pour le modèle (Cf. 4.5.5 Augmentation des données),
  - Éliminer les caractéristiques redondantes ou peu informatives,
- Division du Jeu de Données pour l'analyser et tester le modèle :

    - Diviser le jeu de données en ensembles d'entraînement, de validation et de test,
- Documentation du jeu de données facilitant la connaissance et l'auditabilité du modèle :
    - Documenter le processus de préparation des données, en particulier les règles de transformation des données,
    - Créer et maintenir des métadonnées décrivant le jeu de données,
- Sécurité et Confidentialité :
    - Assurer la conformité aux réglementations de confidentialité des données (ex : RGPD),
    - Anonymiser les données sensibles si nécessaire,
- Backup et Versioning des Données :
    - Sauvegarder régulièrement le jeu de données,
    - Utiliser des systèmes de versioning pour suivre les modifications.

En outre, tous les traitements modifiant les jeux de données pourront être tracés afin d'assurer un historique des modifications apportées.

A noter que la plupart des solutions de Data Science intègrent des modules de préparation de données telles que Alteryx, Dataiku ou Rapidminer. A défaut, certains langages de bases de données proposent également des librairies pour la préparation de données telles que DBT ou Pandas.

4.5.5. Augmentation des données

FEATURE STORE[29]

Comme évoqué précédemment, les « *features* » (ou caractéristiques) sont essentielles pour améliorer les jeux de données et les modèles : elles permettent d' « augmenter » les données en y incluant des données dont les experts connaissent l'importance (par exemple, la moyenne mobile sur une période passée en *trading*) ; l'utilisation de données synthétiques est une autre technique d'augmentation que nous ne traiterons pas ici. Disposer d'un ***feature-store*** permet donc de centraliser, et de standardiser et de partager ces « *features* » afin d'assurer leur qualité et leur réutilisabilité.

[29] Cf. §6. Glossaire

En outre, un *feature-store* optimise les processus de création de modèles en fournissant un accès facile et rapide aux « *features* » déjà préparées et validées, réduisant ainsi le temps et les efforts nécessaires pour la préparation des données.

CADRE D'UTILISATION DU FEATURE STORE

Afin de disposer d'un *feature-store* pérenne et facilement exploitable, il est préconisé de définir **un cadre stipulant les règles et bonnes pratiques pour son usage**. Parmi celles-ci, il est important de préciser notamment :

- Le standard des données et des *features* à stocker (Définition de schémas de données, utilisation de pipelines ETL (*Extract*, *Transform*, *Load*) pour normaliser les données, etc.),
- Les droits d'accès et les usages autorisés selon les profils des utilisateurs,
- L'accessibilité et l'interopérabilité avec les outils de Data Science ou de développement (Usage d'API, langages de programmation autorisés, etc.)
- Les métadonnées associées aux *features* (versions des *features*, source de création, historique de modification, utilisateurs des *features*, etc.)

4.6. Modélisation

4.6.1. Construction du modèle

STANDARD DE CONSTRUCTION DU MODELE

Une procédure ou un standard est un document détaillant les étapes de construction d'un modèle. Il permet une meilleure homogénéité des travaux (facilitant la reprise en main en cas de départ du data scientist ou d'audit) et de s'assurer que certaines étapes clés ne sont pas oubliées par les data scientists.

TEMPLATE DE DOCUMENTATION DU MODELE

Cette procédure peut être complétée par un *template* de documentation du modèle, qui permet d'uniformiser les attendus, de s'assurer que les éléments pertinents de la construction sont expliqués et détaillés, de simplifier la lecture et la compréhension pour des équipes de revue indépendantes (cf. *Model Risk Management*).

TEMPLATE DE CONSTRUCTION DE PIPELINE

Un autre moyen important pour améliorer la traçabilité et la transparence de la modélisation consiste à utiliser un *template* de construction de pipeline, qui permet de décomposer chaque étape et chaque flux d'information, facilitant ainsi la compréhension du système d'intelligence artificielle.

OUTILS DE BENCHMARKING

La définition de cibles de performance pour le modèle est une étape importante : quelle métrique de performance choisir en fonction du cas d'usage ? quelle est la performance que le système d'IA doit atteindre pour être viable ? Les benchmarks peuvent être :

- Le processus existant que le système d'IA cherche à remplacer ;
- L'utilisation de règles ;
- La performance humaine ;
- Des systèmes d'IA basés sur des méthodologies différentes.

L'utilisation d'outils d'auto Machine Learning (AutoML) permet d'industrialiser la création de modèles alternatifs (dits *challenger models*) pour conforter les choix de modélisation et sélectionner le meilleur modèle pour un cas d'usage et des contraintes donnés (ex. : métriques de performances à optimiser, contraintes à satisfaire). Un outil d'AutoML présente généralement comme résultat principal le classement des différents modèles en fonction d'une métrique de performance choisie par le modélisateur.

Parmi les outils d'AutoML disponibles, on peut citer : H2O, Flaml, Autosklearn, Datarobot, Mljar.

N.B. : Il est tout à fait possible de créer ses propres *challenger models* sans utiliser d'outil d'AutoML.

OUTILS DE VERSIONING

Lors de la construction d'un système d'IA, il est judicieux de conserver les différentes versions des modèles développés avec leurs hyperparamètres et les performances associées. MLflow est un exemple d'outil permettant d'enregistrer les expériences et les configurations du modèle, de comparer les performances et les paramètres.

D'autres outils de versioning (ex. : Github, GitLab, SVN) permettent de gérer et de stocker du code de façon collaborative. Les modifications sont ainsi historisées et traçables.

4.6.2. Evaluation du modèle

TABLEAU DE SUIVI DES PERFORMANCES

La première étape dans l'évaluation d'un modèle consiste à établir la liste des indicateurs à optimiser ou à satisfaire :

- Le **choix d'une métrique de performance** à optimiser selon le cas d'usage (ex. : *accuracy*, F1 score, précision, rappel pour la classification, ou R^2 pour la régression. Pour les IA génératives le choix de la métrique de performance va dépendre des capacités générales utilisées. On peut citer l'*exact match accuracy* pour les questions/réponses ou ROUGE-N (*Recall-Oriented Understudy for Gisting Evaluation*) pour les tâches de type résumé. Cette dernière métrique évalue la similarité entre un résumé généré automatiquement et un résumé de référence fait par un humain en comparant les séquences de n mots consécutifs (*n-grams*) entre les deux textes ;
- La **définition de contraintes à satisfaire**. Elles peuvent être techniques (ex : utilisation de CPU ou de GPU, temps de latence, taille du modèle, nombre de *tokens* générés pour les LLM), éthiques (ex. : *fairness*), en lien avec la transparence (ex : capacité à comprendre les décisions du modèle et à s'assurer que la prédiction est faite pour les bonnes raisons), la robustesse ou la quantification de l'incertitude liée aux prédictions.

Ces indicateurs vont permettre d'établir un classement des modèles, appelé également *leader board*.

OUTIL DE DETECTION DES BIAIS

Certains outils permettent de détecter et de traiter les biais (ex : AIF360, Fairlearn), d'autres sont spécifiques à la production d'explications (ex : SHAP, AIX360, interpretML) ou à la mesure d'incertitude (ex : MAPIE).

L'analyse des indicateurs peut se faire globalement ou sur certaines régions du *dataset* pour identifier de potentiels problèmes sur des sous-populations. Par

exemple, il peut être intéressant d'expliquer les prédictions erronées du modèle grâce aux techniques d'XAI[30] pour mettre en lumière des biais potentiels.

4.6.3. Sélection du Modèle

Les modèles candidats, évalués à l'étape précédente, qui ne satisfont pas les contraintes ou dont les performances sont jugées insuffisantes sont éliminés. Parmi les modèles restants, les parties prenantes dans le développement du système d'IA doivent s'accorder sur la liste des critères de sélection principaux (ex. : performance, degré d'interprétabilité, absence de biais) et faire potentiellement des compromis (ex. : performance et explicabilité) en fonction du cas d'usage. L'analyse par le métier de la pertinence des variables explicatives les plus importantes pour les prédictions est clé dans le processus de sélection du modèle. La faisabilité technique liée à l'implémentation en production est un autre critère important dans la sélection du modèle. Par exemple, certains modèles nécessitent l'utilisation de GPU ou de solutions *cloud* pour pouvoir fonctionner correctement ou avec un temps de latence acceptable. Des limitations techniques comme celle du *hardware* de l'environnement de production peuvent réduire le choix des techniques de modélisation.

GOUVERNANCE DE VALIDATION DU CHOIX D'UN MODELE D'IA

La mise en place d'une gouvernance de validation d'un système d'IA est une bonne pratique. Elle permet de définir les modalités de décision et de désigner les personnes responsables de l'approbation. La revue par une équipe indépendante de celle en charge de la modélisation permet de conforter ou de remettre en question le choix du modèle et sa construction.

Dans les institutions financières, cette gouvernance de choix du modèle à mettre en production pourrait s'appuyer sur le cadre de gestion des risques de modèles existant. De surcroît pour les entreprises plus matures, il est recommandé d'intégrer autant que possible la validation dans le processus de développement du modèle.

[30] eXplainable AI

4.7. IT

4.7.1. Déploiement du modèle

Après développement, un modèle est généralement déployé dans un environnement de production. Ce déploiement reste avant tout un projet IT et doit alors suivre les bonnes pratiques et règles de l'entreprise.

STANDARD DE DEPLOIEMENT D'UN MODELE D'IA

Dans le cadre des systèmes d'intelligence artificielle, le MLOps (*Machine Learning Operations*) est un bon exemple de pratiques normalisées visant à déployer des modèles de *Machine Learning* en production de manière fiable et efficace, et couvre généralement toutes les phases de développement, de mise en production ainsi que le monitoring des modèles.

Par exemple, le MLOps couvre les éléments suivants :

- *Data Engineering* : collecte et préparation des données et des *features ;*
- Construction du modèle : modélisation et entraînement du modèle ;
- Validation du modèle : vérification de la performance du modèle et d'autres critères de risques ;
- Déploiement du modèle : intégration du modèle dans l'environnement IT et les processus métiers ;
- Monitoring : vérification continue ou périodique des critères d'acceptation du modèle tels que la performance.

Pour chaque phase, divers degrés d'automatisation sont possibles :

- Processus manuel : les étapes sont réalisées manuellement et le passage d'une phase à une autre l'est également ;
- Pipeline d'automatisation : des étapes d'automatisation sont développées au sein de « pipelines », par exemple, pipeline de sélection de variables, pipeline d'entrainement, etc. ;
- CI/CD : l'ensemble de la chaîne est automatisable au travers d'outils de « *Continuous Integration* » i.e. automatisation et planification des étapes de développement et de tests d'un modèle, et de « *Continuous Deployment* » i.e. automatisation des contrôles et du déploiement d'un modèle.

L'automatisation d'un tel processus n'est possible que si :

- Le processus est clair et documenté de bout en bout ;

- Les rôles et responsabilités sont clairement définis pour chaque acteur et à chaque étape ;
- Les équipes IT sont associées dès les premières phases du projet pour garantir une transition fluide vers l'environnement IT de production, et anticiper les défis potentiels liés à l'intégration et au déploiement ;
- Les coûts du déploiement modèle d'IA sont clairement évalués, car le passage à l'échelle de tels projets engendre souvent des coûts insoupçonnés, ce qui devient un point de blocage (sous-estimation des ressources nécessaires, inefficacités dans la gestion des ressources, absence de support, …) en fin de projet.

TEMPLATE DE DEPLOIEMENT D'UN MODELE D'IA

Afin de faciliter le déploiement d'un modèle d'IA, les informations suivantes devraient être formalisées :

1. **Description du modèle** : Une description détaillée du modèle d'IA, y compris son objectif, les données qu'il utilise et les algorithmes employés ;
2. **Indicateurs de performance du modèle** : Une liste d'indicateurs de performance pour évaluer la précision et la qualité du modèle lors du test et de la validation ;
3. **Modalités de déploiement du modèle** : Des instructions étape par étape sur la façon de déployer le modèle dans un environnement de production, comprenant l'infrastructure utilisée, les configurations logicielles et les pipelines d'intégration des données ;
4. **Modalités de suivi du modèle** : Des détails sur la façon de suivre le modèle en production, en particulier les indicateurs de performance, les outils à utiliser et la fréquence de suivi ;
5. **Plan de mise à jour et de maintenance du modèle** : Un plan pour mettre à jour et maintenir le modèle au fil du temps (en continu ou de manière périodique).

OUTIL DE GESTION ET DE DEPLOIEMENT DE MODELE D'IA

L'orchestration de ces étapes est de mieux en mieux automatisée au travers d'outils facilitant à la fois les déploiements et le suivi des modèles passés en production. Plusieurs outils sont disponibles :

- MLflow assiste les *data scientists* dans le développement et le déploiement de modèles, ainsi que la gestion du cycle de vie et l'évaluation des modèles, le suivi des composants créés (*features*, *pipelines*, configurations, etc.), jusqu'au *monitoring* ;
- D'autres outils existent tels que Kubeflow, Comet, etc., ainsi que les outils nativement intégrés au sein des fournisseurs de service cloud pour gérer les services et applications proposés.

INVENTAIRE DES MODELES D'IA

Un inventaire des modèles d'IA est essentiel pour comprendre où ils sont déployés et utilisés, renforçant la transparence de l'utilisation de cette technologie auprès des lignes métiers et des clients.

Avoir un inventaire permet également de suivre l'évolution de l'utilisation de l'IA dans une entreprise et de la maturité des équipes réalisant ces modèles. Ainsi, une vision claire et transparente aidera également à répondre à des exigences de gestion des risques et de leur suivi.

4.7.2. Maintenance du modèle

GUIDE POUR LE MONITORING DU MODELE

Ce guide accompagne les équipes de 1LOD dans la construction du monitoring de modèle. Il liste les rôles et responsabilités, par exemple qui sera en charge du calcul des métriques de monitoring ou qui décidera de réentraîner le modèle. Le guide de monitoring permet aussi de lister les axes classiques à surveiller (ex. dérive de distribution des données, qualité des données, dérive de la performance, incidents, métriques IT). Il propose et définit ainsi des exemples de métriques classiques employées pour surveiller ces aspects. Il suggère de manière générique des exemples d'actions envisageables suivant les niveaux de ces métriques. Enfin, il liste tous les autres éléments importants à définir suivant la matérialité du modèle, comme la fréquence de monitoring ou l'infrastructure de données requise.

Un *Template* de protocole de monitoring peut ainsi être utile afin d'homogénéiser les méthodes des équipes chargées de monitoring. Cela permet aussi d'accélérer le process d'implémentation du monitoring.

OUTILS DE MONITORING

MLflow permet de suivre les expériences réalisées, comme l'évaluation d'un modèle sur une période de production avec l'affichage des métriques de performance calculées. MAPIE est une librairie Python qui permet de quantifier l'incertitude des prédictions d'un modèle. Une incertitude croissante peut être signe de dérive de distribution dans les données en entrée.

Par ailleurs, comme lors de la construction d'un modèle, des outils d'explicabilité peuvent générer des explications à la demande : ex., SHAP, AIX360.

Enfin, dans une logique de suivi des émissions carbone, des packages Python tels que CodeCarbon ou CarbonTracker peuvent être intégrés au pipeline d'inférence.

4.8. Transfert au métier

Le transfert au métier d'un modèle d'IA est indispensable lors de sa mise en œuvre opérationnelle. Cette phase comporte un ensemble d'éléments qui ne sont pas spécifiques à l'IA, tels que la documentation, la formation, la conduite du changement, la mise en place d'un dispositif de contrôle et de surveillance des usages par le métier.

Dans un premier temps, l'ensemble des individus impactés par le projet doivent être identifiés. Pour un modèle interne à une entreprise, il s'agira des utilisateurs directs du modèle ainsi que de leur ligne hiérarchique.

Il est important que les utilisateurs bénéficient d'instructions claires pour l'intégration du modèle dans le processus métier, qu'ils deviennent autonomes dans l'interprétation des résultats du modèle et en connaissent les limites d'utilisation. Afin de faciliter le transfert, des exemples d'utilisation et des conseils pour la maintenance continue du modèle pourront également être proposés.

Les documents à fournir pour faciliter le transfert d'un modèle d'intelligence artificielle vers un métier peuvent inclure :

DOCUMENTATION DETAILLEE

Globalement, il faut pouvoir fournir une documentation détaillée du modèle qui inclut des informations sur la conception du modèle, sur son déploiement dans les systèmes d'information, sur les processus suivis pour la validation technique et IT ainsi que des informations sur la sécurité et la confidentialité (notamment la

protection des données sensibles), des informations sur les limites d'utilisation du modèle. Cette documentation peut être divisée en plusieurs documents si besoin.

Cette documentation doit être compréhensible par des non data scientists : détails sur l'architecture du modèle, le type d'algorithmes choisi et les paramètres clés. De plus, une documentation sur les données utilisées peut également être proposée au métier, incluant les informations clé sur les jeux de données utilisés pour l'entraînement, y compris leur source, leur qualité, et les prétraitements effectués.

La documentation doit intégrer des **métriques de performance** (résultats d'évaluation du modèle) mesurant sa précision.

GUIDE D'UTILISATION ET DE MAINTENANCE DU MODELE

Ce guide a pour vocation d'être plus succinct que la documentation détaillée et à destination des utilisateurs finaux, il contient des instructions sur la manière d'intégrer le modèle dans les processus métier existants, sur les limites d'utilisation du modèle ainsi que des informations sur la maintenance du modèle, notamment les métriques qui sont suivies et le processus de monitoring.

KIT DE FORMATION

Il s'agit de développer un kit de formation intégrant des exemples d'utilisation concrets montrant comment tirer le meilleur parti du modèle dans des scénarios réels, et les principales caractéristiques du modèle (synthèse des éléments déjà cités). Il est recommandé d'organiser des sessions de formation à partir de ce support.

PROCESSUS DE SUPPORT ET DE FEEDBACK

Enfin pour accompagner l'utilisation opérationnelle de la solution d'IA, il faut définir un processus clair de rôles et responsabilités qui peut inclure notamment les informations suivantes :

Contacts de support : Fournir des coordonnées pour un support technique en cas de questions ou de problèmes

Template de feedback utilisateur permettant à l'utilisateur du modèle de remonter de potentielles faiblesses du modèle

4.9. Phases transverses

Les phases transverses (Figures 4 et 5) visent à assurer la sécurisation de l'application IA (modèle et données). Les tâches classiques de cybersécurité protègent les systèmes d'IA de la même façon que les autres systèmes informatiques en sécurisant les accès ou en protégeant les données. Cependant, les communautés IA et cyber commencent à prendre conscience de la particularité de l'IA qui peut être attaquée de façons différentes : en effet de très nombreuses attaques spécifiques à l'IA (attaques adversarielles) commencent à être répertoriées [31] faisant apparaître de nouveaux risques. Par exemple, des **attaques par empoisonnement** des données d'apprentissage vont dégrader les performances du modèle ; des **attaques par évasion** (ou manipulation pour l'IA-Gen) vont altérer les données pendant la production de façon à modifier les prévisions du modèle ; des **attaques de confidentialité** (ou plus généralement **d'exfiltration**) vont exploiter le système d'IA en production pour en extraire des données (personnelles notamment) ; des **attaques de violations d'abus** (pour IA-Gen) vont réorienter le système par injection de prompt pour promouvoir haine ou discrimination, disséminer du code malveillant etc.

Des pistes de mitigation sont proposées dans le document du NIST, et aussi dans celui de MITRE[32]. Cependant, dans beaucoup de cas, surtout pour l'IA Générative[33], il n'existe pas encore de solution complète ou infaillible pour protéger les modèles.

La meilleure méthode resterait évidemment de concevoir et développer des systèmes d'IA sécurisés « by design », ce que recommandent l'ANSSI avec les grands organismes de cybersécurité mondiaux [34] : malheureusement, ces préconisations restent encore peu opérationnelles face aux risques, dont on estime qu'ils sont en forte croissance.

[31] Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Anderson. Adversarial Machine Learning. NIST. January 2024. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf

[32] https://atlas.mitre.org/matrices/ATLAS

[33] https://owasp.org/www-project-top-10-for-large-language-model-applications/

[34] ANSSI. Recommandations de sécurité pour un système d'IA générative. Guide. 29 avril 2024. https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative
UK National Cyber Security Centre. Guidelines for secure AI system development. 2023. https://www.ncsc.gov.uk/files/Guidelines-for-secure-AI-system-development.pdf

4.9.1. Sécurisation des modèles et des données

CADRE DE SECURITE IT

La sécurisation d'un modèle se construit à toutes les étapes du cycle de vie d'un modèle IA car les sources de vulnérabilités ou potentielles surfaces d'attaques sont multiples :

- En premier lieu, un **cadre global de sécurité IT** (ou cybersécurité) doit être mis en place, ce qui est généralement le cas pour les établissements bancaires déjà soumis à diverses réglementations. L'objectif est notamment de s'assurer de la robustesse des composants (matériels et logiciels) utilisés au cours du cycle de vie des modèles d'IA, par exemple, empêcher des accès non autorisés aux données servant à l'entraînement d'un modèle, aux plateformes de production, ou aux équipements réseaux.
- Le développement de modèles d'IA s'appuie fortement sur des **composants open-source**, pouvant eux-mêmes être des vecteurs d'attaque car porteurs de vulnérabilités ou de virus. Il convient de vérifier la provenance de ces composants, et de réaliser des scans pour détecter d'éventuels virus ou vulnérabilités dans les librairies.
- De la même manière que pour les composants open-source, des **modèles IA open-source** peuvent être utilisés pour développer un nouveau modèle ou intégré dans un logiciel. La vérification de la source, et détecter la présence de virus est nécessaire afin de ne pas introduire des éléments perturbateurs dans le système d'information.
- Des contrôles complémentaires peuvent également être réalisés au moment de l'utilisation des modèles afin de détecter des **attaques dites « adversarielles »**, telles que le *data poisoning*, le *prompt injection* ou la manipulation de données en entrée afin de perturber les valeurs générées par un modèle d'IA. Ces types d'attaques sont en recrudescence et font l'objet d'une attention particulière des équipes de cybersécurité et de recherche aujourd'hui.

Par ailleurs, des vérifications portant sur ces différents aspects liés à la sécurité devraient être régulièrement réalisées afin d'identifier les contrôles inopérants et ainsi renforcer ce même dispositif de contrôles.

4.9.2. Protection des données personnelles

L'entraînement d'un modèle d'IA nécessite un ensemble plus ou moins important de données. Il convient de s'assurer que ces données peuvent être réellement utilisées, c'est-à-dire que les données sont licites et ne présentent pas un problème juridique associé, par exemple, dans certains cas, obtenir le consentement des personnes afin de pouvoir utiliser leurs données personnelles. Nous avons couvert dans la section 4.5 les risques liés à la gouvernance des données, nous nous concentrons ici sur les données personnelles.

Les politiques de protection des données diffèrent selon les régions et les pays. Voici une liste non exhaustive de politiques existantes :

- **Règlement Général de Protection des Données** (RGPD) permet de collecter de manière directe ou indirecte des données personnelles, sous des conditions de collecte, d'information ou de consentement desdites personnes.
- Le « ***EU-U.S. Privacy Shield Framework*** » est un ensemble de principes de protection des données personnelles auxquelles les entreprises établies aux Etats-Unis d'Amérique sont libres d'adhérer. Les entreprises établies dans l'Espace économique européen peuvent transférer les données personnelles qu'elles traitent à destination des sociétés américaines figurant sur la liste « *EU-U.S. Privacy Shield Framework* », de la même manière que s'opèrent les transferts vers les pays reconnus comme "adéquats" par la Commission européenne.
- Le « ***Act on the Protection of Personal Information*** » au Japon est considéré comme compatible avec le RGPD européen.

De plus en plus de pays mettent en place des règlementations similaires de protection des données personnelles, néanmoins, par rapport à l'Europe, seule une petite partie est considérée comme « adéquate ». Pour les autres pays, les transferts de données personnelles doivent être encadrées par des outils ou contrats de transferts.

CADRE DE CONTROLE DE L'UTILISATION DES DONNEES PERSONNELLES

Le travail d'identification des impacts des réglementations sur le modèle d'IA doit être réalisé en amont et pendant le développement du modèle, notamment avec des Référents Informatique et Liberté, ou toute autre fonction pouvant répondre à ces problématiques.

4.10. Risques spécifiques à l'IA Générative

Dans cette section, nous évoquons succinctement les risques spécifiques qui sont associés à la conception de solutions s'appuyant sur l'IA générative (il est à noter que l'expérience sur ces sujets est beaucoup moins développée que pour les systèmes d'IA prédictive, largement déployés). Globalement, nous avons identifié trois grandes familles de risques.

Tout d'abord, **les risques afférents aux données** dans lesquels nous retrouvons :

- **Risque de violation de la propriété intellectuelle**, au niveau de l'obtention des données d'entraînement, mais également au niveau du contenu généré (droit d'auteur, copyright). Il est à noter que ce risque est renforcé par l'usage des modèles pré-entraînés, dont on ne sait pas sur quelles données ils ont été entraînés ;
- **Risque de reproduction et de propagation des biais** présents dans les données d'entraînement ;
- **Problèmes de confidentialité des données**, possibilité de fuite de données, manque de fraicheur des données d'entraînement (*~model Staleness*) ;
- **Manque de transparence** : données d'entraînement, explication du texte généré (RAG).

La seconde catégorie concerne **le modèle** en lui-même et les erreurs et comportements déviants possible. Nous identifions notamment :

- **Risque d'hallucinations** : génération d'un contenu plausible, mais faux ;
- **Génération de contenus toxiques ou nocifs** : nécessité de mettre en place des garde-fous.

Enfin, nous avons rassemblé dans une troisième catégorie **les risques transverses** qui sont largement facilités par l'IA générative ainsi que les possibles effets néfastes pour la société :

- **Failles de cybersécurité** régulièrement mises au jour (domaine peu mature encore) ;
- **Facilitation des activités malveillantes de cybersécurité** : utilisation malveillante, *jailbreak attempts*, *prompt injection ;*
- **Augmentation du risque réglementaire** : génération de textes inappropriés (par rapport à une réglementation spécifique comme la protection de la clientèle) ;

- **Impact environnemental**, que ce soit sur la consommation d'énergie ou d'eau notamment ;
- **Influence sur la souveraineté** : dépendance à des tierces parties (*LLM provider*) ;
- **Impacts économiques et sociétaux** : effets à moyen terme sur le monde du travail, sur certaines industries.

4.11. Synthèse des outils d'opérationnalisation

Le déploiement et l'utilisation des outils précédemment décrits au sein d'une organisation peuvent être accélérés par l'emploi d'une solution logicielle « sur étagère » du marché. Ces solutions sont de deux types :

- **Solution en tant que plateforme** : suite d'outils disponible en tant que plateforme utilisable par l'organisation. La plateforme est supportée par son éditeur et est accessible moyennant une souscription / un abonnement / un contrat auprès de celui-ci. La plateforme sera majoritairement proposée en tant que service SaaS et accessible depuis un navigateur web. Certains éditeurs proposent des prestations de services complémentaires (formation, conseil, accompagnement) venant compléter leur offre.
- **Solution en tant que service de conseil outillé** : proposés par des prestataires, ces services sont destinés à accompagner une organisation dans la production de ses systèmes d'IA. Afin de réaliser leurs activités, les prestataires peuvent s'appuyer sur des outils, sans nécessairement en donner l'accès aux clients. Ces outils peuvent être la propriété du prestataire (développés par ses soins, pour son unique usage), ou bien acquis (souscription à un service tiers, ou achat de propriété intellectuelle).

Le Hub France IA, avec le groupe de travail Banque et Auditabilité, a cherché à évaluer l'offre existante en France : nous avons pour cela élaboré une méthodologie d'analyse des offres et interviewé un panel d'éditeurs de solutions. Le travail mené est décrit par le Hub France IA dans la note technique[35] publiée en juin 2024. Cette note présente les conclusions de l'analyse des solutions et se positionne comme un guide dans la prise de décision quant au choix d'une solution devant répondre

[35] Hub France IA. Juin 2024. https://www.hub-franceia.fr/telecharger-le-pdf-operationnaliser-la-gestion-des-risques/

aux besoins spécifiques d'une organisation. Il propose également une méthodologie d'évaluation des solutions.

Cette étude, réalisée dans le cadre du groupe Banque et Auditabilité, reflète naturellement l'expertise du secteur financier, qui est particulièrement régulé notamment au niveau de la gestion de modèles. Les résultats de l'analyse peuvent cependant fournir des conclusions applicables pour d'autres secteurs économiques.

Dans un premier temps, nous décrivons en détails la méthodologie employée par les experts pour identifier, évaluer et analyser les solutions. Puis, les résultats de l'analyse sont exposés dans le chapitre suivant.

Note : L'analyse des solutions citées dans cette étude ne reflète pas la qualité absolue de celles-ci. Cette analyse illustre le niveau de couverture de chaque solution vis-à-vis d'attentes propres au processus d'entreprise décrit dans le présent document, c'est-à-dire pour sa capacité à opérationnaliser le processus de mise en conformité à l'AI Act. Ainsi, la note globale présentée pour chaque solution est à mettre en perspective selon son contexte d'utilisation.

La production de l'analyse s'est organisée en 3 temps :

1. Construction de la grille d'analyse à utiliser pour évaluer la solution
2. Sélection et analyse des solutions à l'aide de la grille d'analyse
3. Traitement des données et restitution par l'intermédiaire de la note technique

L'exercice a débuté en octobre 2023 et s'est terminé lors de la publication de la note. Le Hub France IA en est l'unique contributeur.

Etape 1 : Construction de la grille d'analyse

Les solutions sont analysées à l'aide d'une grille construite à partir du processus de production d'un système IA. Chaque élément d'analyse représente une étape du processus. L'analyse consiste, pour chaque étape, à évaluer si la solution étudiée propose des services relatifs à l'étape en question.

Pour chaque étape, est attendu un **niveau de couverture** :

- **Couvert** : la solution propose les services permettant l'exécution de l'étape ;
- **Partiellement couvert** : la solution propose des services pouvant contribuer à l'exécution de l'étape OU des futures évolutions sont prévues sur la solution pour permettre l'exécution de l'étape, ceci dans un avenir proche ;

- **Non couvert** : la solution ne propose pas de services permettant l'exécution de l'étape.

Etape 2 : Analyse des solutions

Notez que cette étude ne propose pas un référencement exhaustif des solutions répondant au besoin de pilotage du processus de production de modèles IA, mais repose seulement sur un panel.

L'analyse des solutions se déroule en trois étapes successives :

1. L'i**dentification** : recensement des solutions analysées et citées dans le document ;
2. La **démonstration** : présentation générale de chaque solution par l'éditeur ;
3. L'**analyse** : remplissage de la grille d'analyse.

L'identification des solutions retenues est menée par la communauté des adhérents du Hub France IA. Cette identification, puis la sélection de chaque solution, a été motivée par les raisons suivantes :

- Une organisation membre du Hub France IA est déjà utilisatrice de la solution ;
- L'éditeur de la solution est une entreprise adhérente du Hub France IA.

Etape 3 : Traitement des données

À la suite de la phase d'analyse engendrant les grilles complétées et cohérentes, le Hub France IA traite les données, pour produire les résultats présentés dans la note technique.

Ces résultats sont :

- **Quantitatifs** : calculés à partir des niveaux de couvertures de la grille d'analyse ;
- **Qualitatifs** : issus des commentaires justifiant la couverture de chaque étape du processus, ils apportent une description textuelle de chaque solution.

Un « radar » permet notamment de positionner chaque solution par rapport à sa couverture des besoins de chacune des phases du processus de production du système d'IA.

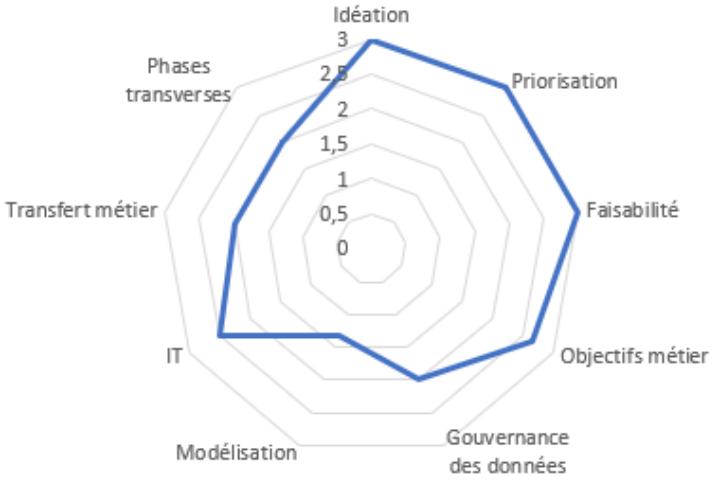

Figure 7 – Radar de couverture d'une solution sur les phases du processus IA

HUB FRANCE IA

5. Conclusion

5. Conclusion

L'analyse des risques que nous venons de décrire fait apparaître **beaucoup d'outils** à mettre en place tout au long du processus de mise en œuvre du système d'IA. Nous avons montré, sans être exhaustifs sans doute, l'importance d'outils de suivi capables d'accompagner le processus qui viendront s'ajouter aux outils informatiques habituels et qui devront faciliter la mise en conformité nécessaire avec l'arrivée de l'AI Act.

Le marché voit apparaître beaucoup de **solutions logicielles d'opérationnalisation de l'AI Act**, mais, comme nous l'avons vu à travers les solutions que nous avons analysées, la couverture de ces solutions reste encore parcellaire. Il n'y a pas encore de solution qui s'est imposée, on doit bien sûr s'attendre à ce que le marché se renforce dans les années à venir. Par ailleurs, nous n'avons pas cherché à aligner l'adéquation de ces outils avec les **obligations juridiques de l'AI Act**, qui n'était pas encore publié au moment de nos travaux (les normes harmonisées et les actes délégués non plus). Il est encore difficile de prévoir comment l'AI Act se mettra en place et impactera la pratique de la gestion des risques : le domaine est en pleine évolution.

Enfin, **l'IA générative** commence à se déployer dans les entreprises. Il est déjà évident qu'elle fait apparaître de nouveaux risques, mais avec encore beaucoup de flous juridiques (droit d'auteur, copyright par exemple). Ce domaine est donc à suivre dans les années (mois ?) à venir pour mettre en place les outils nécessaires au contrôle des risques.

Nous espérons que ce travail, fait au groupe de travail Banque du Hub France IA apportera des éléments utiles à tous les secteurs qui vont se trouver confrontés avec l'AI Act.

6. Glossaire

AutoML	Aussi appelé *Machine Learning* automatisé. Automatise les tâches de développement d'un modèle *Machine Learning*, par exemple la préparation des données, la sélection des variables, l'entrainement, etc.
BCBS239	*Basel Committee on Banking Supervision's standard number 239* : Principes aux fins de l'agrégation des données sur les risques et de la notification des risques
ChatGPT	Chatbot développé par OpenAI, fondé sur un grand modèle de langage
Chunk	Bloc d'information extrait d'un ensemble de données plus large
CI/CD	*Continuous Integration / Continuous Delivery* ou intégration continue/déploiement continu regroupe des pratiques d'intégration automatique des modifications de code dans un référentiel, d'intégration et de tests de ces modifications, puis de déploiement automatique dans les environnements de production.
CNIL	Commission Nationale de l'Informatique et des Libertés
CPU	*Central Processing Unit* : microprocesseur principal d'un ordinateur
Dataset	Ensemble des données utilisées dans l'une des phases de modélisation
Deep Learning	Sous-ensemble du *Machine Learning* fondé sur l'utilisation de réseaux de neurones dits profonds, c'est-à-dire utilisant de multiples couches de neurones
Embedding	Un *embedding* est une représentation vectorielle de grande dimension pour un objet tel qu'un mot, un document ou une image. Ces représentations contribuent au calcul de l'attention dans les modèles Transformers et permettent de savoir quels mots sont proches sémantiquement

ETL	*Extract, Transform, Load* : logiciel permettant de collecter des données provenant de multiple sources
Feature	Variable explicative en entrée du modèle
Feature store	C'est une base de données qui permet de stocker, de partager et de gérer de manière efficace les « *features* », afin notamment de faciliter leur réutilisation entre projets et modèles, pour l'entraînement ou en production.
Few-shot learning	L'intégration dans le *prompt* de l'IA Générative de quelques exemples de la tâche à effectuer est appelée *few-shot learning*.
Fine-tuning	Le *fine-tuning* d'une IA Générative pré-entraînée consiste à lui faire exécuter un entraînement supplémentaire sur des données labellisées spécifiques d'une tâche ou d'un domaine particulier afin d'améliorer sa performance
GPT	*Generative Pretrained Transformer* : c'est une famille de *Large Language Models* développée par OpenAI
GPU	*Graphics Processing Unit* : processeur spécialisé dans le rendu d'image, le traitement d'image 2D/3D, et de calculs mathématiques
Guardrails	Ce sont des protections qui permettent de contrôler les entrées et sorties d'une IA Générative afin de réduire les risques liés à son utilisation
Hallucination	Information fausse, inexacte ou incohérente créée par une IA Générative
Human-in-the-loop	Intervention humaine dans le processus de décision
IA Générative	Sous-ensemble du *Deep Learning*, visant à produire du contenu, que ce soit du texte, une image, de l'audio ou une vidéo, à partir de données en entrée (on parle alors de *prompt*), elles-mêmes du texte, une image, de l'audio ou une vidéo.

IA Générative pré-entraînée	Une IA Générative est dite « pré-entraînée » si elle a été entraînée sur de très larges volumes de données, lui permettant d'acquérir des connaissances générales et les bases pour générer du contenu pertinent.
KPI	*Key Performance indicator* : Indicateur de performance
Large Language Model (LLM)	Un type d'IA générative capable de générer et d'analyser du texte (par exemple : langage naturel, langage de programmation…)
LOD	*Line Of Defense* : niveau de contrôle composant le contrôle interne d'un établissement
Machine Learning	Apprentissage automatique à partir d'un ensemble de données
MLOps	Ensemble de pratiques qui vise à déployer et maintenir des modèles de *Machine Learning* en production de manière fiable et efficace
Model Owner	Acteur clé qui a la responsabilité de s'assurer que le développement du modèle d'IA, son implémentation, son usage, et son suivi dans le temps soient conformes avec les politiques et procédures de la banque
Model Risk Management	Gestion du risque de modèle
Override	Décision humaine d'outrepasser et de changer un résultat donné par un système
POC	*Proof Of Concept*. Désigne une réalisation ayant pour but de démontrer la faisabilité d'un projet
Prompt	Le *prompt* est l'instruction ou la requête en langage naturel fournie à l'IA Générative dans le but d'obtenir une réponse
RAG	Retrieval *Augmented Generation* : génération augmentée via la récupération d'informations d'une base de connaissances qui n'a pas été utilisée lors de l'entraînement de l'IA Générative
RGPD	Règlement Général sur la Protection des Données

ROUGE-N	Rappel basé sur des N-grams entre un résumé généré et un ensemble de résumés références
SLA	*Service Level Agreement* : contrat de service entre un prestataire informatique et un client
SSI	Sécurité des Systèmes d'Information, voir la norme internationale ISO/CEI 27001 ainsi que l'autorité nationale de sécurité des systèmes d'information (ANSSI)
Supervised Learning	Méthode de *Machine Learning* où un modèle est entraîné sur un ensemble de données labellisées ou avec des valeurs cibles, i.e. à chaque entrée est associée un label ou valeur cible, permettant ainsi au modèle d'apprendre à prédire les labels ou valeurs cibles sur de nouvelles données.
Température	La température dans le cadre d'une IA Générative est un paramètre du modèle permettant de gérer le caractère aléatoire d'un texte généré (par exemple). La température varie généralement entre 0 et 1 ; une valeur proche de zéro générera un texte quasi identique à chaque génération, alors qu'une valeur proche de 1 générera un texte avec plus de créativité ou variabilité.
Token	Sous-ensemble d'un mot constituant une unité de traitement par un *Large Language Model*
Unsupervised Learning	Méthode de *Machine Learning* où un modèle est entraîné sur un ensemble de données sans étiquettes ou valeurs cibles. Le modèle cherche à découvrir des relations entre les données pour par exemple les grouper (*clustering*) ou réduire la dimension du problème
Zero-shot Learning	Le *zero-shot learning* consiste à alimenter l'IA Générative avec un *prompt* « sec », sans exemple de la tâche à effectuer

7. Remerciements

Contributeurs

- **Audrey Agesilas**, Superviseur – Model risk Audit, Société Générale ;
- **Benjamin Bosch**, Manager - Model risk Management – Data Science, Société Générale ;
- **Thomas Bonnier**, Model Risk Manager – Data Science, Société Générale ;
- **Pierre Dehaene**, Data & IA Strategist, La Banque Postale ;
- **Lea Deleris**, AIR Tech, RISK x Compliance, BNP Paribas ;
- **Jérôme Lebecq**, Data Science Coordinator, BNP Paribas ;
- **Cyril Nicolotto,** Chef de projets, Hub France IA ;
- **Chloé Pledel,** Responsable des affaires européennes et réglementaires, Hub France IA.

Relecteurs

- **Caroline Chopinaud**, Directrice Générale, Hub France IA.
- **Françoise Soulié-Fogelman**, Conseiller Scientifique, Hub France IA.

La touche finale

- **Mélanie Arnould**, Responsable des opérations, Hub France IA.

BNP PARIBAS

SOCIETE GENERALE

LA BANQUE POSTALE

Opérationnaliser la gestion des risques des systèmes d'intelligence artificielle

GT Banque et auditabilité

Juillet 2024

HUB FRANCE IA