LIVRE BLANC
GT Banque et Auditabilité
Hub France IA

Contrôle des risques des systèmes d'Intelligence Artificielle

Octobre 2022

BNP PARIBAS

LA BANQUE POSTALE

SOCIETE GENERALE

HUB FRANCE IA

CONTROLE DES RISQUES DES SYSTEMES D'INTELLIGENCE ARTIFICIELLE

GROUPE DE TRAVAIL - BANQUE ET AUDITABILITE - HUB FRANCE IA

1 TABLE DES MATIERES

2 INTRODUCTION

Le succès du recours à l'Intelligence Artificielle (IA) et la multiplication de ses usages dans la majorité des domaines industriels et scientifiques s'accompagnent naturellement de **l'émergence de nouveaux risques**. Dans ce contexte, la définition ou le renforcement de réglementations encadrant l'IA se profile dans de nombreuses régions, notamment au niveau de l'Union Européenne, pour favoriser un **développement maîtrisé** de ces techniques.

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, souhaite partager ses réflexions et son retour d'expérience en matière de gestion des risques liés à l'IA.

Ces travaux présentent des réponses à certaines problématiques clés et se positionnent comme un guide de bonnes pratiques pour évaluer et maîtriser les risques des solutions basées sur l'IA. Le processus IA sera couvert de bout en bout, grâce aux regards croisés des trois lignes de défense représentées dans ce groupe de travail.

Cette organisation spécifique en **lignes de défense** est propre au secteur financier, qui est par ailleurs particulièrement régulé notamment au niveau de la gestion de modèles. Cette organisation en lignes de défense permet de structurer l'approche des établissements en matière de **gestion des risques**. Ils s'appuient sur une organisation et un cadre éprouvé. La déclinaison de cette approche dans le cas de l'IA peut dès lors fournir des pistes de réflexion pour d'autres secteurs économiques. Au-delà de l'organisation, le principal apport de ces travaux réside probablement dans la déclinaison opérationnelle des dispositifs de maîtrise envisagés. Ils sont conçus ici à dire d'expert, et non en réponse explicite par rapport à l'une ou l'autre des réglementations en gestation.

La démarche adoptée consiste dans un premier temps à décrire le processus IA dans son ensemble, y compris sur les aspects relevant de sa mise en conformité. Dans une deuxième section, les risques spécifiques apportés ou exacerbés par l'IA ont été identifiés et illustrés. Enfin, dans un troisième temps, dix risques ont été sélectionnés, à la fois sur la base de leur importance mais aussi de leur lien spécifique avec l'IA. Pour chacun, des propositions d'évaluation d'impact et de remédiation ont été détaillées.

Le travail 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.

1 LE PROCESSUS IA

1.1 DEFINITION 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 »[1]. Elle comprend deux grandes familles : **l'IA symbolique et l'IA numérique**, qui ont chacune connu des périodes de succès et des « hivers », comme le montre le schéma ci-dessous représentant l'activité de ces deux familles depuis les années 50.

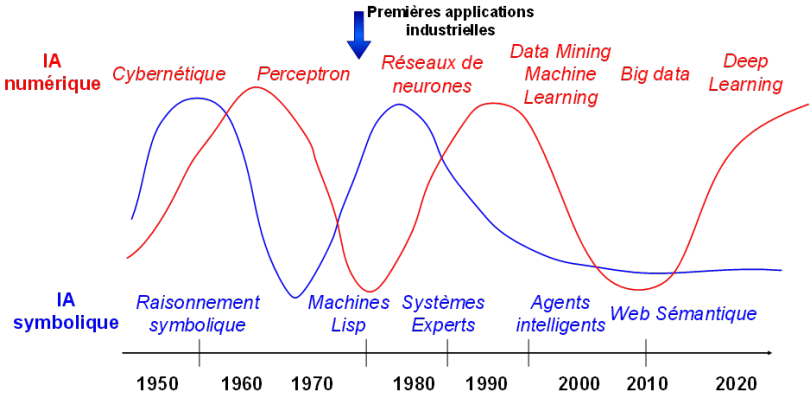

Figure 1 – IA symbolique et IA numérique

Nous ne parlerons pas de l'IA symbolique, peu utilisée dans le monde bancaire. Nous nous concentrerons sur **l'IA numérique**. Depuis 2012 et le succès des techniques *Deep Learning* à la conférence *ImageNet*[2], les techniques les plus répandues aujourd'hui, et notamment dans les applications bancaires, sont les techniques du ***Machine Learning*** (ou apprentissage automatique), dont fait partie le *Deep Learning*. Dans ce livre blanc, lorsque l'IA sera évoquée, il s'agira donc de *Machine Learning*, sauf mention contraire explicite.

1.2 L'IA DANS LA BANQUE

L'intelligence artificielle accompagne de plus en plus de processus bancaires, la tendance s'étant fortement accélérée ces dernières années.

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. Des cas d'usage sont développés par exemple pour lire et traiter automatiquement des documents dans le domaine juridique ou à l'entrée en relation avec les clients. D'autres usages sont dédiés à la classification des e-mails entrants avec proposition de réponse, réduisant ainsi le temps de traitement des demandes clients. L'IA facilite également l'extraction et la structuration de gros volumes de données, afin par exemple d'alléger les travaux des analystes crédit.

L'usage de l'IA vise également à **améliorer l'expérience client**, avec l'apparition d'agents conversationnels qui assistent les clients dans leurs opérations (Chatbots, Voicebots), parfois

[1] https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257
[2] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems pp 1097-1105. 2012. papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

appelé *selfcare*. Dans le domaine du marketing, l'IA permet d'affiner la connaissance des clients afin de leur proposer des produits et des conditions tarifaires plus personnalisés, en permettant de mieux anticiper leurs besoins et prévoir l'attrition. En outre, l'IA facilite l'émergence de nouveaux services pour les clients, tels que par exemple des moteurs de recommandations pour l'investissement (*robo-advisors*).

En améliorant la précision des algorithmes, le recours à l'IA contribue à l'**atténuation de nombreux risques**, notamment **le risque opérationnel** (e.g. détection des fraudes commises sur les clients, détection d'anomalies dans les données), le **risque de crédit** (e.g. détection des clients les plus risqués à l'octroi de crédit) et la **conformité** (e.g. dans le cadre de la lutte contre le blanchiment d'argent et le financement du terrorisme, la détection des « nouvelles négatives » à propos des clients dans le cadre du KYC[3]). On trouve aussi des applications sur les modèles financiers et la gestion de ces risques, ainsi qu'en ALM (*Asset and Liability Management*).

Enfin, il est important de rappeler que l'évaluation de la pertinence d'une IA doit être effectuée au regard de l'appétence au risque ou l'exposition au risque tolérée de l'institution. Cette évaluation de la pertinence doit aussi être effectuée en fonction de l'ensemble des risques réduits ou accrus par l'usage de l'IA, notamment en complément du risque de modèle, les risques opérationnels, de conformité, de crédit ou de marché.

Si pour l'intelligence artificielle il n'existe pas à ce jour de cadre précis et globalement partagé définissant l'ensemble des éléments relatifs aux impacts et risques des systèmes d'IA, il faut cependant noter que le développement et le déploiement de systèmes d'intelligence artificielle s'appuient sur des **dispositifs internes** en forte adhérence, plus avancés dans leur encadrement :

(i) Gestion de la donnée et notamment de la qualité de la donnée ;
(ii) Cyber sécurité ;
(iii) Gestion des risques de modèles au sein des institutions financières.

Ces dispositifs bénéficient d'une réglementation, de pratiques professionnelles et d'expertises en matière de gestion du risque bien établies qui inspirent les propositions qui suivent pour l'IA.

En outre, les institutions financières sont très aguerries au développement de modèles, notamment au **contrôle de la discipline de modélisation**. Il apparaît que si ce sont des risques pertinents pour les modèles d'intelligence artificielle (*feature selection*, *hyper parameter tuning*, surapprentissage, manque de documentation), ils ne sont pas prépondérants dans notre industrie. En revanche, il est important d'y prêter attention lorsque les modèles sont développés en externe ou dans des équipes traditionnellement éloignées de la modélisation.

1.3 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 sélectionner un algorithme 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

[3] *Know Your Costumer*, i.e. Connaissance client

courant étant aujourd'hui python. Les données utilisées pour l'apprentissage sont nécessairement des données du passé. On utilise aussi un *dataset de validation*, différent du *dataset* d'apprentissage, mais ayant la même structure, pour pouvoir comparer des modèles entre eux et choisir le meilleur. En général, on découpe aléatoirement l'ensemble de toutes les données disponibles en trois parties, par exemple 70% / 15% / 15%. Pour l'apprentissage, on utilise la première partie pour produire des modèles, et la seconde pour choisir le meilleur modèle (choix des hyperparamètres). La troisième partie est réservée pour tester le modèle et ne sera jamais vue pendant l'apprentissage. Si l'on dispose d'un nombre trop limité de données, des techniques de validation croisée sont employées.

- **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. On utilise donc des données du passé pour déterminer un **comportement prédictif** du futur. Il est recommandé de collecter au fil de l'eau ces données et les résultats associés, que l'on compare à ce qui se produit vraiment, ce qui permet de mesurer l'erreur de prévision (on dit qu'il y a une *erreur* s'ils sont différents). On peut ensuite, à la fréquence souhaitée, relancer l'apprentissage du modèle en incorporant ces nouvelles données. On met ainsi en place une **boucle de réapprentissage** qui permet d'améliorer le modèle au cours du temps.

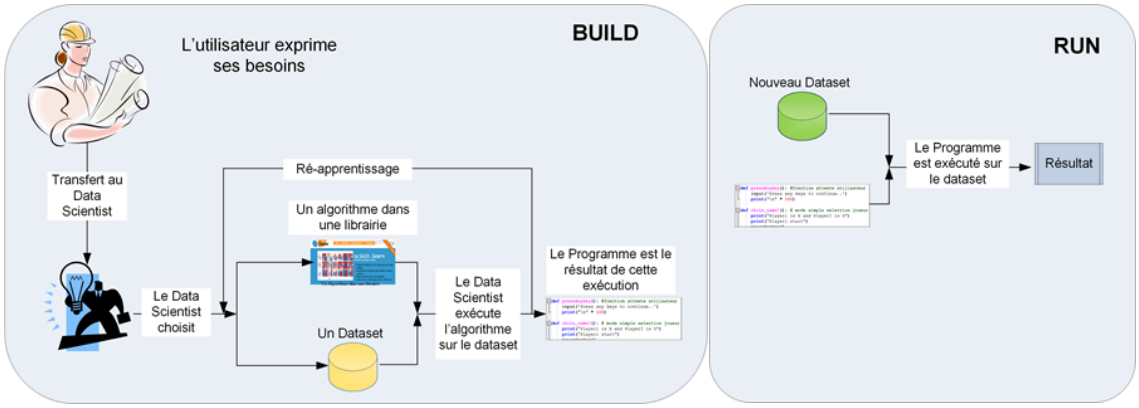

Figure 2 – Les deux étapes de production et utilisation d'un modèle de *Machine Learning*

Les algorithmes d'apprentissage correspondent essentiellement à deux grandes approches. Dans la première, on connait la bonne réponse associée à une donnée et dans la seconde, on ne la connait pas.

Ces deux approches sont définies plus précisément comme suit :

- L'**apprentissage supervisé** : les données comprennent chacune une étiquette, que l'on appelle une **labellisation**. Pour les algorithmes de classification ; il s'agit de l'étiquette de classe (par exemple *chien* ou *chat* pour des images). Pour les algorithmes de régression, il s'agit d'une étiquette numérique (par exemple le montant de la fraude sur une carte). L'apprentissage supervisé a pour but de réduire l'erreur sur le *dataset* d'apprentissage. On fait passer l'un après l'autre les données du *dataset*, puis pour chaque donnée, on compare le résultat obtenu par le modèle à l'étiquette et on mesure l'erreur. L'apprentissage consiste à tester itérativement pour un algorithme de nombreuses combinaisons des données d'entrées pour retrouver les données de sortie qui correspondent aux labels. L'approche itérative permet de progressivement diminuer ces erreurs, et ainsi d'apprendre. La qualité de l'apprentissage peut fortement dépendre du nombre d'événements « connus »

(c'est-à-dire les données labellisées). S'ils sont en quantité suffisante, on pourra obtenir un bon modèle. Sinon, il faudra utiliser une autre approche

- L'**apprentissage non supervisé** : les données ne comprennent pas d'étiquette. Il y a de nombreuses approches, consistant en particulier à regrouper les données similaires (algorithmes de *clustering*) de manière à identifier des associations, des différences entre données sans *a priori*.

Il existe d'autres approches moins utilisées, comme **l'apprentissage semi-supervisé** ou **l'apprentissage par renforcement**. Les techniques de vision, de reconnaissance de la parole ou de traitement de la langue naturelle (NLP, *natural language processing*) font très souvent appel à des techniques à base de **réseaux de neurones**, profonds en général. Ces derniers permettent de construire une **représentation** des données dans un espace de dimension généralement plus faible. On parle d'***embedding*** dans l'espace de représentation. D'autres techniques d'*embedding* permettent de représenter du texte (word2vec[4]) ou des graphes (graph2vec[5]). On peut alors par exemple catégoriser des documents, en faire des résumés automatiques, ou encore automatiser leurs identifications.

Une fois en production, on s'attend à ce que le modèle se comporte sur les données de production avec des performances comparables à celles obtenues pendant la phase de construction du modèle : on dit que le modèle « *généralise* bien ». Cette propriété est fondamentale, et il est primordial de s'assurer que le modèle que l'on utilise ne soit pas en situation d'***overfitting*** (ou sur-apprentissage). En effet, lorsqu'un modèle est trop complexe, il va pouvoir « apprendre par cœur » le *dataset* d'apprentissage et ne pourra pas généraliser correctement en production. La Figure 3 ci-dessous illustre ce phénomène. Lorsque l'on augmente la **complexité** (techniquement il s'agit de la dimension de Vapnik Chervonenkis[6]) d'un modèle (par exemple le nombre de neurones d'un réseau de neurones, ou la profondeur d'un arbre de décision), l'erreur sur le *dataset* d'apprentissage décroît, en même temps que l'erreur sur l'ensemble de validation. Quand le modèle devient trop complexe, l'erreur d'apprentissage continue à décroître (le modèle a appris le *dataset* par cœur) mais l'erreur en validation se met à augmenter : le modèle ne généralise plus, il « **sur-apprend** » (ou « *overfit* ») le *dataset* d'apprentissage.

[4] Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint,* 2013. https://arxiv.org/pdf/1301.3781.pdf
[5] Narayanan, Annamalai, Mahinthan Chandramohan, Rajasekar Venkatesan, Lihui Chen, Yang Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. arXiv preprint, 2017. https://arxiv.org/pdf/1707.05005.pdf
[6] Vladimir Vapnik – Estimation of Dependences based on empirical data. Springer. Information sciences and Statistics. Reprint of 1982 Edition with afterword. 2006.

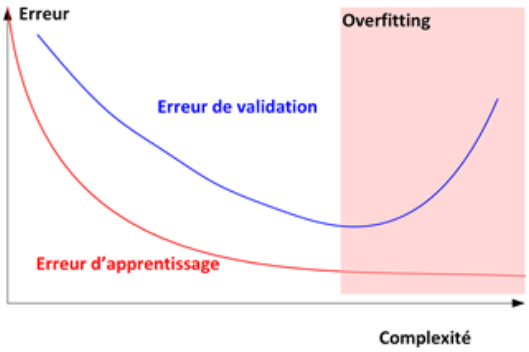

Figure 3 – Erreurs d'apprentissage et de généralisation en fonction de la complexité du modèle

Quel que soit le mode d'apprentissage, le nombre total d'algorithmes n'est pas très élevé (avec beaucoup de variantes). On pourra consulter la *cheatsheet*[7] qui les résume en cinq pages (denses !) ou la synthèse[8].

Il faut noter que l'on utilise en pratique **trois familles d'indicateurs de performance** : les **indicateurs techniques** sont utilisés en apprentissage pour optimiser le modèle IA, les **indicateurs métier** permettent de mesurer la valeur métier générée par l'utilisation du modèle, et enfin aussi les **indicateurs plus opérationnels**, comme la durée du calcul, le temps de latence, le nombre de variables et la complexité du modèle, voire même le coût des variables s'il y en a qui sont achetées.

1.4 LE PROCESSUS DE PRODUCTION DU MODELE

Le **processus de production** d'un modèle, décrit dans la figure ci-dessous, comprend différentes étapes, avec potentiellement des retours en arrière pour itération tant que l'on n'est pas satisfait du résultat. La figure suivante (Figure 4) illustre la situation réelle avec les itérations et échanges potentiels. La répartition des tâches entre métier et IT peut différer selon les organisations.

[7] https://github.com/aaronwangy/Data-Science-Cheatsheet/blob/main/Data_Science_Cheatsheet.pdf
[8] https://towardsdatascience.com/overview-of-supervised-machine-learning-algorithms-a5107d036296

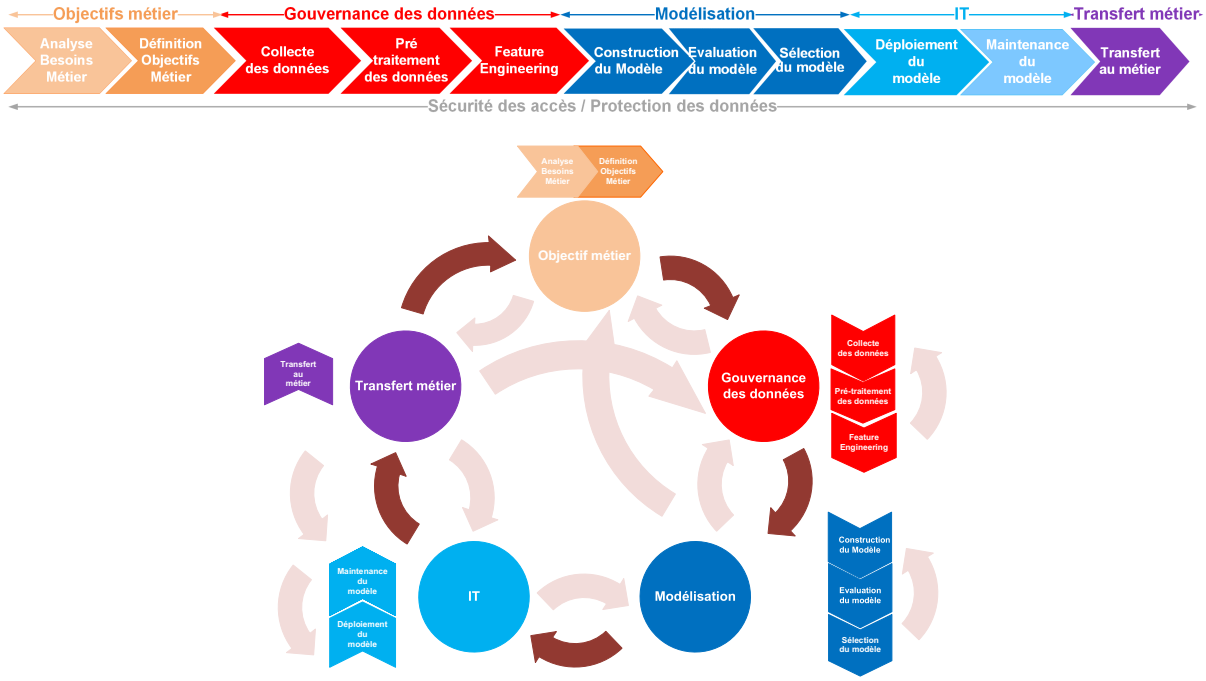

Figure 4 - Processus de production d'un modèle en représentation linéaire (en haut) et avec itérations (en bas)

L'évaluation des risques d'une solution IA doit donc se faire **tout au long du processus** et il est essentiel de s'assurer que les bons intervenants soient positionnés là où cela est nécessaire.

1.5 LES BIAIS

La **notion de biais** est essentielle à définir, car elle reviendra à plusieurs reprises dans ce livre blanc. Les biais ne sont pas propres aux systèmes d'IA. Ils peuvent être cependant amplifiés ou cachés par la complexité de certains modèles.

Le biais peut être défini de différentes façons suivant son origine. Historiquement, au sens statistique, le biais d'un modèle est la différence entre **l'espérance d'un estimateur $\hat{f}$ et la grandeur à estimer f.**

$$Biais(\hat{f}) = \mathbb{E}(\hat{f}) - f$$

Mesurer un biais revient à quantifier des erreurs systématiques dans un modèle.

Le biais statistique n'est plus le sens visé quand on parle aujourd'hui de **données biaisées**. Ce sera le cas si un groupe, par exemple les femmes, est sous-représenté dans le *dataset* d'apprentissage. Un *dataset* biaisé produira alors un modèle qui n'a pas pu apprendre correctement le groupe sous-représenté. Le modèle est alors susceptible de **produire des discriminations** au détriment de ce sous-groupe. Les biais peuvent prendre différentes formes dans les données et dans leur manipulation sur tout le cycle de vie d'un système d'IA. Ils sont donc présents à plusieurs niveaux :

- **Biais sociétal** : **problème de représentativité** qui provient du fait que les données ont été recueillies dans un cadre social et historique spécifique n'étant plus adapté à la situation actuelle d'utilisation du modèle. Par exemple la proportion de femmes parmi les 500 CEOs des plus grandes entreprises mondiales a beaucoup évolué depuis les 50 dernières années.

Cela limite la pertinence de cette information pour une modélisation vouée à être appliquée au monde actuel ;

- **Biais de sélection** : **problème de représentativité** des populations car certains jeux de données ont été créés à partir d'un sous-groupe de la population globale qui ne représente pas pleinement le périmètre d'application. Par exemple, dans le cas des modèles d'octroi, il est fréquent d'utiliser uniquement les données des crédits acceptés et financés dans le passé pour construire les modèles. Ainsi tous les dossiers rejetés ne sont pas pris en compte dans la modélisation, créant ainsi un **biais de sélection**[9] ;
- **Biais cognitif** au cours de l'annotation des données : l'annotateur peut reproduire une méconnaissance sociale ou culturelle dans le choix des étiquettes à utiliser sur des images par exemple ;
- **Biais d'association** : ce biais peut être présent lorsque des variables du modèle sont corrélées avec des attributs sensibles ou protégés qui ne peuvent pas être utilisés dans le modèle. Un système d'IA peut dans certains cas appuyer ses décisions sur des variables latentes (i.e. non observables directement) représentant des sous-groupes d'individus. On parle également de **biais encodé** car la variable sensible est encodée par les variables sélectionnées dans le modèle ;
- **Biais d'évaluation** : lorsque la mesure de performance n'est pas faite sur un jeu de données de test indépendant des données d'entraînement, la mesure du pouvoir de généralisation du modèle est faussée (à cause, par exemple, de l'utilisation d'une proportion importante des mêmes observations entre les données d'entraînement et de test) ;
- **Biais algorithmique** : le choix de l'algorithme peut également **influencer les prédictions** car certains algorithmes peuvent amplifier une sous/sur-représentativité d'une classe d'individus par exemple ;
- **Biais d'automatisation** : lié au fait que les utilisateurs pourraient **privilégier les résultats de systèmes automatisés** par rapport à ceux issus de systèmes non automatisés, faisant abstraction de leur pensée critique ;
- **Biais d'interaction** des utilisateurs : les utilisateurs vont **orienter les futures prédictions** du modèle en renseignant des données spécifiques, notamment pour les modèles à réapprentissage continu. C'est le cas pour beaucoup d'applications web[10] (**biais de position, biais de popularité**, etc.) ;
- **Biais de rétroaction** (*feedback bias*) : l'utilisation des résultats de l'IA peut également créer un biais quand une personne va suivre ce résultat, et que ce dernier est réinjecté dans l'apprentissage conduisant au renforcement de ce résultat de façon automatique. C'est un cas particulier de **biais d'interaction**.

Cette liste de biais n'est pas exhaustive, mais elle met en exergue que le biais peut avoir un impact sur l'IA qui est construite à partir des données. Par ailleurs, il est difficile de corriger un biais indépendamment des autres, du fait de leur **interdépendance**.

[9] Des techniques comme la *reject inference* peuvent aider à atténuer le biais de sélection en prenant en compte les demandes de crédit rejetées dans la base d'entraînement.

[10] Baeza-Yates, R. Bias on the Web. Communications of the ACM 61, no. 6: 54-61. 2018. http://classes.eastus.cloudapp.azure.com/~barr/classes/comp495/papers/Bias-on-web.pdf

1.6 LES ACTEURS

Le développement, la mise en production et le suivi d'un système IA fait appel à de nombreux acteurs, dont certains sont spécialistes de *data science* / IA et d'autres ne le sont pas. Le *data scientist* doit toujours avoir des **compétences métier et des compétences IT** en plus de son expertise.

Comme le montre la figure 5, il y a trois grandes familles d'acteurs :

- Les profils « métier » ;
- Les spécialistes de *data science* ;
- Les profils IT.

Notons que chaque banque peut avoir des noms d'acteurs qui diffèrent, voire découper différemment ces rôles sur différents acteurs. Les descriptions suivantes sont donc à adapter à chaque cas.

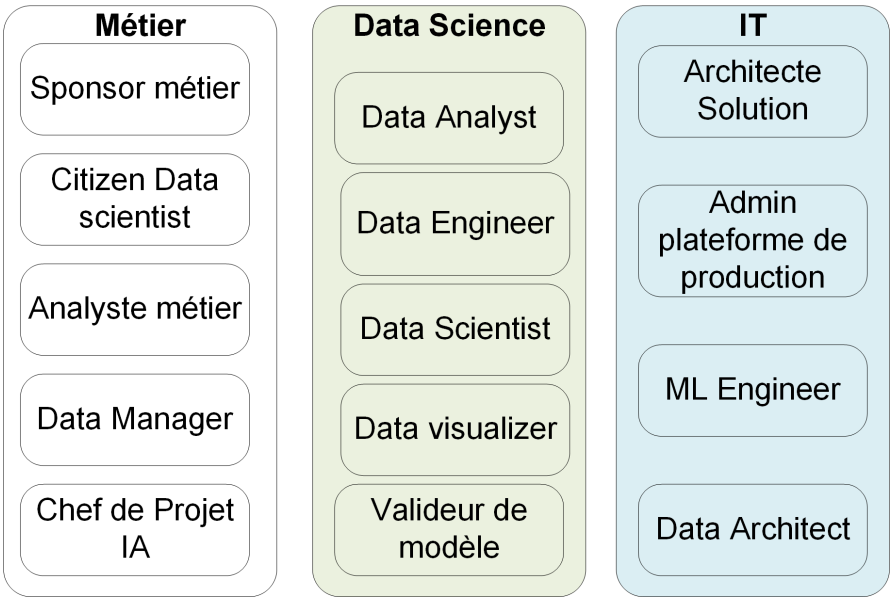

Figure 5 – Les profils d'acteurs

- **Profils « métier »**
  - **Sponsor métier** : l'entité ou la personne qui émet une expression de besoin pour un nouveau projet et qui est responsable de **l'évaluation métier** du projet en production ;
  - ***Citizen data scientist*** : il peut exécuter des **processus de production de modèles** avec des outils simples (*no-code*) pour démontrer/ explorer la valeur métier ; ce n'est pas un *data scientist,* mais un **opérationnel métier** ;
  - **Analyste métier** : il analyse les **besoins métier** et identifie les KPI métier (*key performance indicator,* ou indicateur de performance) ;
  - ***Data manager*** : Il assure **l'organisation et la gestion des données** sur son domaine de responsabilité. Il agit comme un référent pour les données. Selon les organisations, il peut faire partie du métier ou de l'IT ;

  - **Chef de projet IA** : il accompagne le projet dans **l'expression fonctionnelle des besoins**, pilote leur traduction en **réalisation technique** et leur exécution. Il s'assure de la conformité à la réglementation en matière de données (RGPD, etc.).

- **Profils *data science***
  - ***Data analyst*** : il **traite, exploite et analyse** les données ;
  - ***Data engineer*** : il **analyse les besoins en données**, il définit les processus de collecte et *monitoring* selon les modes d'interaction (*batch*, fil de l'eau) ;
  - ***Data scientist*** : **expert en data science**, il conçoit le pipeline de construction du modèle et produit / sélectionne le modèle ;
  - ***Data visualizer*** : il élabore des **visualisations des données** et des résultats à destination des *data scientists* et des spécialistes métier ;
  - **Valideur de modèle** : il **s'assure de la validité du modèle** et évalue ses conséquences non désirées éventuelles. Il assure une revue du modèle indépendante du développeur du modèle (*data scientist*).
- **Profils IT**
  - **Architecte Solution** : il propose une architecture de solution, comprenant les **contraintes de cybersécurité** ;
  - **Administrateur plateforme de production** : il est responsable de la **plateforme informatique** où s'exécutent les modèles. Il assure le traitement des alertes et les reprises sur incidents ;
  - ***ML engineer*** : il intègre les pipelines ML (c'est-à-dire la succession des étapes de développement du modèle) dans des processus MLOps (Machine Learning Operations);
  - ***Data architect*** : il conçoit les **infrastructures et les solutions Data**, il identifie les différentes sources de données.

Les **responsabilités** aux différentes étapes du cycle de vie de l'applicatif IA sont réparties comme suit :

- ***Model / Product Owner*** : personne qui prend la responsabilité du **modèle en production** (et donc de sa mise en production). Cela correspond à la personne côté métier qui va bénéficier de l'utilisation du modèle et qui peut également jouer le rôle de sponsor ;
- ***Model Developer*** (*data scientist*, *data analyst*, *data visualizer*) : personne ou groupe de personnes en charge du **développement du modèle** (équipe de modélisation) ;
- ***System Developer*** : personne ou groupe de personnes en charge du système IT qui va implémenter et ensuite permettre d'utiliser le modèle ;
- ***Model Validator*** (*data scientist* avec le rôle de validation) : personne en charge de la **revue indépendante** d'un modèle ;
- ***Model Monitor*** : personne en charge du **suivi d'un modèle** qui est en production. Cela peut inclure le suivi côté métier comme un suivi côté équipe de modélisation ;
- ***Model User*** : personne qui utilise le modèle dans son quotidien.

1.7 LES CONTROLES

Le **dispositif de contrôle** a pour but **l'encadrement et la maîtrise du risque**. Les contrôles visent à **identifier les risques, qualifier leurs impacts et évaluer leur matérialité**. Ils peuvent déboucher sur l'application de mesures de mitigation des risques.

Le **risque de modèle** est défini et strictement encadré par les superviseurs bancaires. Ceux-ci ont formulé des **exigences minimales** en matière de gestion du risque de modèle, qui s'appliquent à toute entité utilisant des modèles, en fonction du type d'utilisation.

On peut citer notamment les cadres suivants :

- (US) SR Letter 11-7, Supervisory Guidance on Model Risk Management[11], 4 April 2011, défini par la Réserve Fédérale (Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency), qui donne des **directives sur la gestion du risque de modèle** ;
- (Europe) La réglementation (EU) No. 575/2013 (CRR – *Capital Requirements Regulation*)[12] qui vise à améliorer la **transparence** sur les **risques encourus** par les institutions financières.

En 2019, un groupe d'experts de haut niveau sur l'intelligence artificielle mandaté par la Commission européenne, a publié les lignes directrices[13] **d'évaluation de l'éthique** d'une intelligence artificielle. Pour cela, le groupe a listé **sept critères** assurant une **IA digne de confiance**, ainsi que la **protection des droits fondamentaux** de chacun. Cette liste de critères[14] ainsi que les questions associées peuvent être utilisées afin d'effectuer un auto-diagnostic de son IA.

L'Autorité de Contrôle Prudentiel et de Résolution (ACPR) a publié un document de réflexion[15] sur la **gouvernance des algorithmes d'intelligence artificielle** dans le secteur financier. Il propose **quatre principes d'évaluation** des algorithmes d'IA basés sur :

- Le traitement adéquat des données d'entrée ;
- La performance des algorithmes ;
- La stabilité de la pertinence du modèle dans le temps ;
- Les différents degrés d'explicabilité en fonction des parties prenantes.

La Commission européenne a proposé en avril 2021 un **règlement établissant des règles d'harmonisation** dans le domaine de l'intelligence artificielle. Ce règlement a une vocation transverse, au-delà des établissements financiers. Il[16] classe les applications d'intelligence artificielle en fonction de leurs risques et les réglemente en conséquence. Les applications à faible risque n'y sont pas réglementées. Les IA à risque moyen et élevé nécessiteraient une auto-évaluation obligatoire avant d'être mis sur le marché, et certaines applications critiques nécessiteraient une évaluation indépendante par des tiers. La proposition viserait par ailleurs à interdire certains types d'IA (e.g. la surveillance biométrique de masse ou la notation sociale). Ce règlement introduit donc un **cadre de contrôle très général**, dépendant du niveau de risque de

[11] https://www.federalreserve.gov/supervisionreg/srletters/sr1107a1.pdf
[12] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32013R0575
[13] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai
[14] Critères d'auto-évaluation d'une IA : Intervention et supervision humaine ; robustesse et sécurité techniques ; protection et gouvernance des données ; transparence ; diversité, non-discrimination et équité ; bien-être sociétal et environnemental ; responsabilité.
[15] https://acpr.banque-france.fr/gouvernance-des-algorithmes-dintelligence-artificielle-dans-le-secteur-financier
[16] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

chaque IA. Le présent document ne se substitue pas à ce cadre de contrôle mais propose des dispositifs de maîtrise de certains risques spécifiques à l'IA.

Notons par ailleurs que plusieurs définitions de l'IA coexistent au niveau des instances européennes. Le Joint Research Center de la Commission européenne propose une définition basée sur une **approche taxonomique**[17], alors que le rapport[18] adopté par le Comité AIDA en préparation de la délibération de mai 2022 du Parlement Européen adopte **une définition plus globalisante** mais au périmètre moins sûr au plan juridique.

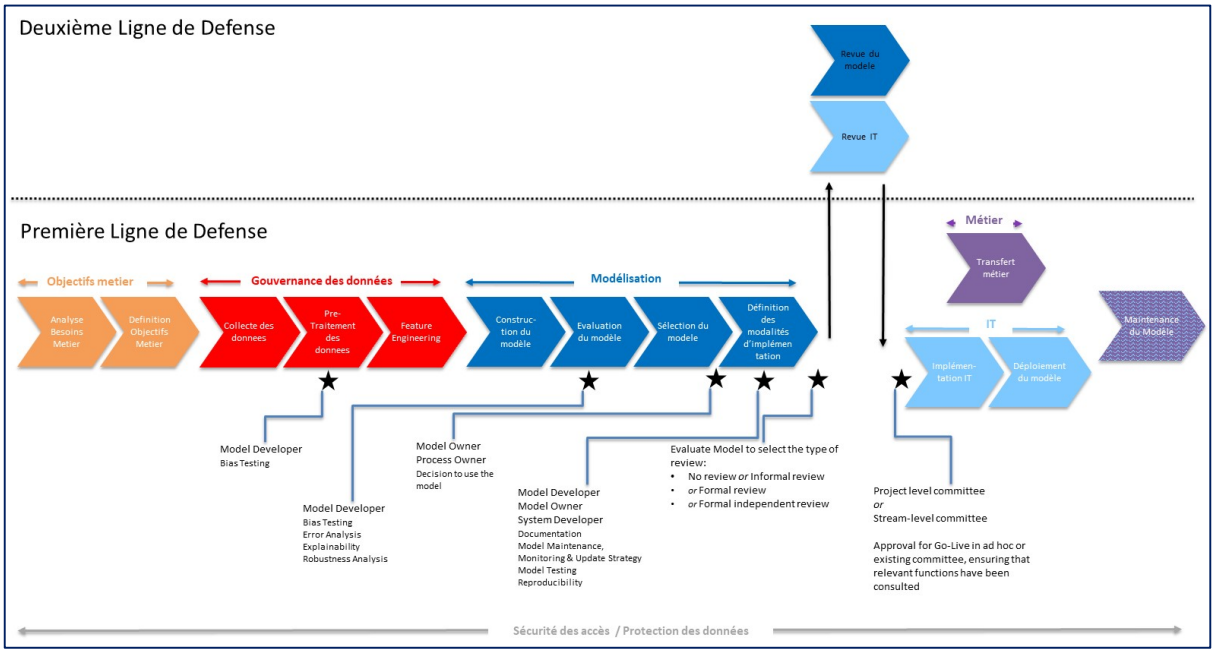

Figure 6 – Processus IA et lignes de défense

Dans le schéma ci-dessus (Figure 6), nous avons souligné le rôle des divers acteurs dans le processus de validation des modèles d'intelligence artificielle. Nous présentons les éléments importants de ce processus au niveau de la **première ligne de défense** (partie inférieure du diagramme) ainsi qu'au niveau de la **seconde ligne de défense** (partie supérieure du diagramme). La seconde ligne de défense n'est pas systématiquement un préalable à la mise en production.

Comme c'est déjà le cas pour la gestion des autres risques d'une institution financière, et avec des ajustements d'une banque à l'autre, l'organisation de la gestion des risques des modèles et donc des risques qui émanent de l'utilisation de l'intelligence artificielle, est organisée selon **trois lignes de défense**.

- La **première ligne de défense** LoD1 est représentée par les « ***Model Owners*** ». Ce sont les personnes qui prennent la responsabilité de l'utilisation des modèles et sont donc les premiers à devoir s'assurer que les risques afférents au modèle ont bien été pris en compte lors des diverses étapes du développement, déploiement et de l'utilisation du modèle. Ces personnes s'assurent que le processus de développement, de la documentation, et du

[17] Nativi, S. and De Nigris, S. AI Standardisation Landscape: state of play and link to the EC proposal for an AI regulatory framework, EUR 30772 EN, Publications Office of the European Union, Luxembourg, 2021, ISBN 978-92-76-40325-8, JRC125952. 2021. https://data.europa.eu/doi/10.2760/376602

[18] Report on artificial intelligence in a digital age. Special Committee on Artificial Intelligence in a Digital Age. 2020/2266(INI), April 2022. https://www.europarl.europa.eu/cmsdata/246872/A9-0088_2022_EN.pdf

suivi (*monitoring*) du modèle sont **conformes aux standards** de l'institution financière associée. Selon le type d'utilisation du modèle, il peut exister des procédures spécifiques qui stipulent plus précisément les actions et contrôles à mener. Ces gouvernances spécifiques ne sont pas nécessairement organisées par type de modélisation, mais plutôt **par usage** (par exemple Crédit, Assurance). Dans le cadre de l'IA, sur le schéma ci-dessus, les étoiles illustrent pour chaque étape les rôles et les aspects à vérifier et à documenter.

Notons que la toute nouvelle norme ISO Norme ISO/IEC 38507 : 2022[19] d'avril 2022 traite des implications de gouvernance de l'utilisation par des organisations de l'intelligence artificielle.

- La **deuxième ligne de défense** LoD2 est constituée par les équipes de **revue indépendante**, les équipes en **charge de la gouvernance** et de la **supervision du portefeuille des modèles** et si pertinent, des personnes participant aux comités d'approbation et de revue des modèles. Leur rôle est de collectivement s'assurer que la première ligne de défense joue bien son rôle de gestion des risques de modèle, mais aussi de mesurer et de « reporter » les risques de modèle agrégés au niveau d'un périmètre défini. Sur le schéma ci-dessus, la deuxième ligne de défense joue notamment un **rôle important de revue du modèle** avant la phase d'industrialisation. Elle doit intervenir au niveau de tous les risques identifiés et suffisamment matériels pour le justifier, que ce soit au niveau de la gouvernance des données, de la gouvernance des modèles et de l'implémentation IT, et tout au long de la vie du modèle.

  Ceci est d'autant plus vrai dans un **cadre d'industrialisation** des cas d'usages où les analyses de risques doivent être le plus possible insérées dans le **processus de construction** des modèles plutôt qu'intervenir en aval, créant potentiellement un goulet d'étranglement. Dans le cadre de modèle d'IA, il est fréquent que cette revue s'accompagne également **d'une revue de la solution IT** qui va intégrer le modèle dans le but par exemple de s'assurer de **la protection et sécurité des données** ou de la **robustesse du système** par rapport à des attaques cyber. Ces revues ne sont pas nécessairement menées par les mêmes équipes. Enfin, la fonction de gestion des risques devrait tenir informé l'organe de direction de l'établissement bancaire des hypothèses utilisées dans les modèles, de l'analyse des risques ainsi que leurs éventuelles lacunes.

- La **troisième ligne de défense** LoD3 est constituée des **équipes d'audit interne** (parfois dénommée inspection générale). Leur rôle est[20] **d'évaluer la conformité des opérations**, du niveau de risque effectivement encouru, du respect des procédures, de l'efficacité et du caractère approprié des dispositifs d'identification et de gestion des risques. Elles peuvent donc être amenées à vérifier que les travaux réalisés aux deux premiers niveaux sont **conformes aux règles** en vigueur au sein de l'établissement, ce qui implique de procéder à une **réévaluation de modèles** développés, voire dans certains cas, de challenger les contrôles réalisés au moyen de modèles alternatifs. La troisième ligne de défense évalue également le **dispositif de gestion du risque de modèle**. Selon les orientations de l'European Banking Authority (EBA) en matière de gouvernance interne[21], la fonction d'audit interne devrait en particulier vérifier **l'intégrité des processus** garantissant la fiabilité des méthodes

[19] ISO/IEC 38507 : 2022. Gouvernance des technologies de l'information — Implications de gouvernance de l'utilisation par des organisations de l'intelligence artificielle ; Avril 2022. https://www.iso.org/fr/standard/56641.html
[20] Arrêté du 3 novembre 2014 relatif au contrôle interne des entreprises du secteur de la banque, des services de paiement et des services d'investissement soumises au contrôle de l'Autorité de contrôle prudentiel et de résolution
[21] Rapport final sur les orientations en matière de gouvernance interne, EBA/GL/2021/05 2 juillet 2021. https://acpr.banque-france.fr/sites/default/files/media/2021/12/07/20211207_orientations_eba-gl-2021-05.pdf

et techniques de l'établissement, ainsi que des hypothèses et des sources d'information utilisées pour ses modèles internes. Elle devrait également **évaluer la qualité et l'utilisation des outils qualitatifs de détection et d'évaluation des risques,** et les mesures prises pour atténuer les risques. L'IA entre pleinement dans ce cadre général.

2 IDENTIFICATION DES RISQUES

Afin de mettre à profit notre expérience collective de l'IA dans la banque, chacune des trois banques participant au groupe de travail a commencé par établir un **inventaire des risques** encourus à chaque étape du processus de développement et de déploiement d'un modèle d'IA. Au-delà de la description du risque, nous avons cherché à documenter également qui était **partie prenante** (Métier, *data scientist*, Utilisateur, IT ou plus généralement l'établissement) et quel était le **type d'impact de ce risque** (i.e. efficacité opérationnelle, conséquences financières, risque de réputation, risque règlementaire). Nous avons également indiqué pour chaque risque les mesures de contrôle qui pouvaient être mises en place sans en chercher nécessairement l'exhaustivité.

Une fois ces informations recueillies, nous avons collectivement revu les propositions respectives afin d'identifier les doublons, et de repréciser ou nuancer certains risques. Cette **méthodologie itérative** nous a permis d'aboutir à une **vision commune des risques** que nous décrivons dans les sections suivantes.

Au cours de cette revue commune, nous avons essayé de nous attacher aux risques qui sont soit spécifiques pour les systèmes IA, soit présentant un risque accru. Par exemple, la sécurité des systèmes d'information ou la gouvernance des données sont deux problématiques qui sont généralement traitées globalement par l'entreprise et englobe les systèmes IA.

Parmi les trois banques participantes, des représentants des trois lignes de défense ont participé à cet exercice de revue des risques, couvrant ainsi le processus de bout en bout.

Conformément au processus présenté dans la Figure 6, nous avons analysé les **5 étapes successives** : objectifs métier, gouvernance des données, modélisation, IT et transfert au métier. Pour chaque étape, nous avons étudié les sous-tâches et les risques associés. Les éléments de contrôle seront approfondis dans la suite du document (section 3).

2.1 OBJECTIFS METIER

2.1.1 ANALYSE DES BESOINS METIERS

Les systèmes d'IA permettent de **répondre à des besoins spécifiques** de certains métiers. La première étape consistant à bien définir ces objectifs est aussi le **premier facteur de risque**. Ce n'est pas un risque propre à l'IA, mais il est potentiellement plus élevé pour plusieurs raisons qu'il convient d'identifier pour en limiter l'impact.

2.1.2 COMPREHENSION DE L'IA

Une des causes principales est intrinsèquement liée à une mauvaise compréhension de ce qu'est l'IA, notamment sur ce qu'il est réellement possible de faire, sur la **fiabilité des résultats** ou leur **pertinence**. En phase de conception d'un système d'IA, les équipes impactées par ce risque sont les métiers et les *data scientists*. Le risque peut se matérialiser de plusieurs manières,

avec en premier un objectif métier non atteint, ou encore une performance financière inférieure à l'attendu.

2.1.3 CHOIX DE LA VARIABLE CIBLE

Un système d'IA a le plus souvent comme objectif de **produire un résultat, appelé variable cible** (il peut y en avoir plusieurs). Si cette variable cible est mal définie, par rapport au cas d'usage ou au contexte, le système d'IA peut produire des résultats non conformes au besoin. Les *data scientists* peuvent être impactés car la modélisation du système sera inadaptée, les métiers utilisant les résultats peuvent pâtir d'une mauvaise performance.

En complément des mesures relevant du point précédent, une attention toute particulière doit être portée à la définition de la variable cible :

- Description exhaustive et partagée de la cible, de sa mesure ;
- Contrôle de l'adéquation entre la variable cible et la réalité opérationnelle.

2.1.4 L'IA POUR OPTIMISER UN PROCESSUS

Outre le calcul de variables cibles ou la prévision par apprentissage, l'IA peut aussi être utilisée comme **moteur d'optimisation** (d'un portefeuille, d'un processus ou la détection de cas atypiques). Dans ce cas, la résultante du système d'IA doit être adaptée aux besoins métier. En particulier, la marge d'erreur issue du modèle d'IA est-elle compatible avec l'objectif ? La capacité à trouver un optimum correspond-elle à la cible ? La probabilité de ne pas trouver la solution est-elle acceptable d'un point de vue métier ?

Là encore l'impact du risque peut être financier ou sous optimal du point de vue du processus à optimiser, voire induire en erreur et donner une mauvaise solution à un problème. Quantifier le risque lié au modèle suppose de le faire par rapport à une finalité qui est définie par le métier et l'usage qu'il fait du modèle. La définition de l'usage et des limites d'emploi du modèle sont donc essentielles pour évaluer leur compatibilité avec les hypothèses de fonctionnement théorique et la calibration réalisée. Le risque du modèle est mesuré par l'écart potentiel à la réalité, pour le réfuter il faut donc disposer d'une réalité observable tangible définie par le métier sur des métriques pertinentes. Pour limiter les risques, un certain nombre de mesures peuvent être envisagées, telles que :

- La bonne définition du problème à optimiser et son périmètre ;
- L'analyse d'impact des erreurs théoriques de prévision ou d'optimisation ;
- La bonne appréhension par le métier du risque d'erreur et le choix de métriques « métier » pertinentes.

2.1.5 L'INADEQUATION DES BESOINS METIERS AVEC D'AUTRES OBLIGATIONS

Les objectifs du métier peuvent être légitimes, pour autant le recours à un système IA peut s'avérer contraire à la réglementation, à la déontologie ou aux valeurs de l'entreprise. Dans ce cas le risque dépasse le tandem métier – *data scientist* car il affecte potentiellement l'entreprise dans son ensemble (amende, réputation…).

Les causes racines sont multiples, pour la plupart non spécifiques à l'IA par exemple utilisation de données interdites ou encore ségrégation intempestive des individus. La plupart de ces risques font l'objet de dispositifs de maîtrise par ailleurs. Il convient néanmoins de :

- S'assurer que la mise en œuvre de système d'IA dispose bien du même niveau d'assurance que les autres projets (notamment sur des volets type RGPD, déontologie, processus nouveau produit, sécurité, etc.) ;
- Prendre en compte la **réglementation propre** à l'usage de l'IA (à date cette réglementation est encore en projet : cf. le projet Artificial Intelligence Act[16] de la Commission européenne) ;
- Spécifiquement pour les systèmes d'IA, tester les effets indésirables, en particulier les comportements indésirables qui ne se manifestent qu'avec l'usage en conditions réelles. Cela peut être par exemple la discrimination de certaines catégories de clients sur des critères apparents d'âge ou de domicile, quand bien même ceux-ci ne feraient pas explicitement partie des données de calibration du modèle ;
- Mettre en place une **gouvernance** pour évaluer si le senior management connaît les usages.

2.2 GOUVERNANCE DES DONNEES

2.2.1 COLLECTE DES DONNEES

Le **processus de collecte des données** en amont des modèles est essentiel pour la pertinence et la qualité des systèmes d'IA. La qualité des données en entrée d'un modèle est un sujet important mais qui n'est pas spécifique aux modèles basés sur l'IA puisqu'une qualité déficiente aura des impacts également sur des modèles statistiques classiques ou des systèmes de *reporting*. Par ailleurs, la qualité de la donnée fait déjà l'objet d'une **réglementation spécifique**, (Article 82 de la Directive 2009/138/CE du parlement européen et du conseil dans le cadre de Solvabilité II[22] pour les assurances par exemple, BCBS239 du Comité de Bâle sur le contrôle bancaire[23], Arrêté du 25 février 2021 modifiant l'arrêté du 3 novembre 2014 article 104[24] par exemple). Le sujet n'est donc pas spécifique aux modèles liés à l'IA.

Le Comité de Bâle définit en particulier **14 principes répartis selon 4 thèmes**. Les cinq premiers impliquent de mettre en place :

- Une **gouvernance** en lien avec les capacités d'agrégation des données de risque et ses pratiques de notification des risques ;
- Une **architecture des données** et **infrastructure informatique** permettant de renforcer ses capacités d'agrégation des données de risque et ses pratiques de notification des risques, non seulement en situation normale, mais aussi en période de tensions ou de crise ;
- Une **exigence d'exactitude et d'intégrité** des données de risque ;
- Une **exigence d'exhaustivité** des données de risque ;
- Une **exigence d'actualité** pour pouvoir produire rapidement, agréger et mettre à jour des données sur les risques.

Si la finalité de ces principes est de permettre une meilleure mesure des risques, elle induit une organisation au sein des établissements bancaires contribuant à une vision transverse de la qualité de la donnée, qu'elle soit liée aux reportings, aux modèles d'IA ou toute autre finalité métier.

[22] https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32009L0138&from=sv
[23] https://www.bis.org/publ/bcbs239.pdf
[24] https://www.circulaires.gouv.fr/loda/id/LEGIARTI000043224879/2021-06-28/

2.2.1.1 DISPONIBILITE DES DONNEES

La capacité à disposer des données essentielles aux modèles doit être évaluée. La **disponibilité des données utiles** se retrouve à deux niveaux : lors de la **phase de modélisation** (*build*) et en **exploitation** (*run*).

Lors de la phase de modélisation, les enjeux de disponibilité se posent notamment lors de l'apprentissage, quand des données essentielles ne sont pas disponibles, comme par exemple un historique tronqué ou incomplet pour des raisons techniques ou réglementaires. Il peut s'agir également d'une donnée dont la **durée de conservation** en base active est fixée par le responsable de la protection des données à trois ans, mais pour laquelle une durée d'au moins dix ans serait nécessaire afin de modéliser correctement des processus de moyen terme.

Lors de la phase d'exploitation, certaines données peuvent être manquantes, invalides ou disponibles avec retard par rapport à l'usage du modèle.

Ces risques peuvent être limités par une analyse des données en entrée lors de la modélisation et par un *monitoring* en continu des données lors de la phase d'exploitation. L'évaluation du risque doit prendre en compte ces dispositifs éventuels.

2.2.1.2 DONNEES EXTERNES

L'entité responsable d'un modèle ne maîtrise pas nécessairement la qualité de ces données. Les facteurs de risque associés proviennent d'un **éventuel manque de transparence** sur la constitution et la définition des données, les règles de mesure, d'agrégation ou de calcul utilisées. Contrairement aux données internes où les producteurs de données sont en général garants de la qualité, il incombe ici au *data scientist* ou au *Model Owner* de s'assurer de la **qualité** des données externes utilisées.

Le risque est présent lors de la **calibration des modèles** mais aussi en **phase d'exploitation**.

2.2.1.3 UTILISATION DE DONNEES PERSONNELLES SANS AUTORISATION

Les systèmes d'informations bancaires comportent de nombreuses **données personnelles ou sensibles**. Comme prévu en l'application de la réglementation sur la protection des données (RGPD), le recueil du consentement à l'exploitation et au partage de certaines données est essentiel dans certains cas. Il faut aussi dans ce cadre envisager les usages du modèle car le consentement se fait pour des finalités explicites lors de l'accord. Un autre facteur de risque réside dans **l'utilisation de données historiques**, en veillant à ne pas exploiter de données ayant dépassé leur durée d'archivage réglementaire ou sujettes au droit à l'oubli. La durée de détention est un risque à analyser spécifiquement, avec des risques de conflit entre des réglementations globales et locales.

Les données peuvent concerner des variables individuelles mais aussi des cookies et autres traceurs par exemple.

Ces risques doivent être adressés en amont de la constitution des jeux de données et s'appuyer sur une gouvernance de la donnée stricte. Il faut en la matière contrôler la bonne information des *data scientists* et les fonctions en charge de la gouvernance des données personnelles à propos des usages prévus.

2.2.1.4 ANONYMISATION DES DONNEES

La norme ISO/IEC 29100 : 2011[25], relative à la protection de données personnelles identifiables (*personally identifiable information*), définit l'**anonymisation** comme un procédé par lequel les données personnelles identifiables sont transformées de manière **irréversible**, de sorte que les personnes concernées ne soient plus identifiables [26]. Par contre, la **pseudonymisation** permet de rétablir une identification par le biais d'informations supplémentaires.

L'utilisation de données anonymisées[27] pour la calibration des modèles est fortement recommandée, sauf lorsque cela détruit la pertinence du *data set*, ou lorsque l'anonymisation est incompatible avec le but recherché, tel que le traitement de données granulaires (détaillées) et non agrégées. Dans ce cas, des techniques de pseudonymisation doivent *a minima* être appliquées, en s'assurant de l'adéquation entre le consentement des personnes et la finalité visée. Il est ainsi nécessaire d'évaluer le **risque de ré-identification** et la compatibilité des techniques utilisées avec les réglementations applicables en la matière, par exemple la réglementation sur la protection des données (RGPD).

2.2.1.5 PROTECTION DES DONNEES

La **protection des données** fait partie des politiques de sécurité des systèmes d'information des établissements. Elle est naturellement renforcée pour les données personnelles ou sensibles. Les conditions de développement et de mise en œuvre de systèmes d'IA peuvent laisser des **brèches de sécurité**, à tous les niveaux : accès aux modèles, accès aux moteurs de calibration, utilisation de composants externes ou ouverts, exposition des actifs à l'extérieur d'un environnement sécurisé, etc.

La modélisation et le déploiement de son usage doivent également se faire en accord avec les politiques de sécurité des systèmes d'information.

2.2.2 PRE-TRAITEMENT DES DONNEES

2.2.2.1 DONNEES INCOMPLETES OU BIAISEES

La nature même des données, leur mesure ou leur codage peut biaiser les informations, ce qui peut fausser les résultats lorsqu'il s'agit d'une variable cible ou par exemple contribuer à une mauvaise classification lorsqu'il s'agit d'une variable en entrée.

Ainsi, les **données en entrée du modèle** sont à évaluer soigneusement lors des phases de modélisation et d'inférence. Elles peuvent être parcellaires et induire des **biais de calibration**, notamment à cause d'un mauvais échantillonnage. Elles peuvent **manquer de représentativité**, soit en profondeur d'historique, soit en proportion de certaines catégories. Les données historiques peuvent refléter des décisions de précédents modèles et contenir des biais. Enfin, elles peuvent aussi provenir d'un pré-traitement (par exemple de catégorisation, labellisation ou normalisation) erroné.

---

[25] ISO - ISO/IEC 29100:2011 - Technologies de l'information — Techniques de sécurité — Cadre privé . https://www.iso.org/fr/standard/45123.html

[26] Avis 05/2014 sur les Techniques d'anonymisation. 0829/14/FR WP216, Groupe de travail « Article 29 », chapitre 2.2). https://www.cnil.fr/sites/default/files/atoms/files/wp216_fr.pdf

[27] L'anonymisation de données personnelles, CNIL, 19 mai 2020. https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

2.2.3 FEATURE ENGINEERING ET FEATURE SELECTION

Dans le processus de modélisation, une fois que les données ont été identifiées, partagées, protégées et analysées en termes de qualité de données, l'étape suivante consiste à s'appuyer sur ces données en entrée pour définir des données supplémentaires qui s'appuient sur ces données brutes (***feature engineering***) mais aussi, dans un second temps de sélectionner parmi un large choix de candidats quelles données vont être utilisées dans le modèle (***feature selection***). Il s'agit donc de deux étapes de manipulation des données qui peuvent comporter des risques afférents.

Ces étapes ne sont pas spécifiques à la modélisation par du *Machine Learning*. Ce qui est peut-être plus spécifique à l'IA est la capacité des modèles d'ingérer un grand nombre de variables **d'entrée**, ce qui impose une **vraie discipline de modélisation** si l'on souhaite maîtriser les risques.

2.2.3.1 UTILISATION DES DONNEES SENSIBLES

Dans le choix des données utilisées dans un modèle, le risque le plus évident consiste à utiliser des données pouvant **induire un biais ou une discrimination**, comme le genre ou l'âge. Une liste de données dites sensibles est indiquée dans la réglementation RGPD. Néanmoins, les *Model Owner* et *Model Developer* se doivent de connaître les données qui ne sont pas tolérées comme variables d'entrées directes de modélisation. Il y a certains cas spécifiques, par exemple dans l'assurance, où l'utilisation de données comme l'âge sont acceptées alors qu'elles ne le sont pas en général. L'utilisation de données sensibles présente un **risque réglementaire** ainsi qu'un **risque de réputation**.

2.2.3.2 MAUVAISE INTERPRETATION DES DONNEES

La fragmentation de la gestion de la donnée dans les entreprises peut également mener à une situation ou l'équipe de modélisation a une **mauvaise compréhension ou interprétation des données manipulées**. En effet, les personnes en charge de fournir les données à l'équipe de modélisation (IT) sont souvent séparées des personnes ayant la connaissance de la donnée (typiquement les *Data Managers, Model Owner* et *Model User*). Ainsi l'équipe de modélisation peut se retrouver à utiliser comme variable cible une variable appelée « Défaut » mais qui en fait ne correspond pas à la variable défaut réelle qu'il faudrait modéliser et qui se nomme peut-être « D_reel ». La conséquence est que la modélisation se fait sur des bases approximatives et mène donc à des **résultats potentiellement erronés**.

2.2.3.3 PROFUSION DE VARIABLES EXPLICATIVES

Grace à l'accès plus simplifié à la donnée, grâce aux capacités de calculs qui permettent l'apprentissage de modèles avec un très grand nombre de paramètres, le nombre de variables explicatives en entrée d'un modèle n'est plus un facteur qui limite la modélisation. Cela permet de travailler sur des **modèles plus riches et plus nuancés,** mais la profusion de données en entrée est aussi associée à des **risques** tels que : (i) l'augmentation du nombre de sources de données qui peut avoir un effet négatif important en **augmentant la dette technique** et (ii) **l'accroissement des vulnérabilités potentielles** du modèle en cas de défaillance d'une des sources.

2.2.3.4 FEED-BACK LOOPS

Dans le cadre de l'apprentissage (en continu ou régulier) des modèles supervisés, **l'utilisation du modèle en production** génère automatiquement des données supplémentaires qui permettent de réentraîner un modèle rapidement afin de lui permettre de se mettre à jour au fil de l'eau. L'apprentissage en continu génère des interactions entre les données et le modèle qui peuvent entraîner des **conséquences négatives**.

Nous illustrons trois phénomènes de feed-back :

- Dans le cas où l'utilisateur ne donne son feed-back sur la proposition du modèle que si la prédiction est par exemple positive (comme dans les vérifications de suspicion de fraude), il est évident que le **feed-back sera biaisé** et peut mener à une dégradation de la performance du modèle, notamment pour les faux négatifs. Ceci s'apparente à **une forme de biais de sélection** ;
- Dans le cadre des systèmes de recommandation, par exemple les recommandations de films sur les plateformes de vidéo à la demande, le modèle influence directement la sélection de ses propres futures données. Ce cas de figure n'est pas limité aux systèmes de recommandation, un modèle de prédiction des prix de l'immobilier par exemple peut aussi engendrer ce phénomène si les acteurs économiques utilisent le modèle comme base de leur choix de prix. Dans ces cas-là, la performance observée du modèle tend à s'améliorer mais c'est un effet de **biais de sélection**. Le risque est donc d'enfermer le modèle dans un **phénomène de bulle** ;
- Dans des cas, moins fréquents mais possibles, où deux systèmes d'IA interagissent, il faut être attentif à l'interaction entre ces deux modèles. On peut citer par exemple le cas des **modèles en cascade**, où la sortie d'un modèle est utilisée en entrée d'un autre modèle. Ce cas peut par exemple, survenir lorsqu'un premier modèle est utilisé pour créer des scenarii d'alertes en protection contre la fuite de données et qu'un second modèle permet de disqualifier certaines alertes afin de réduire les faux positifs. Il faut avoir conscience que modifier l'un des deux systèmes peut rendre déficient le second.

2.3 MODELISATION

2.3.1 CONSTRUCTION ET SELECTION DU MODELE

La **phase de construction** d'un système d'IA présente différents points d'attention. Les **hyperparamètres du modèle** doivent être calibrés en évitant les écueils méthodologiques. Il est également essentiel de surveiller le renforcement de potentiels biais en sortie du modèle. Les indicateurs de performance doivent être alignés avec l'objectif initial. Enfin, la **documentation** joue un rôle clé pour assurer la justification et la transparence des choix de modélisation.

2.3.1.1 CALIBRAGE DES HYPERPARAMETRES

Dans un système d'IA, les hyperparamètres peuvent provenir des algorithmes d'optimisation (par exemple, le *learning rate*), de problèmes non convexes / non différentiables (par exemple, le choix d'une fonction d'activation, de l'architecture d'un réseau de neurones) ou de considérations statistiques (par exemple, les paramètres d'un *kernel*). Il est important de pouvoir les identifier et d'évaluer leur potentiel impact sur les performances et la robustesse du système d'IA.

A la différence des paramètres du modèle (par exemple, les poids dans les réseaux de neurones ou les coefficients d'une régression linéaire) qui sont appris via une procédure d'optimisation d'une fonction objectif en lien avec le cas d'usage, les hyperparamètres sont initialisés en utilisant une connaissance *a priori*. Ils nécessitent la mise en place **d'un processus itératif d'exploration** de l'espace des hyperparamètres. Ce processus doit se faire sur un **jeu de données de validation**, différent des données d'entraînement utilisées pour l'apprentissage des paramètres du modèle, et différent également des données de test utilisées pour obtenir une **mesure non biaisée** de la performance.

Le processus de calibration des hyperparamètres introduit indirectement des informations du jeu de données de validation dans les paramètres du modèle, que l'on appelle le *target leakage*. Pour la calibration des hyperparamètres, il faut diviser les données de modélisation en deux parties : **entraînement et validation**. Les données de validation indépendantes servent donc à l'optimisation des hyperparamètres sans risque de *target leakage*. Cependant, l'utilisation de la technique de validation croisée permet aussi de traiter ce problème tout en conservant un échantillon de données plus important pour l'entraînement final. La validation croisée « *k-fold* » partage aléatoirement les données en k groupes d'échantillons. Au cours de k itérations, le modèle est entraîné sur la collection de k-1 groupes et validé sur le groupe restant. Le groupe restant est différent à chaque itération. La performance finale est la moyenne des performances observées sur le groupe restant lors des k itérations. Le modèle final peut ainsi être ré-entraîné sur toutes les données avec les hyperparamètres présentant les meilleurs performances moyennes.

Les techniques usuelles de recherche des hyperparamètres optimaux (*grid search*, *random search*, *hyperband* et optimisation bayésienne) sont **coûteuses en temps de calcul**. Le *data scientist* doit souvent faire un compromis entre temps de calcul et exploration de l'espace des combinaisons d'hyperparamètres possibles. Une bonne pratique consiste à utiliser des « *learning curves* » pour traquer les améliorations de performances en fonction des hyperparamètres choisis. En effet, pour les hyperparamètres numériques et critiques comme le nombre d'itérations (e.g. réseau de neurones, *gradient boosting*) ou le nombre d'estimateurs (e.g. *random forest*), il est conseillé d'afficher les deux courbes d'erreur suivant le nombre d'itérations pour l'entraînement et la validation. En augmentant le nombre d'itérations, l'erreur continuera à se réduire sur les données d'entraînement. Pour éviter le sur-apprentissage et choisir le nombre pertinent d'itérations, il suffit de se placer au moment où la courbe d'erreur en validation commence à augmenter (figure 3). Cette technique visuelle permet de sélectionner facilement l'hyperparamètre considéré. Les *learning curves* sont également utiles pour le choix des *learning rates* des algorithmes d'optimisation et la visualisation des vitesses de convergence de ces derniers.

2.3.1.2 CREATION OU AMPLIFICATION DES BIAIS

Certains choix de modélisation, visant à optimiser les performances, peuvent **créer des biais** ou potentiellement **amplifier des biais** préexistants dans les données brutes.

Le choix de l'algorithme de *Machine Learning* peut être lui-même un facteur de biais en privilégiant un certain type de variable par rapport à un autre. Par exemple, un modèle de type *Random Forest* aura tendance à accorder plus d'importance aux variables continues ou aux

variables catégorielles à forte cardinalité[28]. Certains systèmes d'IA peuvent, par construction, s'appuyer sur des variables latentes (synthétisant plusieurs variables explicatives), et sont ainsi susceptibles d'identifier et d'exploiter des sous-groupes de la population, sans que cela soit explicite.

Lorsque l'on choisit un modèle, il faut donc ne pas utiliser exclusivement la performance, mais aussi analyser les biais potentiellement créés et choisir en fonction du compromis entre les deux critères.

2.3.2 EVALUATION DU MODELE ET INADEQUATION DES KPI

L'évaluation de la performance est une étape essentielle avant le déploiement du modèle et tout au long de son cycle de vie. Cela permet d'éviter d'implémenter ou de maintenir en production un modèle sous-performant, écartant ainsi la matérialisation du risque de modèle.

Pour cela, une attention particulière doit être portée sur le **choix de l'indicateur de performance**, qui dépend du type de modèle (e.g. classification ou score) et de l'usage prévu du modèle. L'indicateur choisi ainsi que le seuil associé doivent être compris par le métier afin de pouvoir approuver le modèle de manière éclairée, comme précédemment mentionné dans la section sur les besoins métier.

De plus, les indicateurs de performance sont multiples et peuvent donner lieu à des **interprétations** différentes pour un même modèle.

Par exemple, dans un problème de classification (positif / négatif), les KPI de performance font souvent référence à la **matrice de confusion**[29]. L'exactitude (ou *accuracy*) de la classification peut être excellente, si une classe est prépondérante, mais ne pas refléter correctement la performance du modèle. D'autres métriques extraites de la matrice de confusion telles que la précision, le rappel (ou *recall*), la spécificité ou le F1-score, sont à étudier en fonction de la nature du problème à résoudre et du besoin métier (par exemple, minimiser les faux positifs, les faux négatifs ou les deux).

Enfin, il est important de mesurer la performance du modèle sur un échantillon indépendant de celui qui a été utilisé pour la construction du modèle afin d'identifier les **problèmes de surapprentissage**. Dans ce cas, le modèle présente de bonnes performances sur l'échantillon d'entraînement, mais une baisse significative apparaît sur l'échantillon de validation/test. Par ailleurs, l'échantillon de test doit être représentatif du périmètre d'application du modèle afin de fournir une **vision réaliste** de la performance du modèle après déploiement. Ensuite, en phase de déploiement, les données doivent respecter la distribution des données d'apprentissage. L'échantillon test doit être représentatif du périmètre d'application du modèle, dans le cas contraire il s'agit d'une **dérive de la distribution**. Quand cela cesse d'être le cas, ce que l'on pourra détecter par une baisse des KPI métier mesurés (on parle de dérive ou de *drift*), il faudra envisager une **phase de réentraînement**.

[29] La matrice de confusion est le résumé des résultats de la classification, elle compare données observées et données prédites et indique le nombre d'observations bien prédites selon la classe (vrais (TP) et faux (FP) positifs / vrais (TN) et faux (FN) négatifs). De cette matrice, peuvent être extraites différentes métriques de performance, telles que la précision (TP/TP+FP), le rappel (*rappel* ou sensibilité TP / TP+FN), le F1 score (2 x *rappel* x *précision* / *[rappel + précision]*).

2.3.3 ABSENCE DE PISTE D'AUDIT (DOCUMENTATION, LISTE DES LIBRAIRIES UTILISEES)

Une documentation imprécise du modèle peut questionner sa raison d'être et la **fiabilité du mécanisme de décision**. Dans un souci de transparence des choix de modélisation, la documentation joue donc un **rôle clé de justification**. Les hypothèses émises et les choix de modélisation doivent s'appuyer sur des **preuves théoriques, métier et expérimentales**. Il est donc recommandé de former la première ligne de défense sur les éléments requis dans la documentation modèle. Il est notamment clé de décrire l'objectif et le périmètre d'utilisation, le cadre réglementaire, les hypothèses émises, le support théorique et métier des choix de modélisation, les facteurs du modèle, les choix des métriques de performance, les conditions et résultats expérimentaux, et les limites du modèle. Par ailleurs, le pipeline de transformation des données doit être documenté afin d'assurer son **auditabilité et sa réplicabilité**, notamment afin de **réduire le risque opérationnel** dans un scénario de départ des personnes ayant développé le modèle. Enfin, la documentation du modèle doit être maintenue tout le long de son cycle de vie.

Un décalage entre la documentation du modèle et le code est un risque majeur. En effet, il est possible de questionner ce qui sera finalement implémenté. Les codes de développement et d'inférence doivent donc être documentés et alignés à la documentation du modèle. Ceci favorisera notamment sa maintenance. Afin d'assurer la traçabilité des changements, le ***versioning*** du code doit être correctement appliqué. Les données ayant servi à la conception du modèle doivent être archivées dans la mesure du possible. Par ailleurs, la version des librairies doit être mentionnée. Étant donné la **nature inductive** de l'IA, il est clé de préciser les états aléatoires utilisés dans un **souci de réplicabilité**. Enfin, en cas d'utilisation de librairies tierces, le risque d'outrepasser les termes et conditions d'utilisation est important. Il est donc essentiel de vérifier au cours du temps que l'usage reste conforme à la licence.

2.4 IT

Après l'étape de construction d'un modèle, et comme toute application et service IT, un modèle doit être déployé et intégré dans l'écosystème IT de production existant afin que celui-ci puisse être utilisé par les métiers, intégré à une application existante ou encore en tant que service utilisable par plusieurs métiers ou applications. Il est courant de voir un modèle développé sur une plate-forme dite de ***Data Science***. Le modèle est ensuite pris en charge par des équipes informatiques afin de le mettre en production. Cette dichotomie entre les deux activités peut engendrer un certain nombre de problèmes et d'inefficacités.

2.4.1 DEPLOIEMENT DU MODELE

Comme évoqué ci-dessus, la finalité d'un modèle est de **fournir un service** (une prédiction) et d'être utilisé au sein d'un ou plusieurs processus métier. Le déploiement d'un modèle consiste principalement en sa **préparation** pour y être inséré dans un environnement IT de production. Les activités couramment rencontrées sont :

- **Préparation de l'environnement cible** dans lequel le modèle va être utilisé (exemple : environnement de l'application qui va intégrer le modèle, environnement type *docker*, etc.) ;
- **Automatisation des pipelines**, en particulier relatifs aux données (exemple : pipeline de préparation et modification des données avant d'être utilisées par le modèle) ;
- ***Versioning*** du modèle, des métadonnées associées ;

- Dans certains cas, le modèle peut être **recodé** pour des problématiques de langage, de performance ou d'environnements.

Après cette phase de préparation, le déploiement peut être réalisé, le plus souvent par les équipes IT responsable de la production. Plusieurs stratégies (définies préférablement en amont et en accord avec les responsables métier) peuvent être suivies, comme le remplacement d'un modèle existant, ou l'utilisation de méthodes telles que :

- ***Shadow mode*** : le nouveau système d'IA est déployé en parallèle du processus actuel mais sa sortie n'est pas utilisée pour la production. Ceci permet d'évaluer la **stabilité et la robustesse** des performances du nouveau système, mais requiert des ressources IT pour faire fonctionner les deux processus en même temps.
- ***Canary*** : le déploiement est fait de manière graduelle sur un nombre d'utilisateurs ou d'opérations de plus en plus important. Dans cette stratégie, certains utilisateurs testent ainsi le nouveau système en conditions réelles.
- ***A/B Testing*** : les utilisateurs sont divisés aléatoirement en deux groupes A et B qui utiliseront des versions différentes du système d'IA en production. Ceci permet d'évaluer empiriquement sur des métriques métier quel système est le plus performant dans des conditions réelles.

Cette liste de stratégies de déploiement n'est pas exhaustive. Il appartient donc au donneur d'ordres (*Model Owner*) de choisir la **stratégie de déploiement proportionnée et adaptée** au contexte du risque du modèle.

Le déploiement d'un modèle IA n'est pas uniquement du ressort de l'IT. Il doit être préparé le plus en amont possible car plusieurs problématiques peuvent se poser :

- Au moment de la préparation du modèle pour l'environnement cible de production, il convient de s'assurer que l'environnement de production sera suffisamment bien dim**ensionné** pour que le modèle ait un temps de réponse satisfaisant au regard du nombre d'utilisateurs et « d'appels » au modèle.
- Le déploiement du modèle doit suivre les **règles de sécurité et les standards de mise en production** de l'entreprise sous peine de compromettre la stabilité et/ou la sécurité de l'environnement de production. Certains outils de *Data Science* permettent de faire des « mises en production » rapides réalisées par le *data scientist*. En fonction des entreprises et des environnements, ceci peut être permis selon certains critères (par exemple, démonstration ou encore POC). Dans un environnement de production sécurisé, le déploiement d'un modèle doit suivre les standards spécifiques de l'entreprise.
- En relation avec les standards de mise en production et de déploiement d'une entreprise, il convient de s'assurer au travers d'un **process et de contrôles** que les modèles qui sont déployés correspondent bien au modèle qui a été développé, testé et validé par le métier, sans modification majeure.

2.4.1.1 UTILISATION DU CLOUD

L'utilisation du *cloud* comme plate-forme de développement et de déploiement de modèles d'intelligence artificielle est fréquente, notamment car les fournisseurs de *cloud* proposent des outils qui rendent les tâches d'entraînement et déploiements accessibles aux *data scientists*.

Lorsque le *cloud* auquel ont accès les *data scientists* est intégré officiellement dans la gouvernance IT de l'entreprise, les risques afférents ont été considérés et les processus qui encadrent l'utilisation de ces ressources font qu'il n'y a pas de risque particulier spécifique à l'intelligence

artificielle. Dans la plupart des cas en effet, ces clouds sont soit hybrides soit privés et offrent certaines garanties en termes de protection des données.

En revanche, il faut prêter attention à l'utilisation de ressources de type *cloud* en dehors de cette gouvernance IT. En effet, dans les cursus universitaires, les enseignements s'appuient souvent sur les *cloud* publics comme bacs à sable pour les travaux pratiques et projets des élèves. Les jeunes diplômés développent donc une expertise dans l'utilisation de ces ressources mais sont rarement sensibilisés au fait que dans le cadre d'une entreprise privée, certains aspects – exposition de données confidentielles, exposition à des réglementations extraterritoriales, propriété intellectuelle - peuvent être problématiques. Il faut donc s'assurer d'une **certaine sensibilisation** des nouveaux arrivants. S'il est nécessaire de faire appel à un *cloud* public pour un besoin d'accès à la puissance de calcul suffisante, notamment de *GPU (Graphics Processing Unit*), alors il faut s'assurer qu'un **processus de chiffrement** est bien en place. En tout état de cause, les outils utilisés doivent être au catalogue IT et il ne faut pas déployer, sans accord de la sécurité informatique, des outils non référencés (*shadow IT*).

Notons qu'il existe un environnement réglementaire pour le cloud, dont le projet européen DORA[30].

2.4.1.2 CYBERSECURITE DE L'IA

Un système d'IA, au même titre que les autres actifs informatiques, est sensible à la plupart des attaques relevant de la cybersécurité. Le niveau de risque est fonction de **l'exposition et de l'accessibilité** du modèle et des données utilisées pour entrainer le modèle, et intrinsèquement lié à la sécurité mise en place par l'entreprise pour sécuriser les systèmes IT.

Les principaux types d'attaques sont :

- ***Poisoning*** : les données d'entraînement sont par exemple modifiées afin d'induire un changement dans les résultats du modèle. Ceci suppose que l'attaquant dispose d'un accès aux données d'entraînement ;
- **Oracle** : l'attaquant tente **d'extraire des informations** sur le modèle IA, voire sur les données utilisées pour son entraînement, en « requêtant » le modèle un grand nombre de fois et en analysant les résultats ;
- **Evasion** : l'attaquant, en supposant qu'il ait connaissance du modèle IA ou de la façon dont ce dernier a été créé, **modifie** certaines données en entrée du modèle IA afin de produire un résultat différent de celui que le modèle aurait dû donner.

En particulier, on peut produire des **attaques « adversarielles »** en construisant des exemples « adversariels » (ou quelques fois adverses), c'est-à-dire des exemples auxquels on a ajouté une perturbation imperceptible (la **perturbation « adversarielle »),** et qui, du fait de cette perturbation, se retrouvent mal classifiés. Par exemple un *spam*, dans lequel on modifie un caractère, est désormais classé comme non-*spam*. Le problème est de savoir déterminer la **perturbation** qui apportera le résultat escompté. Ces attaques peuvent être mises en œuvre pendant l'entraînement (*poisoning*) ou en production (évasion).

[30] Proposition de règlement du Parlement européen et du Conseil sur la résilience opérationnelle numérique du secteur financier. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0595

2.4.2 MAINTENANCE ET MONITORING DU MODELE

Quand un modèle est en production et fournit un service, il convient de s'assurer que ce service correspond toujours à celui qui est attendu. Ceci fait intervenir des équipes mixtes IT et métier. Outre le suivi IT qui permet de s'assurer que le service est bien disponible, il faut également s'assurer que le résultat d'un modèle est toujours pertinent pour l'usage qui en est fait pour l'entreprise et pour le public.

- De la même manière que la performance du modèle a été évaluée au cours de sa construction, il faut s'assurer que cette **performance reste la même** tout au long de son utilisation. Ceci implique de définir des **indicateurs pertinents** pour suivre cette performance, et pouvoir la suivre dans le temps. Ceci est d'autant plus vrai si un modèle se réentraîne automatiquement afin d'identifier une **éventuelle dégradation** des performances, alors que la vérité terrain n'est pas encore disponible (exemple en fraude ou blanchiment). Une solution est de conserver un échantillon qui est systématiquement **analysé manuellement**. Cela implique des coûts et donc la nécessité de dimensionner correctement l'échantillon;
- Ce suivi de la performance doit être accompagné d'un **processus opérationnel** qui traite les dégradations de performance des modèles, avec la définition a priori de **seuils significatifs**. Ce processus devrait impliquer l'équipe qui a créé le modèle afin de déterminer les **causes de la dégradation** de performance : est-ce que les données utilisées sont toujours représentatives de l'objectif recherché ? Est-ce que l'algorithme utilisé doit être changé pour améliorer la performance ? Y a-t-il un changement des comportements des utilisateurs ? Y a-t-il des problèmes de qualité des données (ou un changement de données externes) ? Une **analyse** doit être réalisée afin de déterminer la cause de la diminution de la performance par rapport au modèle initial et les solutions à apporter.

Outre la performance du modèle IA, son **coût carbone** doit également être évalué et documenté. Ainsi, le **type de hardware** (e.g. GPU), la **mémoire utilisée** et la **performance en temps de calcul** doivent être renseignés pour les phases expérimentales, en particulier lors de l'entraînement du modèle. En effet, l'énergie consommée pour alimenter le hardware et la mémoire dépend principalement du type de matériel et du temps d'utilisation. En outre, l'efficience **du data center et sa localisation** détermineront l'impact final en grammes de gaz à effet de serre émis.

2.5 TRANSFERT AU METIER

Le transfert au métier d'un modèle correspond à sa **mise en œuvre opérationnelle**. Cette phase comporte un ensemble d'éléments qui ne sont pas spécifiques à l'IA : déploiement de nouveaux outils, formation, conduite du changement, mis en support, dispositif de contrôle et de surveillance.

Cependant, l'usage de l'IA implique des **risques nouveaux** et des **adaptations particulières** des dispositifs. Le risque peut être augmenté lorsque le cadre de développement du modèle ne correspond pas au cadre habituel de gestion des projets et mise en production usuels auquel les utilisateurs sont habitués.

2.5.1 INTERPRETATION ET EXPLICATION DES RESULTATS

La façon dont le modèle basé sur l'IA produit des résultats n'est pas nécessairement intuitive et sa logique de fonctionnement peut être obscure. Le **premier niveau de risque** se situe lors

du **transfert** par les experts modèles aux métiers. Ces experts doivent donc bien être en relation étroite avec les métiers. Quand bien même le **risque d'adéquation des besoins** aurait été traité en amont, le résultat final et la capacité du modèle à répondre au besoin présente potentiellement des écarts. Parmi les **points de vigilance** à traiter notamment lors de la formation, on note :

- L'utilisation du modèle par le ***Process Owner*** métier : le risque réside dans un **mauvais positionnement** de la brique basée sur l'IA au sein d'un processus métier. Ce risque peut apparaître lorsque le responsable du processus n'a pas la bonne compréhension de la finalité du modèle proposé ;
- La description du type de résultat produit en fonction des **données entrées, la nature et la marge d'erreur** possible sur la production du modèle et le regard critique à apporter sur la cohérence de la réponse. Les modèles produisant un **indice de fiabilité** en plus de leurs résultats modèrent ce risque. Ce risque peut être mitigé par une **acculturation à l'IA** de manière générale ainsi qu'une formation spécifique des utilisateurs métier plus poussée, avec des cas limites illustrant la **marge d'interprétation** à employer. Le risque se situe ici au niveau des **utilisateurs directs** du modèle ;
- La bonne utilisation des résultats produits : le modèle a été conçu dans un but et un contexte précis. Il faut contrôler le **risque de mauvais usage**, notamment celui d'utiliser les résultats du modèle pour une autre finalité qu'initialement prévue.

2.5.2 ROLES ET RESPONSABILITES

Lors de la construction et du déploiement du modèle, les rôles et responsabilités peuvent être **mal définis ou mal compris**, ce qui peut engendrer un risque sur le métier. Parmi les points de vigilance à surveiller propres à l'IA, le *Model Owner* doit s'assurer que l'ensemble des tâches suivantes ont bien été définies et attribuées :

- Suivi (***Monitoring***) du modèle en BAU : qui a la charge de suivre la performance du modèle ? Quels sont les indicateurs qui sont suivis ? Quel outil est utilisé pour ce suivi ? Quels sont les seuils ou règles appliqués pour déclencher des alertes ? Qui alerter, dans quel contexte ?
- **Soutien utilisateur** : quel est le bon interlocuteur en cas de résultats aberrants ou inattendus : le support utilisateur classique ou les concepteurs du modèle ?
- **Formation utilisateur** : qui est en charge de la formation des nouveaux utilisateurs et qui est garant du maintien du savoir-faire ?
- **Processus post-recalibration** : quel type de validation après une recalibration du modèle ou de changement de version (par exemple lorsque le modèle exploite des librairies transverses qui sont mises à niveau) ?

3 MESURES DE CONTROLE POUR LE TOP10 DES RISQUES

Après avoir défini, comme décrit précédemment, une liste de risques, notre groupe de travail a cherché à établir lesquels paraissaient être les plus importants du point de vue des **conséquences associées**, prenant en compte la maturité des sujets et les contrôles mis en place en plus de la probabilité et des conséquences directes.

Pour ce faire, chacune des trois banques a revu la liste des risques et a ordonné de 1 and 10 les risques présentés. Nous avons ensuite revu conjointement nos notations, afin d'ajouter des nuances à notre estimation. Le résultat de cette **homogénéisation qualitative** est présenté

dans cette section, ordonnée pour faciliter la lecture selon les étapes globales du processus présenté en Figure 4 (et non selon l'importance relative de ces 10 risques).

3.1 RISQUE D'INADEQUATION DE L'IA AU BESOIN METIER

Nos expériences communes nous ont menés à mettre en avant le risque de **non-alignement ou d'inadéquation de l'IA aux cas d'usage identifiés**. Quantifier le risque lié au modèle suppose de le faire par rapport à une **finalité** qui est définie par le métier et l'usage qu'il fait du modèle. La **définition de l'usage et des limites d'emploi** du modèle sont donc essentielles. Si cette base est mal définie, le reste du projet sera naturellement impacté.

Comme présenté dans la section précédente, la source de ce risque réside dans la combinaison de plusieurs facteurs, notamment :

- La **méconnaissance des métiers sur l'IA** en général et des situations dans lesquelles ces techniques sont pertinentes, des situations dans lesquelles elles ne le sont pas, et les risques associés ;
- Le **mauvais cadrage du projet** où la variable cible est mal définie par rapport au cas d'usage et au contexte ;
- La **définition ou choix des indicateurs de performance** ne représentant pas correctement la performance de ce modèle, notamment d'un point de vue métier.

Dans ce cadre, il est à noter que **l'exposition médiatique** dont jouit l'intelligence artificielle tend à exacerber ce risque en incitant des personnes peu documentées à s'engager dans des projets liés à l'intelligence artificielle pour ne pas être en reste, sans pour autant avoir la capacité (i) à bien comprendre ce que cela peut leur apporter en général, (ii) à évaluer la difficulté des tâches et (iii) à évaluer la pertinence des solutions proposées.

Une des manières les plus directes de diminuer ce risque réside dans **l'application de mesures visant à disséminer la connaissance de l'IA et des risques associés au sein de l'entreprise**, que ce soit de l'acculturation globale auprès des métiers demandeurs (formation aux principaux enjeux de l'IA et au système envisagé, compréhension du vocabulaire adopté, etc.) et enfin ciblée vers les *Model Developers* pour leur faire comprendre les enjeux des métiers demandeurs.

Au-delà de l'aspect **sensibilisation**, une autre mesure se situe au niveau de la gestion des projets d'IA, par exemple en passant par la mise en place **d'une gouvernance transverse qui valide la pertinence du cas d'usage** d'IA au moment de l'initiation, une fois le cadrage achevé. Enfin, une dernière mesure, quoique tardive dans le cycle de développement du cas d'usage, réside dans **le contrôle de la qualité de la documentation** décrivant les besoins, les objectifs et le dispositif retenu pour y répondre.

Il est à noter que ce risque diminue aussi grâce à la **présence d'équipes d'IA en interne** qui permettent un accès plus simple à cette expertise.

3.2 ABSENCE DE GESTION DES RISQUES LIES A L'UTILISATION DE DONNEES PROTEGEES ET/OU SENSIBLES DANS L'APPRENTISSAGE DE L'IA

Les algorithmes d'IA ont besoin d'une quantité importante de données pour leur apprentissage. La **collecte de ces données** est une étape importante précédant la modélisation.

Des règles et contrôles doivent être édictés et mis en place afin **d'encadrer la manière dont ces données sont collectées et utilisées** dans le cadre de la construction d'un système d'intelligence artificielle.

Selon les différentes réglementations existantes, **les données doivent être classifiées** en données publiques, personnelles, sensibles, etc. L'utilisation de ces données, hors données publiques, dans un algorithme d'IA doit être **encadrée** car l'utilisation (ainsi que la collecte) de certaines données personnelles (par exemple, l'origine ethnique ou l'orientation sexuelle) est proscrite. Leur utilisation pourrait créer une discrimination fondée sur ces données, ce qui serait préjudiciable pour certaines catégories de personnes et l'établissement créant ce type d'algorithme d'IA.

Cette maîtrise des données protégées et/ou personnelles passe d'abord par des processus externes de **gestion des droits d'accès à ces données**. Des moyens de protection doivent être mis en place afin de garantir leur confidentialité, leur accès et leur utilisation. Des solutions de type ***data masking*** peuvent être envisagées si l'utilisation de certaines données (anonymisées ou agrégées) est nécessaire, réduisant ainsi le risque d'identification des personnes et la divulgation de données protégées et/ou personnelles.

Afin de répondre à ces différents besoins au moment de la collecte des données, une attention particulière, au travers d'un processus de gestion des risques et/ou d'un outil, doit être apportée aux données qui sont nécessaires au modèle.

Outre la sécurisation et l'accessibilité de ces données, une **certaine discipline doit être suivie afin de limiter le nombre de données utilisées** au cours de l'apprentissage de l'IA. Compte tenu de la diversité et quantité de données susceptibles d'intégrer les modèles, une première mesure de maîtrise du risque consiste à s'assurer que l'ensemble bénéficie d'une **gouvernance et d'un pilotage adapté**. Les établissements peuvent s'appuyer sur les dispositifs de mise en conformité avec les réglementations existantes en la matière, comme le cadre général du BCBS239 ainsi que celui du RGPD. Au-delà de ce dispositif de contrôle, on peut vérifier en quoi il est adapté à la volumétrie, à la sensibilité des données, à la diversité des sources par rapport au dispositif standard. Ces contrôles peuvent se faire sur la base d'entretiens et de revue des contrôles opérationnels (objectifs et résultats). On peut s'assurer que le dispositif de collecte est dimensionné pour la **volumétrie atypique de l'IA** : test de charge, analyse des incidents de production, revue des logs d'alimentation, contrôle des taux de qualité des données les plus sensibles.

La prolifération de données en entrée du modèle n'est pas forcément garante d'une meilleure performance du modèle, celui-ci devenant plus complexe à comprendre, sans compter le besoin en ressources pour l'entraîner. Le principe de minimisation (***need-to-know***) peut être appliqué : seules les données nécessaires au(x) but(s) recherché(s) devraient être accessibles et utilisées. Les ***Model Developers*** doivent également être **sensibilisés aux problématiques de l'utilisation de données personnelles** dans les algorithmes d'IA.

Même si les règles définies par les diverses réglementations doivent être mises en place pour des traitements de données « standards », la volumétrie de données traitées dans le cadre de la création d'un algorithme d'IA peut **complexifier l'implémentation des contrôles** des données garantissant le plein respect de ces réglementations. En effet, les contrôles et l'analyse des données ou de leur contenu ne peuvent pas forcément être réalisés au niveau des données individuelles. Il est alors important de mettre en place des contrôles permettant d'avoir une **vision globale** (identification et utilisation) des données à caractère personnel utilisées par l'algorithme d'IA. Ces contrôles doivent être capables eux-mêmes de traiter un **volume de**

données important (données en entrée, ou logs de traitement). Ces contrôles doivent également inclure des analyses de la qualité des données présentant un caractère personnel car leur utilisation, outre leur encadrement réglementaire, peuvent générer des **effets non désirés** (par exemple un biais). Ces contrôles, si automatisés, doivent également être complétés par des **revues régulières** afin de vérifier la pertinence des contrôles et leurs résultats.

Une dernière mesure se situe au niveau de la **revue du modèle,** durant laquelle la **pertinence des données** utilisées par le modèle, notamment celles sous le joug de la règlementation sur les données personnelles mais pas exclusivement, doit être évaluée.

3.3 ABSENCE D'IDENTIFICATION D'UN BIAIS (DIRECT OU INDIRECT) ET CREATION OU AMPLIFICATION LIEE A L'UTILISATION D'UNE OU PLUSIEURS DONNEES EN ENTREE DANS L'APPRENTISSAGE DE L'IA

Au cours de la création d'un algorithme d'IA, il est indispensable de comprendre les **biais potentiels** qui peuvent exister dans les données utilisées pour l'apprentissage, et ensuite être amplifiés par la **modélisation**. Un algorithme d'IA a pour objectif de reproduire un raisonnement et/ou une prédiction à partir de données. Ces données proviennent du monde réel, sont parfois, voire souvent, incomplètes et ne représentent généralement qu'un sous-ensemble d'une population auprès de laquelle les données ont été collectées. En se fondant sur ces données, un algorithme d'IA aura forcément tendance à reproduire ces biais.

La première étape est donc de les **identifier**, non seulement dans les données en entrée, mais aussi **réduire des biais** qui pourraient être introduits dans le processus de modélisation lui-même. En effet la modélisation, qui s'appuie nécessairement sur des hypothèses et des approximations, peut influencer (en réduisant ou en amplifiant) le **caractère discriminatoire** des données.

Il est donc indispensable d'inclure **l'identification des biais dans le processus de construction d'une IA**. L'identification d'un biais permettra de comprendre s'il aura un impact sur la prédiction et l'objectif recherché. L'identification peut se faire via **l'analyse de métriques de performance** sur des sous-groupes. Pour une tâche de classification, il existe différentes métriques de mesure de biais en sortie du modèle. Parmi les plus courantes, la ***demographic parity*** reflète que, dans le cas d'un modèle d'octroi, le taux d'acceptation du modèle pour deux sous-populations doit être le même. **L'*equal opportunity*** reflète la même idée en étant conditionnée à la population n'ayant pas fait défaut. Par exemple, le taux d'acceptation du modèle doit être le même pour les femmes et pour les hommes ayant bien remboursé leur emprunt. Il est souvent impossible d'être bon en même temps sur toutes les mesures, car elles peuvent être incompatibles[31]. Il n'est pas toujours évident d'estimer quel écart est autorisé pour les valeurs de ces métriques entre deux populations. Une approche est d'utiliser le ***disparate impact***. Il correspond au ratio d'une métrique de mesure de biais (par exemple, la *demographic parity*) entre la population protégée et celle privilégiée. Le seuil de 80% ou loi des 4/5 est communément utilisé, sans que cela en fasse un seuil réglementaire pour les modèles d'IA (sauf aux USA).

La deuxième étape consiste à **identifier la source du biais**, c'est-à-dire quelles sont les variables incriminées ou susceptibles d'expliquer la différence de comportement du système d'IA. Com-

[31] Kleinberg J, Mullainathan S, Raghavan M. Inherent trade-offs in the fair determination of risk scores. arXiv preprint. Sep 19 2016. https://arxiv.org/pdf/1609.05807

parer la distribution des variables du modèle conditionnellement à l'appartenance à une population privilégiée ou non est une méthode efficace pour caractériser et comprendre la source du biais.

La dernière étape est de traiter le biais en étudiant les **possibilités de remédiations** qui auront un impact plus ou moins important sur la performance du modèle. L'objectif est donc de rendre indépendantes les variables sensibles et les sorties du modèle. Il existe **trois types d'approche**, toujours à manipuler avec précaution, pour ne pas risquer de créer d'autres biais :

- ***Pre-processing*** : ces méthodes visent à résoudre le biais en amont, en modifiant les données d'apprentissage via la suppression de variables, l'ajout de données ou la pondération des observations. La suppression d'une variable sensible peut être suffisante pour corriger le biais du fait des corrélations de la variable sensible avec d'autres variables explicatives (biais d'association). La **technique de *Reweighting*** réalise une repondération des échantillons par groupe et label pour viser l'indépendance de la variable sensible versus le label ou la classe ;
- ***In-processing*** : ces méthodes visent à résoudre le biais en modifiant l'apprentissage du modèle d'IA, par exemple en ajoutant un terme de régularisation dans la fonction de coût. Il est souvent clé de choisir un objectif (par exemple, *equal opportunity*) car ils sont rarement tous satisfaits. La méthode **d'*adversarial debiasing*** apprend à un estimateur à prédire correctement l'étiquette cible tout en minimisant la capacité d'un adversaire à prédire la variable protégée ;
- ***Post-processing*** : ces méthodes visent à résoudre le biais en aval, en sortie du modèle. Par exemple, la méthode de ***Reject Option Classification*** corrige les prédictions incertaines suivant le groupe d'appartenance (protégé ou privilégié). Une alternative serait de calibrer un seuil de décision par population.

Le contexte Big Data implique qu'il est plus complexe d'identifier et de traiter les biais pour de multiples sous-groupes, c'est-à-dire pour une large combinaison de variables sensibles. L'élimination de biais multiples est pratiquement impossible. Ainsi les mesures de contrôles qui permettent de limiter ce risque incluent tout d'abord celles qui s'appliquent au risque abordé précédemment (3.1 Risque d'inadéquation de l'IA au besoin métier). Il s'agit donc de s'assurer de la mise en place **de dispositifs de gestion et de gouvernance des données** par les *data office* et la conformité à destination des acteurs de la première ligne de défense (par exemple informations et caractéristiques des données, sensibilisation des modélisateurs, formation et boîtes à outils pour identifier, mesurer voire compenser un biais).

Une certaine discipline **d'identification des biais est aussi à exiger au niveau de la modélisation** en s'appuyant sur les techniques que nous avons brièvement décrites dans cette section. Cela peut s'accompagner tout d'abord de **mesures de sensibilisation** des *Model Developers* à cette problématique et leur donner accès à des outils ou librairies pour faciliter cette tâche. Des outils existent et sont proposés par IBM (AI Fairness 360), Microsoft (Fairlearn) ou Google (*What-If tool*). Une sensibilisation plus large des *Model Owners* / métiers est souhaitable pour les aider à eux aussi poser les bonnes questions au *Model Developers* lors des discussions autour de la modélisation. Cependant, la recherche exhaustive de tous les biais possibles et leurs combinaisons reste une utopie.

Naturellement, un deuxième contrôle s'inscrit dans **la revue indépendante du modèle** dès lors que le modèle s'appuie sur des données personnelles (revue généralement réalisée par la deuxième ligne de défense).

Enfin, la troisième ligne de défense réalise généralement des revues des dispositifs de contrôle mis en place par les deux autres lignes de défense afin de s'assurer de leur bon fonctionnement et leur adéquation avec les réglementations existantes.

Les considérations précédentes sont applicables dans le cas où le modèle est développé de bout en bout par l'établissement. Le cas de la réutilisation d'un modèle pré-entraîné se pose car on dispose rarement d'informations détaillées quant aux données utilisées pour la construction du modèle ou encore quant à la façon dont le modèle a été construit (algorithme, hyperparamètres, etc.). Au regard des questions posées précédemment et de la future réglementation européenne (*AI Act*), un fournisseur d'une solution IA devrait mettre à disposition des informations relatives aux données utilisées et fournir des résultats d'analyse des biais afin de garantir un minimum de **transparence** et des résultats qui ne créeront ou n'amplifieront pas des biais préexistants.

3.4 MAUVAISE COMPREHENSION OU INTERPRETATION DES DONNEES UTILISEES DANS LA MODELISATION

Le risque spécifique que nous cherchons à souligner ici est celui lié à la **mauvaise compréhension ou interprétation** des données utilisées dans la modélisation.

Ce risque n'est pas spécifique à l'IA, mais il est plus important car les volumes de données qu'il est possible d'utiliser pour modéliser impliquent qu'il y a plus de chances que certaines des données utilisées ne soient pas correctement comprises. En d'autres termes, la masse de données en entrée des modèles limite la capacité d'analyse intrinsèque quant à leur nature.

De surcroît, nos capacités actuelles de suivi des données dans les infrastructures existantes (source, transformation et normalisation) restent limitées, notamment en raison des évolutions des systèmes d'information de manières disparates et dans des environnements souvent hétérogènes. La bonne interprétation de l'information est souvent sujette à la disponibilité d'experts humains ayant une connaissance pratique de cette information.

Ce risque a une importance particulière car il est difficile à identifier *a posteriori*. En effet, la **complexité** des modèles, voire leur **faible interprétabilité** ne permet pas toujours d'identifier les caractéristiques pouvant avoir un impact significatif (qualité, représentativité, erreur de mesure, etc.).

Afin de prévenir ces quiproquos dans la compréhension de l'information, plusieurs mesures de contrôles et bonnes pratiques sont à considérer :

- **Réunions avec le métier** pour comprendre le cas d'usage et permettre à l'équipe de modélisation de faire des contrôles sur les données fournies en entrée en termes de volumétrie et de distribution ;
- Construction d'un **dictionnaire exhaustif des données** utilisées pour l'apprentissage, l'inférence et le *monitoring* (incluant les filtres) ;
- **Réunions avec l'IT** pour comprendre l'architecture IT, sources, les formats/unités des données dans les bases ;
- Le processus de **revue indépendante du modèle** permet aussi de s'assurer qu'un regard indépendant est porté sur les variables utilisées en entrée du modèle.

3.5 RISQUES LIES AU REAPPRENTISSAGE AUTOMATIQUE EN CONTINU

Certains systèmes qui embarquent des modèles de *Machine Learning* incorporent un mécanisme de **réapprentissage automatique en continu**, afin de permettre au modèle de s'adapter au mieux à l'évolution des requêtes et ainsi apprendre en continu plutôt que par exemple à date fixe. Cette configuration est particulièrement pertinente dans les cas d'usage qui peuvent changer rapidement, comme la détection de la fraude. Il faut cependant s'assurer que la configuration du réapprentissage automatique en continu est bien pensée et que les risques suivants sont contrôlés :

- **Risque de divergence** rapide du modèle qui doit être suivi ;
- **Risque de manipulation** du modèle accru (cybersécurité) ;
- **Renforcement du biais de sélection** : le modèle influence directement la sélection de ses propres futures données d'entraînement.

Il est nécessaire qu'un **dispositif de contrôle de ce réapprentissage en continu** existe et qu'il permette non seulement la mesure du risque mais aussi son pilotage, c'est-à-dire par exemple la capacité à basculer sur un autre modèle potentiellement moins performant mais plus robuste le cas échéant. Une mesure de contrôle potentiellement coûteuse mais efficace par rapport au problème évoqué consiste à mettre en place une **stratégie d'échantillonnage spécifique et indépendante** (manuelle), qui permet de mesurer la performance du modèle. Par exemple, dans le cas de la fraude, une sous-partie des transactions peut être systématiquement revue par un humain.

Dans la revue du modèle, il est aussi important d'estimer dans quelle mesure le flux de données d'entraînement au cours du temps dépend des sorties du modèle. Il s'agit de comprendre les **interactions potentielles** entre **prédiction et données d'entraînement.**

Des **méthodes alternatives** peuvent être utilisées pour suivre et comparer les performances d'un modèle dans le temps, telle la création d'un ou plusieurs ***challenger models*** (i.e. solutions et/ou implémentations différentes essayant de résoudre un même problème), ou l'utilisation d'un modèle non-supervisé pour le comparer à un modèle supervisé.

3.6 DEFICIT D'INTERPRETABILITE / EXPLICABILITE DES SYSTEMES D'IA

Dans une régression linéaire, il est aisé de déterminer le poids de chaque variable et le sens positif ou négatif de l'impact grâce aux coefficients. De manière similaire, la structure des branches d'un arbre de décision illustre la séquence des règles permettant d'aboutir aux prédictions. Un modèle d'IA est souvent considéré comme une boîte noire car il est très difficile voire impossible pour un être humain de prédire la décision du modèle sur un nouveau jeu de données. Ainsi ce **manque de transparence intrinsèque** peut limiter la capacité du métier à valider si un modèle d'IA a un sens métier en termes de variables sélectionnées et d'influence de ces dernières sur les décisions.

La méthode d'explicabilité doit également être adaptée au but recherché et au public visé. Quatre niveaux d'explication ont été proposés par l'ACPR en juin 2020[32] et un Tech Sprint sur l'explicabilité des algorithmes d'intelligence artificielle a été réalisé durant l'été 2021. Parmi les

[32] https://acpr.banque-france.fr/gouvernance-des-algorithmes-dintelligence-artificielle-dans-le-secteur-financier

résultats, les **principes d'intelligibilité** (i.e. compromis entre la fidélité et la sobriété de l'explication, concision de l'explication et tenter de concilier les explications locales et globales) et **d'interactivité** (i.e. adapter l'explication aux objectifs du destinataire) ont été identifiés.

Pour un cas d'usage donné, les *data scientists* (*Model Developers*) et le métier (*Model Owner*) doivent **évaluer conjointement le degré d'interprétabilité requis**, que ce soit pour la validation du modèle ou les besoins de transparence pour les utilisateurs du système d'IA. En fonction des hypothèses faites sur la relation entre les variables explicatives et la variable cible, le degré d'interprétabilité changera. Les modèles fondés sur des relations linéaires et monotones ou les arbres de décision ont un fort degré d'interprétabilité. Au contraire les modèles basés sur des relations non-linéaires et non monotones auront un degré très bas d'interprétabilité.

L'idée principale derrière les **techniques d'explicabilité** (XAI[33]) est de fournir un certain nombre de métriques (par exemple, l'importance des variables) ou d'éléments (par exemple, l'extraction de règles) qui permettront aux modélisateurs et au métier de mieux comprendre les décisions d'un modèle. La recherche en XAI est très active, et certainement pas encore aboutie.

L'explicabilité peut être faite à différents niveaux. L'explicabilité globale permet de comprendre les décisions du modèle en moyenne, au travers du classement de l'importance des variables (*Feature importance*), de la forme de la relation entre les variables explicatives et la variable cible (*Partial Dependence plot*), des interactions (ou synergies) entre les variables explicatives ou de l'extraction de règles globales (*Anchors*), via un modèle de substitution simple (*Global surrogate model*). L'explicabilité locale permet de comprendre les décisions sur une observation en particulier ou un sous ensemble d'observations. La décomposition en valeurs de Shapley et les analyses contrefactuelles sont des outils utiles pour l'explicabilité locale.

L'explicabilité peut être soit spécifique à un type de modélisation (par exemple, *DeepLift* pour les réseaux de neurones), soit agnostique. Les méthodes d'explicabilité, dites agnostiques, consistent à séparer les explications du modèle et sont donc indépendantes du type de méthodologie utilisée. En fonction des méthodes d'explicabilité choisies (par exemple, LIME et SHAP), l'explication peut être différente voire incohérente. Il est donc primordial de s'assurer que l'on peut avoir confiance dans l'explication, en mesurant sa fiabilité, de connaître les hypothèses et les limites des méthodes d'explicabilité utilisées.

Pour accompagner la **sensibilisation des *Model Developers*** sur le sujet, il est recommandé de fournir également des outils qui facilitent la construction d'explications. La librairie Shapash[34] par exemple permet d'évaluer la pertinence d'une explication en calculant trois métriques :

- La **stabilité locale**, en comparant l'explication issue de différentes méthodes d'explicabilité sur des instances similaires ;
- La **cohérence**, en vérifiant si les explications provenant de différentes méthodes d'explicabilité sont similaires en moyenne ;
- La **compacité**, en analysant si quelques-unes des variables sont suffisantes pour expliquer les décisions du modèle.

La documentation théorique de la conception, contenant notamment la quantification de l'influence des variables sur les prédictions, la mise à disposition d'une documentation à la

[33] eXplainable AI : IA explicable
[34] https://github.com/MAIF/shapash

portée du métier ainsi que des interactions fréquentes entre les *data scientists* et le métier sont indispensables pour assurer une bonne transparence dans la construction d'un système d'IA.

3.7 DEPLOIEMENT D'UN MODELE INSUFFISAMMENT STANDARDISE, SECURISE ET CONTROLE

Après la conception et l'implémentation du modèle, celui-ci est **déployé dans le système d'information** afin d'être accessible aux utilisateurs.

Ce déploiement doit être réalisé selon le même processus de déploiement qu'une application « standard » dans le système d'information, c'est-à-dire suivre des **processus de validation et de contrôles relatifs au déploiement d'applications**. En effet, des tests d'intégration et de validation restent toujours nécessaires afin de s'assurer du bon fonctionnement du modèle dans l'environnement de production qui diffère parfois des environnements de développement. Ces contrôles sont nécessaires pour détecter d'éventuels effets avec d'autres modèles existants (par exemple pour le cas de modèles dépendants les uns des autres), des incompatibilités (par exemple, des versions de librairies non validées et non disponibles en production).

Les procédures et les standards de déploiement d'applications doivent prendre en charge le déploiement de modèles IA afin de s'assurer qu'un modèle n'entraîne pas de risques non-identifiés liés à la cybersécurité, l'insuffisance ou la surconsommation des ressources IT par exemple. Certaines plateformes de modélisation facilitent aujourd'hui le déploiement d'un modèle, par exemple au moyen du déploiement d'une API exposée sur cette même plateforme. Dans certains cas de tests, POC ou par dérogation, un tel déploiement peut être accepté par les équipes infrastructures et sécurité de l'entreprise, mais ceci devrait soit s'inscrire officiellement et en toute sécurité dans les procédures IT de l'entreprise, soit rester un fonctionnement par exception.

Il est ainsi à noter qu'une **revue des processus IT existants** peut être nécessaire afin d'y intégrer les spécificités de l'IA et donner accès aux développeurs de modèles à un processus adapté à leurs contraintes. Le caractère parfois novateur d'un modèle l'IA ne devrait pas conduire à l'absence d'applications des règles existants dans l'entreprise quant à la mise en production des modèles. **Une sensibilisation des équipes amenées à utiliser les plateformes de modélisation** peut s'avérer nécessaire afin de les accompagner dans le déploiement d'un nouveau modèle. Dans ce cadre, il faut préconiser **la mise en place de processus MLOps** qui permettent une intégration plus forte des équipes métier, des équipes d'IA et des équipes d'IT.

3.8 ABSENCE DE KPIS ET/OU ABSENCE D'UN PROCESSUS DE MONITORING

Une fois le modèle déployé, il est essentiel de réaliser son suivi dans le temps. En cas d'absence de gouvernance ou de métriques de suivi pertinentes, une **déviation des performances** risque de passer inaperçue.

Il est tout d'abord clé de **définir une gouvernance de *monitoring* des modèles** en production. En effet, des métriques de suivi sans rôle spécifié exposent à un manque de réactivité dans les actions à mener. Ainsi, il est important de définir et documenter les responsabilités ainsi que les métriques, les seuils et actions associées. Par exemple, si la performance du modèle passe sous un certain niveau prédéfini, le responsable du *monitoring* devra escalader afin que soit décidé par exemple un réentraînement du modèle. Afin que les actions soient menées avec réactivité, la fréquence de suivi doit être alignée avec la matérialité du cas d'usage. La granularité des observations prend par ailleurs tout son sens. Avec un niveau d'observation trop global,

des biais régionaux pourront passer inaperçus. A l'opposé, si le niveau est trop granulaire, la comparaison dans le temps sera complexe. Par ailleurs, la profondeur des périodes de suivi doit être pertinente afin de pouvoir comparer les métriques à une période de référence (comme la période de développement du modèle). Des périodes plus larges peuvent être suivies ponctuellement, par exemple dans le cadre d'un exercice de *back-testing* annuel approfondi. Enfin, les résultats du *monitoring* doivent être discutés régulièrement entre les différentes parties prenantes, et notamment entre les modélisateurs et les experts du domaine.

En second lieu, **l'infrastructure de *monitoring* doit être suffisamment robuste** pour éviter toute perte d'information. Les données utilisées doivent refléter la production, notamment en termes d'exhaustivité et de granularité. Par ailleurs, le périmètre de suivi doit être aligné à celui d'application.

En troisième lieu, **le choix des métriques et des seuils de rupture associés est tout aussi prépondérant**. Ces derniers doivent être alignés avec l'appétit au risque tout en prenant en compte les limitations propres du modèle. Différents niveaux d'alerte et d'action associés sont recommandés. Par ailleurs, le suivi évolutif dans le temps avec comparaison à la période de référence est un élément clé pour détecter de manière anticipée des déviations émergentes.

Des **métriques globales de production** permettent de suivre l'exposition au risque, par exemple, un volume fortement croissant des demandes d'octroi. Dans un contexte Big Data, il est important de contrôler la disponibilité et la qualité des sources de données. Si le nombre de variables du modèle est élevé, des contrôles de qualité spécifiques peuvent se limiter aux *features* les plus importantes.

Pour éviter toute prédiction incohérente du modèle, il est important de surveiller les ***outliers*** ou les données manquantes. En effet, certains modèles d'IA ou packages d'implémentation sont moins robustes aux anomalies. En accord avec la gouvernance, des actions correctives peuvent ainsi être décidées.

Le contexte Big Data impose aussi de contrôler les temps moyens d'inférence pour éviter tout problème de latence du fait d'une forte augmentation des volumes de production. Par ailleurs, le modèle ayant été entraîné sur une population représentative, il est conseillé de suivre la **stabilité des distributions** des données. En effet, celles-ci pourraient évoluer fortement vers des zones où le taux d'erreur du modèle est plus élevé.

Ensuite, les variables du modèle doivent garder leur **pouvoir discriminant** au cours du temps. Ainsi, une variable qui n'aurait plus d'effet significatif alourdirait la dette technique. La performance du modèle doit être mesurée à l'aide des métriques utilisées lors de la modélisation.

En outre, la comparaison à des benchmarks s'appuyant sur des modélisations alternatives permet de détecter des prédictions incohérentes.

Si un cadre du type ***human-in-the-loop***[35] est appliqué, il est recommandé de suivre les taux des *overrides* (exceptions) et leur justification. Cela permettra de comprendre certaines faiblesses du modèle et de réaliser certains ajustements, e.g. la recalibration d'un seuil ou ajout de nouvelles variables explicatives.

[35] Avec intervention humaine dans le processus de décision

Si le modèle s'applique sur des personnes physiques, il est aussi recommandé de suivre dans le temps des métriques d'évaluation des biais pour s'assurer qu'elles restent dans des intervalles acceptables.

Enfin, il est nécessaire de comprendre si les facteurs influençant la décision sont toujours les mêmes au cours du temps. Ceci peut être réalisé, par exemple, en **agrégeant les valeurs de Shapley** des instances locales et en comparant les résultats à la période de référence.

3.9 RISQUE DU TRANSFERT METIER

Comme il est primordial au début d'un projet de s'assurer de l'adéquation de l'approche d'intelligence artificielle avec les besoins métiers, il est aussi primordial de s'assurer que lors du transfert du modèle au métier, les résultats et les explications du modèle sont pertinentes et comprises.

La maîtrise du risque à l'étape du transfert métier peut concerner plusieurs volets. Pour limiter et maîtriser les risques inhérents à cette phase aval du recours à l'IA, il est possible de s'assurer des éléments tels que :

- **Revue du plan de conduite du changement** lors de l'insertion opérationnelle du modèle dans le processus métier. Ceci peut être évalué directement au moyen d'entretiens et de tests sur les usages effectifs. Une évaluation indirecte peut se faire en analysant les incidents métiers et leur gestion (pour les modèles disposant déjà d'un historique d'utilisation suffisant) ;
- **Efficacité de la formation utilisateur évaluée** par entretien ou questionnaire aux utilisateurs, vérification que les utilisateurs ont formellement accès à l'information nécessaire au bon usage (existence d'un guide utilisateur), analyse de la pertinence des explications, analyse de la cohérence entre explications utilisateurs et réalité du modèle théorique ;
- Présence d'une **cartographie des risques** et d'un **plan de contrôle adapté** ;
- **Contrôle des usages effectifs** des résultats du modèle et de la pertinence des cas où le jugement humain est utilisé pour corriger cet usage ;
- Existence **d'un dispositif de surveillance des usages**. Lorsque le résultat du modèle est embarqué dans un progiciel, le contrôle se fait via les habilitations. Lorsque le modèle produit des données accessibles plus largement, le dispositif peut s'insérer dans celui encadrant la gouvernance de la donnée.

On peut enfin analyser les écarts entre plan de déploiement prévu et la réalisation effective.

3.10 MAUVAISE DEFINITION DE LA GOUVERNANCE : ROLES ET RESPONSABILITES

Lors du transfert au métier, il est important de définir correctement la gouvernance relative à la gestion du système d'IA sur son cycle de vie et de s'assurer que les rôles et responsabilités sont clairement définis. En effet, le passage en production ne marque pas la fin des travaux. Un suivi continu et adapté à la matérialité du cas d'usage doit être mis en place.

Les **dispositifs de gestion de risque de modèle** (*Model Risk Management*) définissent généralement des rôles clefs, dont celui de propriétaire du modèle (*Model Owner*) qui doit s'assurer de la pertinence du système d'IA au cours du temps en s'appuyant sur un dispositif de surveillance.

Les politiques de changement de modèle définissent la **typologie des modifications** (par exemple, changement d'un paramètre), leur matérialité et la nécessité d'une revue par un

tiers indépendant (par exemple, seconde ligne de défense). La définition de ce type de politique en amont du passage en production permet de fluidifier les interactions entre les différentes parties prenantes et d'atténuer les risques liés à des mises à jour.

Dans le **processus de définition des rôles et responsabilités**, les points de contrôle à envisager sont, par exemple :

- Les **règles d'approbation** des recalibrations de modèle et d'information des utilisateurs ;
- Les **interlocuteurs des métiers clairement identifiés** en cas de dysfonctionnement, capacité pour le support de premier niveau de se retourner vers les concepteurs du modèle en cas de valeur inattendue ;
- Les règles de mise à jour de la formation utilisateur et implication des concepteurs du modèle à chaque modification. Contrôle du niveau de formation et de sa mise à jour ;
- Le rôle du ***process owner*** : peut être évalué en contrôlant le bon positionnement de la brique IA au regard du process, notamment dans le respect de l'échéancement des tâches si la brique outille un processus de production par exemple ;
- Les processus à suivre et les personnes à impliquer dans le cadre de la mise en production de nouvelles versions des librairies (et surtout du retrait d'anciennes versions) afin de s'assurer de la **compatibilité des nouvelles versions** et si besoin de faire évoluer les modèles ;
- Enfin, **l'implication et le soutien du management** au moment du transfert est primordiale pour s'assurer que les rôles et responsabilités sont pris au sérieux. Il faut donc faire en sorte que le management communique clairement les **principaux enjeux liés à l'insertion opérationnelle du modèle.**

4 CONCLUSION

L'Intelligence Artificielle est un **levier significatif** d'amélioration des processus bancaires, tant en termes de **création de valeur** pour les clients qu'en matière **d'amélioration de l'efficacité opérationnelle.** Elle s'accompagne cependant de **nouveaux risques** qu'il est nécessaire de comprendre pour pouvoir les gérer et les contrôler, et ainsi tirer pleinement et sereinement parti des **opportunités** proposées par l'utilisation de l'IA. Les processus de gestion des risques et de conformité mis en place dans les banques couvrent d'ores et déjà une grande partie des risques identifiés. La nécessité de gérer les risques induits par l'IA **ne constitue donc pas un changement disruptif** pour les banques ou pour certains secteurs dits « critiques » (nucléaire, aéronautique, véhicule autonome, santé).

En revanche, d'autres secteurs moins coutumiers de ce type de dispositif devront mettre en place leur propre système de contrôle des risques : les analyses présentées ici leur fourniront, nous l'espérons, des pistes de réflexion pour se les approprier et les adapter à leur contexte.

Les nouveaux risques ainsi créés ou augmentés par l'IA sont notamment en lien avec **l'utilisation des données** (par le *Machine Learning*), la **transformation et la conduite du changement** des processus métier qui doivent être modifiés pour incorporer les modèles IA, les **processus IT** au moment de mise en production des modèles IA, ainsi que les **risques de cybersécurité**, qui sont augmentés par l'utilisation des techniques IA rendant les attaques plus efficaces, ou même par de nouvelles techniques IA (attaque « adversarielle »).

Afin d'y remédier, certaines actions doivent être entreprises, comme :

- Mettre en place une **gouvernance forte,** à haut niveau et transverse ;
- **Adapter et standardiser** les processus internes ;

- **Sensibiliser** l'ensemble des collaborateurs, depuis les membres du Comité Exécutif jusqu'aux opérationnels ;
- **Promouvoir** la proximité des équipes métier, *data science* et IT ;
- **Etablir une culture du contrôle des risques** pour obtenir un **niveau de confiance** dans l'IA acceptable de la part des utilisateurs internes comme des clients externes.

Aujourd'hui, les principes de contrôle des risques sont largement formalisés dans les banques, avec une organisation en trois lignes de défense et une répartition claire des rôles et responsabilités, constituant un socle solide pour la gestion des risques induits par l'IA. Cependant, dans la pratique, la mise en œuvre des analyses de détection des risques liés à l'IA et de leur atténuation reste **majoritairement non industrialisée**, en dépit d'une recherche constante de standardisation des processus.

Les risques identifiés dans le cadre de ce livre blanc sont génériques, la démarche adoptée reposant sur un recensement de risques spécifiques à l'IA et une proposition de méthodes de contrôle. L'évaluation de leur impact étant propre à chaque modèle d'IA et chaque contexte d'utilisation, ils ne sont ni quantifiés dans l'absolu ni hiérarchisés. C'est donc volontairement que nous ne fournissons pas de grille de score universelle permettant une appréciation globale du risque d'un cas d'usage. L'AI Act[36] au contraire définit un niveau de risque *a priori* basé sur les usages, pour un ensemble de modèles dépassant le cadre du *Machine Learning* que nous nous sommes fixés ici. Les obligations réglementaires envisagées, pour ceux relevant du « *high risk* » en particulier, visent à maîtriser les risques globaux par un contrôle *ex ante* et une documentation, notamment une piste d'audit particulièrement exhaustive et complexe à mettre en œuvre. Ces obligations ne distinguent donc pas les facteurs de risques et les mesures *ad hoc* à envisager. Les propositions de contrôles présentées dans ce livre blanc ont ainsi pour vocation d'être complémentaires aux préconisations de la règlementation en devenir.

Avec l'arrivée des nouveaux textes de réglementation de l'IA (et notamment l'AI Act de la Commission européenne), les entreprises, y compris les institutions financières, vont devoir mettre en place des **processus systématiques et documentés de contrôle des risques**. Le **coût de la mise en conformité** pourrait alors drastiquement augmenter si l'entreprise ne conçoit pas rapidement un processus global de contrôle des risques qui accompagnera le processus de production des systèmes IA de bout en bout. Un tel processus de contrôle doit être **standardisé et outillé**, afin de le rendre plus efficace et économique afin de répondre aux exigences réglementaires.

En se focalisant sur les trois domaines que sont la **gouvernance, la culture d'entreprise et l'expertise métier,** les entreprises pourront se préparer à l'arrivée des futurs textes réglementaires et ainsi profiter pleinement des bénéfices attendus du déploiement de l'Intelligence Artificielle.

[36] https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-approach-artificial-intelligence

5 GLOSSAIRE

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
CNIL	Commission Nationale de l'Informatique et des Libertés
Dataset	Ensemble des données utilisées dans l'une des phases de modélisation
KPI	Key Performance indicator
GPU	Graphics Processing Unit
Human-in-the-loop	Intervention humaine dans le processus de décision
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 des risques 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.
RGPD	Règlement Général sur la Protection des Données
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)

6 REMERCIEMENTS

Ce livre blanc est le fruit d'un travail amorcé en 2021 par le Hub France IA dans le cadre du Groupe de Travail *Banque et Auditabilité*. Pendant plusieurs mois, ce groupe a croisé les retours d'expérience d'experts des différentes banques membres du groupe.

Nous sommes tout spécialement reconnaissants aux personnes suivantes, qui nous ont donné de leur temps et ont partagé leurs expériences, côté grand groupe et côté start-up.

Contributeurs

- **Léa Deleris**, Head of RISK Artificial Intelligence Research, BNP PARIBAS.
- **Jérôme Lebecq**, Data Science Coordinator, BNP PARIBAS.
- **Ludovic Mercier**, Inspection Générale, Directeur de pôle, LA BANQUE POSTALE.
- **Audrey Agesilas**, Superviseur – Model risk Audit, SOCIETE GENERALE.
- **Benjamin Bosch**, Manager - Model risk Management – Data Science, SOCIETE GENERALE.
- **Thomas Bonnier**, Model Risk Manager – Data Science, SOCIETE GENERALE.
- **Caroline Chopinaud**, Directrice Générale, HUB FRANCE IA.
- **Françoise Soulié-Fogelman**, Conseiller Scientifique, HUB FRANCE IA.

Relecteurs

- **Nathalie Bouez**, Head of RISK Independent Review and Control, BNP PARIBAS.
- **Fabrice Le Chatelier**, Head of Data Science Office, BNP PARIBAS.
- **Michael Rabba**, Model Risk Senior Manager, BNP PARIBAS.
- **Rim Tehraoui**, Group Chief Data Officer & Global ESG Risks Executive, BNP PARIBAS.
- **Pierre Contencin**, Responsable Validation des modèles, LA BANQUE POSTALE.
- **Emmanuel Jouffin**, Responsable du Département Veille Réglementaire Groupe, LA BANQUE POSTALE.
- **Fabien Monsallier**, Directeur innovation du Groupe, LA BANQUE POSTALE et Directeur Général, 115K.
- **Matthieu Olivier**, Chief Data Officer, LA BANQUE POSTALE.
- **Clémence Panet**, Chief Data Scientist, LA BANQUE POSTALE.
- **Julien Bohné**, Chief Data Scientist, SOCIETE GENERALE.
- **Anne-Cécile Krieg**, Deputy Head of Model Risk Management, SOCIETE GENERALE.
- **Julien Molez**, Group Innovation Data & AI Leader, SOCIETE GENERALE.
- **Eric Peter**, Head of Group model's audit, SOCIETE GENERALE.
- **Mélanie Arnould**, Chef des Opérations, HUB FRANCE IA.
- **Pierre Monget**, Chef de Projet, HUB FRANCE IA.
- **Andréa Arnaud,** Chef de Projet, HUB FRANCE IA.

Hub France IA

Octobre 2022

BNP PARIBAS

LA BANQUE POSTALE

SOCIETE GENERALE

HUB FRANCE IA