Décembre 2020

Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier

Analyse des réponses à la consultation

Auteur : Laurent Dupont

Pôle Fintech-Innovation, ACPR

Introduction

Cette consultation de l'ACPR à recueilli 26 réponses écrites, provenant d'établissements bancaires, de prestataires technologies, de cabinets de conseil, d'associations professionnelles, d'instituts de recherche académique. Quasiment toutes les réponses écrites sont complètes, couvrant les 23 questions de consultation, et ont une longueur moyenne de 18 pages. Elles sont complétées d'une dizaine d'entretiens téléphoniques avec les acteurs du secteur sur les sujets couverts par le document de réflexion.

Le nombre relativement faible de réponses reçues étant attribuable au vaste champ couvert par la consultation, un premier enseignement est que l'ACPR aurait dû insister davantage sur la pertinence de réponses partielles. D'autres motifs de non-réponse étaient que certains organismes n'avaient rien à redire aux principes détaillés dans le document de réflexion. Enfin, certaines associations professionnelles n'ont pas pu consolider les contributions de leurs membres en une réponse homogène.

Ce document présente l'analyse de l'ensemble des réponses, en tentant pour chaque question de dégager la tendance générale de l'ensemble des réponses, ainsi que les divergences de positions le cas échéant, en soulignant enfin les points de vue ou propositions les plus éclairants ou sujets à réflexion.

Le reste de ce document reprend la structure du questionnaire (et donc celle du document de réflexion).

Contexte

Question 1 : Expérience en intelligence artificielle (IA) et notamment en *machine learning* (ML)

La plupart des répondants ont su développer en interne des compétences en IA et en ML satisfaisant à leurs besoins en conception et intégration de ces technologies, même si l'ensemble des entités au sein d'une organisation (filiales d'un groupe bancaire, entités locales, etc.) ne partagent pas le même niveau de familiarité et d'expérience. Aussi la quasi-intégralité des répondants ont-ils mis en œuvre des initiatives transverses de formation et de sensibilisation au ML.

Organisation des compétences en ML

Les répondants présentent une grande diversité dans leur organisation des travaux de *data science* : organisations décentralisées ou structures de type *Datalab*, *Model Hub* ou *IA Factory* (où les *data scientists*

conservent un rôle transverse ou se spécialisent dans un métier pour y jouer le rôle de relais). Dans les grands groupes financiers, la conception du ML est souvent réalisée dans les filiales si celles-ci disposent des compétences, et sinon par les experts au niveau du groupe, tandis que la validation tend à être centralisée. Les échanges entre entités et groupe permettent alors de garantir une bonne appropriation du modèle par les équipes utilisatrices.

Profils-types

Les institutions financières qui en ont les moyens disposent à la fois de profils académiques et d'experts en recherche appliquée, tandis que les fintechs ou les fournisseurs de solution se concentrent sur les enjeux pratiques et la construction de plateformes.

Question 2 : Mise en œuvre du ML

Algorithmes

La visualisation suivante représente les types d'algorithmes d'IA mentionnés par les répondants, catégorisés en familles technologiques.

Figure 1 : Algorithmes de ML utilisés par les répondants

Cas d'usage

La visualisation suivante représente les cas d'usage mettant en œuvre de l'IA rapportés par les répondants, catégorisés en domaines métier.

Figure 2: Cas d'usage du ML parmi les répondants

Dans cette visualisation, certains cas d'usage ont été regroupés sous les dénominations suivantes :

- « événements sur compte » : les événements en question sont soit des opérations frauduleuses, soit des anomalies quelconques ;
- « analyse de produits financiers » : les processus concernés sont l'automatisation de l'analyse de prospectus en contrôle dépositaire et la génération de la demande de renseignements sur un produit financier ;
- « sélection médicale » : pré-autorisations de traitements médicaux, assistance au traitement des questionnaires médicaux ;
- « événements clients » : résiliation, rachat (*churn*), moments de vie, etc.

Un point d'attention est le nombre de cas d'usage en NLP (traitement du langage naturel), ensemble de technologies non couvertes par le document de réflexion de l'ACPR qui se focalisait sur les processus métier critiques (par opposition à ceux destinés à la simple optimisation de processus internes de gestion). L'ACPR prend bonne note d'inclure le NLP et ses applications dans ses études futures.

Principe d'explicabilité

L'ensemble des répondants notent un intérêt et des motivations croissants pour l'explicabilité de l'IA, pour les raisons multiples suggérées par le document de réflexion : aider au développement d'algorithmes performants et robustes, répondre aux exigences réglementaires, permettre l'auditabilité interne et externe (autant de tâches qui requièrent maîtrise des technologies et compréhension de leur usage), mais aussi faciliter l'adoption des solutions d'IA en interne.

Question 3 : Définition des niveaux d'explication

Les répondants sont généralement en phase avec les définitions.

Beaucoup notent toutefois une difficulté à voir l'articulation entre les niveaux N2 et N3, qui sont selon eux les niveaux les plus importants, en particulier comment N2 peut être basé sur des résultats (simplifiés) de N3.

N1 est jugé peu important, car il serait principalement utile pour des personnes curieuses d'un modèle mais sans réel besoin de le comprendre.

N4 est vu comme à part des autres : difficile à atteindre dans la pratique, il correspondrait davantage à un test de correcte implémentation ou à une piste d'audit qu'à un besoin d'explication, et serait parfois atteignable par des échanges entre pairs. Il risque même selon certains d'aboutir à des exigences trop importantes (par exemple accès au code développé par un prestataire) et sans réelle utilité pour la compréhension des individus concernés.

Quelques répondants s'interrogent sur la pertinence de hiérarchiser ces niveaux, alors qu'ils semblent complémentaires. Quasiment tous estiment que la définition des niveaux pourrait être clarifiée en illustrant chacun par une mise en situation, selon une méthode parmi celles suggérées :

- en détaillant, pour un modèle particulier de ML (régression linéaire ou *random forest*), ce qui est nécessaire pour atteindre chaque niveau ;
- inversement, en associant à chaque niveau une liste d'algorithmes s'y prêtant (par exemple pas de Deep Learning en N4) ;
- sur un cas d'usage donné, en explicitant chaque niveau par type d'audience : par exemple une décision de refus d'un prêt, qui se déclinerait en un modèle de notice informative, une explication destinée aux *data scientists*, etc.

Un point d'attention concerne les explications contrefactuelles :

- Elles sont souvent peu connues des répondants.
- Certains les jugent contradictoires avec par exemple une pratique de non-divulgation des données précises entrant dans le scoring de décision d'octroi - pratique visant à pallier les tentatives de contournement et de falsification de données déclaratives.
- L'acceptabilité des explications contrefactuelles par le client semble douteuse à certains répondants. A l'inverse, selon un autre répondant les explications contrefactuelles permettent de ne pas révéler l'ensemble des mécanismes du modèle, et ainsi de préserver le secret industriel.

Question 4 : Adéquation des niveaux d'explication

Les répondants considèrent globalement adéquat le choix de ces niveaux, et sont en accord avec le nombre de 4 (bien que certains en préconisent 3). Les 4 niveaux sont jugés intéressants car ils constituent une graduation qualitative et non quantitative.

Comme mentionné plus haut, la frontière entre N2 et N3 est ténue et souvent floue : N2 nécessite en partie N3, la distinction entre méthodes conjointes et post-modélisation[1] est douteuse, si bien que N2 en tant que tel est peut-être superflu, ou alors on pourrait avoir un « N2 simple » et un « N2 détaillé ».

[1]Pour rappel, les méthodes explicatives conjointes à la modélisation consistent pour le modèle à « apprendre à expliquer », cet apprentissage étant intégré à l'apprentissage des prédictions (sorties du

Un répondant note qu'expliquer comment le modèle a été conçu est au moins aussi important que l'explication du modèle résultant. La documentation du processus de création de l'algorithme et la mise à disposition du code et des données sont des conditions nécessaires (quoique non suffisantes) à cette exigence de transparence.

Un répondant corrobore - par un exemple vécu sur le terrain - la proposition du document de réflexion d'inclure dans le périmètre des explications non seulement l'algorithme et le modèle de ML mais aussi les prétraitements des données, notamment le *feature engineering* qui crée des variables intermédiaires *a priori* cachées à l'utilisateur voire à l'analyste.

Facteurs de choix

Les répondants sont globalement d'accord sur les deux facteurs de choix du niveau d'explication décrits dans le document de réflexion, à savoir la criticité du cas d'usage et les destinataires de l'explication.

Un répondant précise toutefois que le niveau attendu devrait aussi dépendre du périmètre d'application du modèle. Ce périmètre peut d'ailleurs souvent être défini par des critères quantitatifs, tels que le pourcentage d'encours sur le total et non pas uniquement le risque associé. L'exigence d'explication devrait selon ce répondant être soumise à un principe de proportionnalité incluant ces différents paramètres.

💡 Plusieurs répondants notent aussi à juste titre que le niveau d'explication devrait intégrer la dimension de robustesse des méthodes explicatives : certaines produisent des explications basées sur des points de données hautement improbables, ou attribuent à deux points très proches des explications radicalement divergentes. Partant, la stabilité des explications elles-mêmes devrait être évaluée, y compris en extrapolant à des données non observées.

Question 5 : Exemples pratiques de niveaux d'explication

Ajustement des niveaux proposés

Les rares propositions d'ajustement des niveaux proposés sont les suivantes :

Cas d'usage « propositions d'indemnisation en assurance » : lorsque l'audience de l'explication est le client, celui-ci pourrait être aussi intéressé par la compréhension du "pourquoi", notamment s'il n'est pas satisfait par le montant d'indemnisation proposé, ce qui implique N2.

Cas d'usage « remontée d'alertes en gel des avoirs » : lorsqu'un collaborateur en niveau 2 (équipe conformité) analyse des alertes remontées par un modèle de ML, il serait nécessaire de lui fournir des explications locales, ce qui implique N3. Le but est d'aider les agents d'une part à développer leur confiance dans les modèles de ML, d'autre part à mieux analyser les alertes en se concentrant sur certaines variables soulignées par le modèle. Un répondant préconise même N4 en raison des risques importants de non-conformité ou de discrimination, et note par ailleurs que certains destinataires des explications ont été oubliés pour ce cas d'usage : les clients bancaires, l'autorité d'investigation financière (TRACFIN en France), et l'autorité mandatée pour la protection des données personnelles (la CNIL en France).

Cas d'usage « modèles internes » : N4 pourrait être exigible pour les équipes de validation qui auditent les scores réglementaires selon Bâle 3 mais pas pour l'ensemble des scores d'octroi ou d'orientation de processus (par exemple aide à la détection de fraude), quant au contrôle permanent N3 est préconisé.

Plus généralement, un répondant indique d'une explication destinée au consommateur final devrait par défaut être N2 ou N3.

---

modèle) ; elles contribuent ainsi à la conception de modèles intrinsèquement plus explicables. Les méthodes explicatives post-modélisation visent pour leur part à produire des explications à partir de modèles préalablement construits et entraînés.

Couverture des cas d'usage

Plusieurs répondants notent la limitation des exemples donnés à des données structurées (tabulaires), notamment l'absence de certains cas d'usage pourtant fréquents dans le secteur parmi les exemples cités : les modèles opérant sur des séries chronologiques, et ceux opérant sur des données non structurées (*chatbots* clients, traitement automatique de documents ou d'images envoyées par les clients, analyse d'emails clients pour détecter les intentions ou irritants, etc.). Dans ce dernier cas en particulier, l'utilité d'une explication sur un modèle de NLP est souvent limitée (en revanche un double contrôle manuel peut être souhaitable pour améliorer les performances du modèle) ; le contrôle de l'algorithme se fait donc par ses résultats, sans essai d'explication.

> 💡 Un répondant expert en LCB-FT (sécurité financière) note qu'il conviendrait d'aborder les usages plus avancés de l'IA dans ce domaine, tels que les méthodes d'analyse de graphes pour réduire les faux positifs (Weber, 2019) ou le SVM non-supervisé pour découvrir de nouveaux schémas de fraude (Tang, 2005). Ces usages sont en effet prometteurs, tout en posant des défis d'encadrement et de gestion des risques.

Outre le niveau d'explication, les attendus pour un cas d'usage donné devraient selon certains répondants inclure la forme de l'explication : par exemple elle peut être en langage naturel (français ou anglais) ou plus technique (par exemple les valeurs SHAP).

Autres facteurs

Pour un répondant, l'angle réglementaire fait défaut car il faudrait notamment examiner :

- l'applicabilité des exigences d'explicabilité induites par l'Article R311-3-1-2 du Code des relations entre le public et l'administration (au moins comme bonnes pratiques pour l'ensemble du secteur, notamment pour les décisions algorithmiques impactant directement les citoyens comme le refus de prêt) ;
- les autres textes réglementaires touchant à l'explicabilité algorithmique : Code monétaire et financier, Article 47 de la Loi Informatique et Libertés (du 8 janvier 1978), Article 22 du RGPD.

L'absence d'angle social est aussi déplorée, il conviendrait ainsi que l'ACPR ou la CNIL précise les niveaux associés aux enjeux sociaux tels que la non-discrimination. S'inspirant de l'« *Algorithmic Accountability Act* » en cours de préparation aux États-Unis, un « *Algorithmic Accountability Report* » pourrait être à présenter avant toute mise en production d'IA dans les cas d'usage à haut risque.

Principe de performance

Question 6 : Mesures techniques de la performance

Les mesures techniques de performance de l'IA sont selon l'ensemble des répondants utilisées principalement pour la conception des modèles (en particulier la sélection entre modèles challengers), leur validation et leur suivi.

Une critique adressée par plusieurs répondants aux exemples de métriques cités par le document de l'ACPR est que ces exemples ne concernent que les problèmes de classification binaire (et encore, sur des données non déséquilibrées). Il convient de garder à l'esprit la pertinence de métriques adaptées :

- à un type de problème
  - classification binaire sur des données déséquilibrées : *log loss* ;
  - classification multi-classe ;
  - régression : RMSE pour une distribution normale, MAE, etc. ;
  - en NLP : par exemple les scores BLEU ou ROUGE en traduction ;
  - le NLG (génération de texte) possède aussi ses propres métriques ;
- à un cas d'usage
  - dans le domaine assurantiel sont particulièrement usités MAPE (modélisation des primes), la déviance Poisson-Tweedie pour une distribution "avec excès de zéro" comme le montant des sinistres, ou encore la déviance Gamma pour une loi Gamma comme le nombre de sinistres ;

- les métriques de ranking telles que PairLogitPairwise or PairAccuracy sont à privilégier en LCB-FT (l'objectif étant de rendre les alertes plus pertinentes pour les clients, le tri est un critère majeur).

Il convient de distinguer les métriques destinées au *data scientist*, qui permettent d'optimiser les paramètres du modèle (erreur moyenne, score GINI, etc.), de celles destinées aux experts métier, plus facilement interprétables (*top-k* taux de vrais positifs, matrices de confusion, etc.). Lorsqu'une métrique est partagée par ces différents destinataires, une des responsabilités des équipes de *data science* en tant que concepteurs de modèle est de s'assurer de sa compréhension par le métier ainsi que de son adéquation.

Un acteur en particulier a opté pour une approche visant à simplifier autant que possible la quantité et la nature des métriques produites. Ainsi en classification, les métriques par défaut se réduiront à la précision (précision en %), la matrice de confusion, et quelques exemples de données mal classifiées.

💡 Un principe important souligné par certains répondants est de privilégier les métriques techniques qui ne nécessitent pas de fixer un seuil de décision à l'avance (par exemple AUC, *lift*, *log loss*) afin de pouvoir comparer des modèles entre eux, indépendamment du choix du seuil de décision, et de laisser le réglage du seuil à la main du responsable de modèle.

Question 7 : Mesures fonctionnelles de la performance

L'avis général des répondants est que les métriques fonctionnelles devraient être définies par les experts techniques (*data scientists*, *data engineers* et spécialistes IA qui jugent de leur faisabilité) en lien avec les experts métiers (qui jugent de leur pertinence), mais aussi avec les instances de gouvernance et les équipes conformité en ce qui concerne le pilotage du risque.

💡 Plusieurs répondants pointent par ailleurs la nécessité de bien distinguer métriques techniques et fonctionnelles :

D'une part la prise en compte de métriques fonctionnelles en apprentissage - en tant que critère d'optimisation - est à éviter. En effet elle impliquerait une redéfinition des algorithmes d'apprentissage dont la réussite n'est pas garantie, car les critères d'optimisation habituels sont basés sur des fonctions complexes, mettant en jeu par exemple des métriques d'entropie, généralement sans lien avec les objectifs métier discutés en amont de la construction du modèle.

À l'inverse, certaines données dites "chaudes" dont l'impact sur la performance technique est important (exemple en assurance : demande du relevé d'information) s'avéreront inutiles voire contre-productives en termes d'objectifs métier (le relevé d'information conduisant à se focaliser sur des dossiers où l'assuré a déjà entamé les démarches pour rejoindre un autre assureur).

Dans le cas d'utilisateurs internes, la performance fonctionnelle est mesurée en termes d'impact sur les processus en jeu (incluant traitements humains et autres traitements automatisés) : temps passé par les agents à analyser les résultats de l'IA, taux d'automatisation du processus global, risques induits par l'IA, réactivité permise aux événements non désirés (fraudes ou anomalies).

Lutte contre la fraude

Les métriques fonctionnelles les plus couramment citées sont la gêne client, le taux de détection en montant, le volume journalier d'alertes.

Sécurité financière

En LCB il s'agit de la pertinence du signalement, du pourcentage de couverture des scénarios (métrique de nature réglementaire), et de l'exposition au risque (définie comme moyenne des scores de transactions non bloquées).

Plus généralement en sécurité financière, on peut distinguer les métriques d'efficacité normative, par exemple :

- amélioration de la détection (taux de vrais positifs, identification de nouveaux schémas) ;
- couverture des scénarios identifiés (risque conformité) ;
- prévention du risque d'adaptation des fraudeurs, etc.

et les métriques d'efficacité opérationnelle :

- réduction du temps de génération d'alertes ;
- amélioration de la qualité des alertes ;
- réduction du temps de traitement des alertes ;
- réduction du risque opérationnel lié à la décision concernant l'alerte ;
- amélioration de la qualité de travail et montée en compétence des équipes ;
- rapidité et souplesse d'intégration de la solution dans le processus métier ;
- transparence et qualité des travaux des fournisseurs technologiques (livrables et SAV) ;
- explicabilité des résultats aux différents types d'audience.

Assurance

Quelques indicateurs de performance clés (KPI) fonctionnels mentionnés sont l'augmentation du nombre de souscriptions et de devis, le taux de reconnaissances des documents (pour l'exemple de la souscription automatisée à une assurance automobile), la précision des décisions positives automatisées (dans un processus de sélection médicale).

Principe de stabilité

Question 8 : Dérive temporelle des modèles

Les sources de dérive mentionnées le plus fréquemment par les répondants sont :

- Modification de la qualité des données en entrée du modèle.
- Changement de la distribution des données (apparition de nouveaux flux, apparition de nouvelles variables ou de *patterns*, etc.).
- Évolution de l'environnement technique autour du modèle (solutions open source ou de tierce partie, etc.).
- Réentraînement du modèle, surtout sur des échantillons déséquilibrés.
- Évolution de l'environnement fonctionnel ou réglementaire autour du modèle (nouvelles règles de filtrage, modification du workflow opérationnel, des systèmes de filtrage en amont du modèle, nouveaux services proposés, nouveaux clients etc.). Le phénomène modélisé peut lui-même évoluer, surtout s'il fait partie d'un système complexe (en économétrie) ou lorsque certains acteurs cherchent à déjouer un système de détection.
- Le contexte économique peut aussi impacter les performances d'un modèle : ainsi, un score d'appétence à l'achat d'un produit peut devenir moins pertinent si le contexte économique ou l'environnement concurrentiel est devenu défavorable à la vente de ce produit.

Les risques le plus couramment associés à la dérive temporelle de modèles sont les suivants :

- Baisse des performances de l'algorithme, jusqu'à rendre son utilisation contre-productive (sous-détection de cas de blanchiments ou de cas de fraudes, une mauvaise estimation du risque de crédit d'un client, ciblage marketing raté, etc.).
- Introduction de biais dans le modèle, menant notamment à la disparition des liens de causalité entre deux événements.
- Risque financier (refus / acceptation à tort, fraude non détectée, dégradation de la qualité de la réponse apportée au client, etc.)
- « Biais d'autoréalisation » : on peut ainsi désigner un biais généralement introduit par l'algorithme lorsque les prédictions influencent l'avenir. Les répondants ont fourni des exemples d'une telle situation allant dans deux directions opposées :
    - Rétroaction positive c'est-à-dire une boucle de renforcement, par exemple en ciblage marketing où, si un profil de client donné n'apparaît jamais dans le "top 10%" des propensions

à l'achat sur le tableau de bord d'un responsable clientèle, ce client achètera *in fine* moins souvent le produit.
    - Rétroaction négative, par exemple pour la détection de l'attrition en créant des faux négatifs (les clients à haut risque d'attrition ayant été « rattrapés » par une action d'un opérateur humain).

Les techniques de remédiation de dérive temporelle les plus fréquemment citées sont :

- *Out-of-sample testing*[2] afin de détecter et d'écarter dès la conception les modèles les moins stables, appliquée par exemple aux différentes entités ou filiales d'un même groupe (ainsi que la méthode plus sophistiquée de *nested cross-validation* permettant d'optimiser simultanément la performance prédictive et le choix des hyperparamètres).
- *Out-of-time testing*[3], méthode plus spécifiquement adaptée aux séries temporelles.
- Attention portée au choix de la période d'apprentissage, par exemple en prédiction des séries temporelles pour capturer les éventuels effets saisonniers.
- Phase de test et d'analyse du comportement du modèle dans le temps avant déploiement, pour déterminer une fréquence (éventuellement une borne minimale) d'actualisation du modèle adaptée.
- Suivi au fil du temps d'indicateurs de stabilité de la performance, de tests statistiques (contrôles de cohérence) sur les données. Dans les activités réglementées, les modèles devraient aussi être accompagnés de tableaux de bord et de rapports de suivi périodiques présentés aux instances de gouvernance, qui peuvent décider de les corriger. Ce suivi peut éventuellement déclencher une mise à jour automatique du modèle lorsqu'un seuil de dérive est franchi. Un répondant note que le déclenchement d'alertes en cas de dérives est un mécanisme pertinent mais à considérer comme un filet de sécurité, car il est souvent trop tard au moment de l'alerte pour éviter une période de transition où l'algorithme sera moins performant dû à la dérive.
- Pour le phénomène de biais d'autoréalisation introduits par l'algorithme et décrits ci-dessus, il convient d'enregistrer les actions ayant engendré une modification du comportement (dans l'exemple de l'attrition, si un discount est proposé au client et qu'il reste).

Un répondant considère l'analyse de stabilité plus complexe pour les modèles non supervisés, nécessitant une définition plus prudente du référentiel de comparaison et des métriques de déviation.

Enfin dans le cas de l'apprentissage au fil de l'eau (*online learning*), le modèle tend à stabiliser voire améliorer sa performance au cours du temps, tant que les modifications dans les données et l'environnement restent incrémentales.

Question 9 - Généralisation des modèles

Les répondants relèvent principalement deux limites au pouvoir de généralisation, outre le problème spécifique du sur-apprentissage : déséquilibre des données d'apprentissage pour les modèles de classification (problème pour lequel des remèdes existent) et biais de sélection des données (difficile voire impossible à remédier).

Le risque de mauvaise généralisation dépend largement du cas d'usage :

- il est souvent présent dans le cas de phénomènes dynamiques : par exemple en lutte contre la fraude, les fraudeurs s'adaptent au système de détection et changent leur comportement, donc la typologie de fraude doit être revue périodiquement ;
- il intervient aussi pour les modèles de prévision de cours de bourse ou d'autre quantité économique ;
- c'est moins le cas si le phénomène modélisé est statique et le corpus stable (par exemple reconnaissance d'image ou *speech-to-text*).

[2] La validation croisée (*out-of-sample testing)* permet d'évaluer la performance d'un modèle prédictif via une méthode de partitionnement des données d'apprentissage et de validation.
[3] On désigne ainsi une forme de validation croisée applicable aux séries temporelles : pour ce type de données, un échantillon de validation obtenu par stratification temporelle donne généralement de meilleurs résultats qu'un échantillonnage totalement aléatoire.

Un répondant attire l'attention sur les stratégies d'échantillonnage des bases d'apprentissage, qui peuvent diminuer la capacité de généralisation du modèle si elles sont inadéquates. Par exemple en détection de fraude, les individus qui ont déjà fait l'objet d'alertes doivent être exclus des ensembles de validation. Ou encore dans les modèles comportementaux des clients, l'échantillonnage sociologique, géographique et démographique des individus doit concorder entre données d'apprentissage et de production.

Quelques principes généraux de conception de modèles visent à pallier le manque de pouvoir de généralisation :

- la complexité des modèles doit être étudiée, et peut être limitée par conception (en s'abstenant par exemple de recourir à du *stacking* de modèles[4]) ;
- le développement de modèles susceptibles d'être adaptés à des cas d'usage spécifiques (via du *Transfer Learning*[5]) contribuerait à maintenir une généralisation élevée ainsi que des coûts de lancement de projet réduits.

Question 10 - Instabilité due au réentraînement

Impact sur la reproductibilité

On peut noter la tension entre d'une part l'exigence de reproductibilité des sorties du modèle avant et après réentraînement, d'autre part l'objectif de stabilité d'indicateurs statistiques globaux. Le principe de reproductibilité, défendu par plusieurs répondants, conduit à éviter un réentraînement trop fréquent des modèles et à privilégier le choix de modèles robustes dans le temps.

Impact sur la stabilité

Un répondant souligne que le réentraînement engendre inévitablement une instabilité dès lors que le phénomène modélisé est sujet à une dérive conceptuelle. Il ne s'agit alors pas de limiter l'instabilité, mais de capturer les modifications de la relation entre variables explicatives et variable cible, tout en évitant de considérer les variations dues au hasard, i.e. en optimisant le compromis stabilité / plasticité de l'algorithme (Hinder, 2020).

À l'inverse, certains répondants estiment que le réentraînement n'entraîne pas d'instabilité s'il est réalisé de manière raisonnée : toute évolution des hyperparamètres ou des jeux de données doit être mesurée, un contrôle de stabilité et des tests de non-régression formels sont appropriés, enfin les réentraînements périodiques peuvent être faits sur des échantillons se recouvrant en partie chronologiquement.

Réentraînement systématique

Un répondant - fournisseur de solutions de ML - indique avoir observé chez plusieurs clients bancaires une stratégie de réentraînement systématique, sur des données complètement nouvelles, sans lien avec le suivi des variables d'entrée. Il convient plutôt de ne rafraîchir le modèle que si la dérive des données est significative ou la performance dégradée. Un autre répondant note d'ailleurs que le réentraînement automatique n'est pertinent que dans des hypothèses d'évolution très rapide des phénomènes à modéliser : les seuls cas requérant du *online learning* à ce jour relèvent de la prévision des comportements digitaux des clients, tandis que dans tous les processus d'aide à la décision, les processus de gouvernance et de validation de modèles induisent une révision beaucoup moins fréquente.

Remèdes

Les palliatifs suggérés par les répondants sont similaires au cas de la généralisation : définir des seuils de qualité sur le nouveau modèle, garantir la réversibilité (*rollback*) à la version précédente du modèle, historiciser les

[4] Technique d'assemblage ensembliste de modèles où la sortie de plusieurs modèles prédictifs est utilisée en entrée d'un autre modèle.

[5] Méthodologie visant à assimiler la connaissance acquise sur un problème donné, puis à la réutiliser sur un problème en lien avec le premier mais différent – et souvent plus spécialisé.

modèles et leur performance, voire utiliser des métriques de type SHAP pour contrôler que les variables discriminantes restent les mêmes.

Principe de traitement adéquat des données

Question 11 : Conformité réglementaire en matière de données

Peu de réponses concernant le traitement des données sont spécifiques à l'IA, les points d'attention les plus récurrents étant l'anonymisation (au moment de la conception) et les questions de rétention et purge des données (après l'apprentissage) afin de satisfaire au RGPD.

La vérification de la conformité réglementaire intervient autant que possible en amont de la construction des modèles de ML, et dans tous les cas avant chaque mise en production : les services juridiques ainsi que les équipes dédiées au RGPD (DPD, RSSI, etc.) se concertent avec les *data scientists* pour vérifier la conformité du cas d'usage, la finalité des modèles et l'utilisation des données, indiquant les contours réglementaires à ne pas dépasser.

Un répondant note la friction - voire la contradiction - entre les contraintes liées à la rétention des données dans le RGPD, et le besoin de garder les logs de décisions individuelles et les données utilisées si l'on veut pouvoir générer des explications.

Dans la phase d'apprentissage sont pris en compte la finalité et le contexte de la collecte de données, la présence éventuelle de données sensibles, les conséquences possibles du traitement, et l'existence de garanties appropriées. Plusieurs répondants indiquent réaliser une revue des variables prédictives de chaque modèle mis en production afin de détecter celles qui pourraient soulever des difficultés, par exemple des biais à caractère discriminatoire – indirects ou latents.

Question 12 : Détection et remédiation des biais

Détection

Les répondants utilisent les méthodes suivantes pour détecter à différents niveaux les biais de nature statistique :

- dans les données : analyses statistiques (qualité, distribution, représentativité des données) ;
- dans les modèles lors de la validation (biais de conformité et d'équité) : *back-testing*, explications de type SHAP, revue manuelle, avec un panel de métriques d'équité (indice de Theil, parité statistique, *Equality of Opportunity*, comparaison des taux d'erreurs de classification entre sous-groupes) ;
- dans les modèles en production : PSI par rapport à la population de modélisation.

Dans la pratique, seule l'analyse des biais liés à des variables disponibles[6] est possible, ce qui limite intrinsèquement la portée des mesures de remédiation envisageables.

> 💡 Un répondant préconise la définition (si possible conjointe par les autorités compétentes, ACPR et CNIL) des métriques d'équité, en étant conscients que l'absence totale de biais est inatteignable en pratique.

Remédiation

La remédiation des biais de données est fréquemment basée sur l'exclusion de données discriminatoires ou le ré-échantillonnage (et sur- ou sous-échantillonnage).

La remédiation des biais de modèles est un domaine moins mature chez les répondants, même si certains utilisent des méthodes de calibration post-modélisation ou de régularisation.

Plusieurs répondants avertissent des conséquences négatives sur la discrimination globale d'un usage systématique de la remédiation, car certains discriminants socio-économiques ne sont que le reflet de la réalité et doivent être pris en compte dans l'exercice normal d'activité. Partant, il convient de distinguer l'IA utilisée

[6] Par exemple l'âge ou la nationalité, mais pas les variables à caractère sensible au sens défini par le RGPD.

pour les interactions avec les clients (par exemple outil d'assistance pour conseiller) des modèles prédictifs, pour lesquels la remédiation de biais ne semble pas nécessaire tant que le niveau de compréhension de l'importance et du sens des variables demeure acceptable.

Il est aussi noté que seule l'analyse des biais liés à des variables disponibles (par exemple âge et nationalité, mais pas les données sensibles au sens CNIL) est possible, ce qui limite grandement la portée des mesures de remédiation envisageables.

Intégration dans les processus

Question 13 : Rôle de l'IA

Intégration de l'IA

Les critères d'évaluation de l'intégration de l'IA les plus fréquemment cités sont :

- gain de temps et de productivité ;
- réduction de la pénibilité pour les salariés ;
- amélioration de la qualité du produit ou du processus, de la relation client ;
- qualité et aisance de la coopération entre humains et machines ;
- facilité d'explication des résultats (transparence, interprétabilité) ;
- conformité avec la réglementation, les règles éthiques en vigueur ;
- capacité à mettre en production et à intégrer l'algorithme (ou la chaîne algorithmique) avec le processus métier ;
- capacité à contredire une décision algorithmique ;
- traitement des données (mise en qualité, traitement sans risque de sécurité ni de confidentialité, etc.) ;
- intégration dans les processus de gouvernance en place ;
- confiance dans l'éthique des méthodes utilisées (absence de biais à caractère discriminatoire) ;
- facilité de maintenance du modèle dans le temps, en particulier la réactivité de l'IA dans une situation nouvelle (données non vues en apprentissage).

Un répondant opère une dichotomie intéressante entre deux scénarios-types d'intégration d'IA :

- Si des algorithmes sont en place depuis longtemps (régressions linéaires pour risque de crédit, règles métier pour la détection de fraudes, GLM pour la tarification en assurance) alors l'IA apporte de la précision mais pas un changement radical du processus métier. L'impact porte surtout sur le processus de modélisation, et donc ceux de validation et d'audit.
- À l'inverse, dans certains cas d'usage l'IA bouleverse le processus métier : les souscriptions d'assurance-vie arrivent avec une recommandation produite par la machine, ou triées selon un score machine, l'IA étant alors un outil d'aide à la décision. Le gain en productivité est plus important, mais s'accompagne d'un risque de rejet par le métier et de forçage par le management. Il est donc essentiel d'intégrer le métier dans le projet de construction de l'IA dès le départ, afin qu'il comprenne le modèle et soit à même de l'améliorer, il devient « manager de l'IA ».

Rôle de l'IA

Un répondant exprime son désaccord sur le caractère « non-disruptif » du ML en sélection de clients prospectifs pour réaliser du démarchage commercial ou de la vente croisée : l'utilisation de la voix ou l'image par exemple, comme source de données supplémentaires pour aider au conseil, voire automatiser le conseil avec du ML, est une rupture majeure dans le processus commercial.

Les questions posées par les répondants pour évaluer rôle de l'IA sont alignées avec celles décrites dans le document de réflexion :

- quelles conséquences aux erreurs de l'IA ?
- quels garde-fous (humains ou systèmes) ?
- quels moyens de remédiation en cas de défaillance de l'IA ?

- l'IA comme aide à la décision ou mode de décision automatique ?
- réversibilité des décisions de l'IA ?
- contrôle humain a posteriori ? si oui, dans quel délai suivant la décision en question ?

Interactions humain/machine

Deux répondants pointent un élément évident mais qui mérite d'être souligné : l'évaluation humaine des résultats de l'IA est essentielle, y compris de la part des clients et consommateurs, ce qui passe par exemple par la récupération des avis des clients sur la qualité de la réponse apportée par l'IA. Ce ressenti de l'utilisateur final (interne ou externe) devrait même être intégré dans l'analyse de performance et le suivi des dérives.

Un désaccord émerge des réponses quant à l'autonomie de l'algorithme :

- Certains conviennent que les humains doivent toujours garder leur libre arbitre vis-à-vis des algorithmes implémentés et doivent être encouragés à remettre en question les résultats.
- Pour un autre, cette autonomie génère du risque : dans l'exemple de la relation client à distance, les conseillers humains n'ont pas un jugement plus qualifié du niveau de risque d'un client que l'IA (et l'analyse aurait montré que les forçages de décision par un humain conduisent à des prises de risque plus importantes).

Plusieurs répondants soulignent l'intérêt des processus parallèles confiés à l'humain : cela peut permettre aux humains de ne pas perdre leur expertise sur des cas qu'ils n'auraient plus l'occasion de traiter, et c'est aussi un moyen d'évaluer l'algorithme en continu. À l'inverse, un répondant fait valoir que le *parallel processing* est coûteux, consommateur en ressources, et peut s'avérer difficile à mettre en place dans certains cas (par exemple lorsque l'ancien outil, sur lequel le processus opérationnel s'appuyait, n'existe plus) ; il aurait donc sa place dans la phase de construction d'un modèle, mais pas dans sa réévaluation continue.

Question 14: Méthodologie de conception de l'IA

Plusieurs répondants estiment que la méthodologie de conception de l'IA doit évoluer dans la même direction générale que celle de l'industrie logicielle, en particulier vers un meilleur contrôle qualité. Une convergence vers des principes de conception de type MLOps, par analogie avec le mouvement DevOps apparu au cours de la décennie passée, est souhaitée autour notamment des quelques axes suivants : reproductibilité, automatisation, traçabilité (via l'historicisation), standards d'écriture de code et méthodologie systématisée de tests.

Certains répondants insistent néanmoins sur une différence essentielle entre l'IA et le logiciel traditionnel, à savoir que la conception d'IA et surtout de ML procède par essais-erreurs. Plus précisément, un enjeu majeur lors du développement d'un logiciel traditionnel réside dans la décomposition en modules et en la bonne interaction entre ces modules. Pour l'IA, la décomposition n'est pas aussi simple : les étapes sont fortement interconnectées (par exemple les variables créées impactent la modélisation) et surtout l'impact d'une modification sur la performance du système est quasiment impossible à prédire. Le processus de développement d'IA s'apparente davantage à de l'essai-erreur qu'à un processus linéaire traditionnel[7].

Sur un plan stratégique, la conception d'une méthode standard d'ingénierie logicielle part en général du postulat qu'il existe une solution au problème posé et qu'il s'agit de l'implémenter. Dans la conception d'algorithmes d'IA, le postulat de base est qu'il existe des solutions concurrentes qui doivent être implémentées simultanément et être mises en concurrence afin de juger de leur efficacité sur les processus décisionnels.

Cette différence méthodologique se traduit aussi par des conséquences en matière d'évaluation et de validation. Ainsi il existe souvent une incertitude que l'algorithme d'IA parvienne à résoudre la tâche confiée avec les données à disposition : l'absence de bug dans le code ne signifie pas que l'algorithme soit sans danger, de plus la dérive est toujours possible, et un processus itératif non seulement sur le modèle mais aussi sur les données

[7] Au sens d'une méthodologie d'ingénierie logicielle de type « *waterfall* » ou même d'une approche itérative suivant les principes Agile telle que que *Scrum*.

est nécessaire. De plus la preuve formelle, difficile à atteindre dans le logiciel traditionnel, est ici illusoire et il convient d'éprouver le système par des techniques plus ou moins sophistiquées, allant des tests classiques (unitaires, fonctionnels, d'intégration) jusqu'à des attaques adversarielles.

Quelques différences plus rarement citées entre l'IA et le logiciel traditionnel sont :

- l'infrastructure spécifique souvent nécessaire pour IA ;
- les règles de sécurité autour des environnements de développement (en particulier la sécurisation des données) ;
- la nécessité d'un monitoring systématique des performances de l'IA en production.

Contrôle interne

Question 15 : Gestion des risques

Impact de l'introduction d'IA

Les répondants estiment que les risques liés à l'IA pour les processus métiers restent globalement les mêmes que pour les modèles statistiques classiques, avec une accentuation de certains risques. Ces changements sont détaillés dans ce qui suit.

Les algorithmes de ML sont des modèles et à ce titre, ils rentrent souvent déjà dans la gouvernance interne de gestion du risque de modèle. Plus généralement, de nombreux répondants estiment que l'IA est déjà couverte par un ensemble de politiques et processus de contrôle éprouvés au sein de leur organisation (gouvernance informatique, gouvernance des données, risques de modèles et risques de tiers), mais que ceux-ci sont amenés à s'adapter à l'évolution des usages et des technologies. L'effort de gouvernance exigible devrait cependant être proportionnel aux risques posés par le système : un *chatbot* de support technique interne ou un système de traduction, bien qu'utilisant tous deux des éléments d'IA, ne présentent pas le même niveau de risque qu'un système d'octroi de crédit à un client.

Il est particulièrement important d'évaluer suite à l'introduction d'IA le taux d'erreurs (de toute nature) dans le processus métier, surtout en le comparant aux systèmes remplacés ou à l'humain.

Le risque opérationnel est généralement amplifié, en particulier les questions de stabilité et le risque de réversibilité.

Enfin, de nouveaux risques réputationnels et juridiques peuvent être engendrés par des décisions automatiques à caractère discriminatoire. La question de la responsabilité du preneur de décision est cruciale. Une erreur de système et une erreur humaine ne sont pas soumises au même cadre juridique : l'attribution de la responsabilité d'une défaillance nécessite donc de déterminer si l'erreur est liée à la définition du processus décisionnel ou imputable au seul algorithme.

Gestion des risques

Parmi les méthodes de remédiation de ces risques nouveaux ou augmentés, les réponses mentionnent :

- Les questions de stabilité impliquent comme déjà mentionné une fréquence de mise à jour des modèles souvent plus élevée pour le ML. Lorsque la validation des modèles doit se faire ex-ante (dans le cadre prudentiel par exemple), certains répondants estiment même qu'il peut être nécessaire de se limiter aux algorithmes statiques (à l'exclusion des modèles auto-apprenants faisant du *Online Learning*).
- L'historicisation des algorithmes, modèles et données est requise pour garantir traçabilité et reproductibilité.
- Les besoins d'explicabilité, de contrôle des biais et de contrôle de robustesse sont eux aussi mentionnés comme essentiels à la bonne gestion des risques liés à l'IA, en ligne avec les quatre principes exposés dans le document de réflexion de l'ACPR.
- Parmi les bonnes pratiques d'ordre général est évoquée la cartographie systématique des risques de modèle avant toute mise en production.

- Pour limiter l'amplification du risque opérationnel par l'automatisation des processus, il convient d'intégrer des garde-fous permettant de garder le contrôle humain et d'éviter la propagation des erreurs algorithmiques. Par exemple, en cas de défaillance, une procédure de basculement (*failover*) peut être mise en place, soit vers un modèle plus simple et plus stable, soit vers une équipe humaine qui reprendrait la tâche avec plus de lenteur mais qui assurerait la continuité du service.
- Concernant le risque juridique, un répondant remarque que le département juridique est rarement impliqué dans les des projets d'IA sur lesquels un accompagnement et un encadrement auraient été nécessaires. Les juristes ne sont par ailleurs que peu informés des risques liés à l'IA, aussi certains acteurs ont-ils entrepris des actions de formation et de sensibilisation non seulement des décideurs mais aussi des juristes de leur organisation.
- Les risques de sécurité sont amplifiés si l'algorithme est exposé à une attaque de type adversarielle ou à une fuite de données.

Question 16 : Validation fonctionnelle

Les répondants s'accordent à considérer la validation fonctionnelle comme nécessaire tout au long du cycle de vie de l'IA, depuis la conception jusqu'au retrait et archivage du modèle.

Dans l'ensemble, le processus de validation en place chez les acteurs est similaire aux modèles classiques, néanmoins il est important que les directions de contrôle interne développent en propre des compétences en IA, de même dans un groupe bancaire ou assurantiel les responsables des entités devraient être familiers des processus de validation de l'IA et de leurs critères.

La validation fonctionnelle doit en outre, selon les répondants, être proportionnée au niveau de risque et de criticité des processus en jeu, et dirigée par les équipes techniques et les équipes en charge de la conformité et des risques.

Certains répondants introduisent une phase pilote avant la mise en production, au cours de laquelle les utilisateurs métier valident en situation réelle les performances de l'algorithme, afin d'ajuster au besoin ses paramètres.

Le monitoring de l'IA en production est indispensable pour tout processus critique ; il doit être d'autant plus poussé si le modèle est une boîte noire.

> 💡 Un répondant recommande de distinguer d'une part les scénarios de remplacement d'un système ou d'une tâche humaine par de l'IA, d'autre part le déploiement d'un nouveau service à base d'IA. Ce dernier cas requiert des spécifications fonctionnelles destinées au département IT et écrites conjointement par les modélisateurs, les utilisateurs finaux et les experts métiers.

Enfin, plusieurs répondants établissent une analogie avec les standards de l'industrie aéronautique ou aérospatiale, plus précisément les logiciels embarqués, suggérant que l'attention à l'ensemble de la chaîne de conception et production doit être renforcée, et les méthodes d'analyse adaptées.

Question 17 : Modèles de risques internes

Utilisation du ML

Un répondant note la tension entre le besoin d'auditabilité (totale explicabilité) des modèles internes dans un secteur réglementé et l'une des promesses de l'IA, à savoir faire le moins d'hypothèses possible et laisser l'algorithme ajuster les paramètres du modèle et les données utilisées par celui-ci. Il conviendrait donc de réfléchir à une documentation sous forme de notice adaptée aux modèles internes les plus complexes.

Les réponses divergent concernant la possibilité et les modalités d'utilisation du ML pour les modèles bâlois et les modèles internes en assurance. Pour certains, le ML peut être utilisé en prédiction si :

- son périmètre d'utilisation est bien défini ;

- des notes détaillées sont fournies sur la démarche scientifique justifiant le choix de tel algorithme de ML ;
- le pouvoir prédictif du modèle initial est amélioré ;
- son interprétabilité est accrue, avec des pistes d'audit permettant de comprendre son comportement (le but est d'obtenir un modèle suffisamment explicable pour permettre son appropriation par le métier) ;
- la maintenance peut être assurée par les responsables du modèle afin de garantir sa stabilité.

Pour d'autres, le ML peut être utilisé indirectement, c'est-à-dire pour la construction de modèles ou pour améliorer des modèles bâlois, mais pas en tant qu'algorithme prédictif déployé en production. Des exemples d'une telle utilisation indirecte sont l'extraction de connaissances ou de règles métier (comme le modèle hybride de probabilité de défaut présenté dans le document de réflexion) ou la création de nouvelles variables explicatives.

Impact de l'IA sur le risque de modèle

Un répondant précise que pour un modèle interne (tel que défini réglementairement par la Directive européenne sur les fonds propres réglementaires ou CRD), toutes les composantes de sa gestion (définies dans un cadre de gestion du risque de modèle - *Model Risk Management* ou MRM) sont impactées par l'IA :

- validation indépendante ;
- quantification de l'impact du modèle (sur les résultats, les RWA[8], la non-conformité, le client) ;
- consolidation du risque ;
- simulation du risque (le jeu de scénarios d'évolution des modèles en fonction de facteurs micro- ou macro-économiques étant plus complexe à mettre en œuvre pour de l'IA).

Un autre répondant en déduit qu'un allègement du processus de validation initiale dans le cas de modèles de ML est irréaliste : en effet le cadre de gestion du risque de modèle suivi par la Direction des Risques impose des procédures de contrôle (double regard, revue indépendante, etc.) incompatibles avec un cycle court de validation, même lorsque le risque associé à l'IA est surtout opérationnel et non matériel.

À l'inverse, certains risques induits par l'IA sont négligés par la plupart des cadres de gestion du risque de modèle, notamment la modification du comportement des opérateurs humains engendrée par une prise de décision algorithmique (par ailleurs décrite dans le document de réflexion).

Il est enfin suggéré que le cas particulier des modèles de risque de crédit permet d'éprouver la pertinence du ML sur le terrain, avant de servir de base pour une évolution plus générale des modèles bâlois.

Politique de changement de modèle

Le ML peut, selon certains répondants, être pris en compte dans la politique de changement de modèle de l'établissement du moment que l'adhésion de toutes les parties prenantes est obtenue (métier, *data science*, DSI, équipes réglementaires et juridiques, etc.).

Une proposition exprimée dans le document de réflexion est aussi confirmée : dans la politique de changement de modèle, étant donné que la déclaration de changement doit être faite dès lors que le changement induit est jugé significatif, il n'y a pas de différence entre les modèles statistiques classiques et les modèles de ML.

Il convient néanmoins, selon les réponses reçues, de s'assurer que lors du changement de modèle un *parallel run* du processus est mis en place afin de réaliser un comparatif (*benchmark* sur une base quantitative et avec des points de contrôle bien définis) et de documenter les apports du nouveau modèle.

[8] *Risk-Weighted Assets*, ou actifs pondérés par le risque.

Question 18 : Validation technique

Le processus de validation technique avant mise en production est, selon les répondants, globalement similaire à celle d'un modèle traditionnel, incluant notamment :

- une validation logicielle classique (tests unitaires, d'acceptation, de non-régression) ;
- une analyse de la robustesse de la conception (protocole de choix du modèle et de ses paramètres) ;
- une analyse comparative des distributions de données en apprentissage et en prédiction, de points de données représentatifs, avec identification des anomalies et *outliers* ;
- l'utilisation de modèles challengers de plus en plus fréquent lors de la mise en production (similaire au *A/B testing* en ingénierie logicielle) afin de suivre les écarts de performance au cours du temps, par exemple entre un modèle traditionnel, un modèle complexe mais précis (par exemple combinaison de modèles), et un modèle simple et stable comme une régression linéaire ;
- il faut évaluer performance brute (sur données de test) et performance opérationnelle (latence de la décision ou prédiction).
- quelques points d'attention supplémentaires par rapport aux modèles traditionnels sont la complexité accrue du *feature engineering*, la détection de biais, la sélection des hyperparamètres, l'interprétabilité et explicabilité.

Le monitoring technique au fil de l'eau devrait inclure :

- la surveillance de la qualité des données ;
- le calcul de métriques de stabilité (Population Stability Index ou indices de concentration comme Herfindahl-Hirschman index) ;
- le calcul d'indicateurs de performance et robustesse (RMSE, f1 score, AUC, test de conservatisme pour les modèles internes), de durée du traitement et de latence décisionnelle ou prédictive ;
- la résilience aux pannes (surtout si une architecture distribuée est utilisée) ;
- une analyse de sécurité avec suivi des incidents ;
- une veille technologique (notamment afin de maintenir l'adéquation entre environnement interne et les dernières versions disponibles de composants tiers).

Sécurité et externalisation

Question 19 : Externalisation

De nombreux répondants soulignent que les risques liés à l'externalisation concernent non seulement la conception et l'implémentation d'un algorithme ou service, mais aussi les modèles entraînés et les données utilisées.

Les risques de tiers les plus fréquemment cités peuvent être regroupés dans les catégories suivantes, qui se recouvrent partiellement :

Catégorie	Risque	Exemples et remèdes possibles
Données	Fuite de données	- Le prestataire utilise les données de son client pour entraîner son propre algorithme, ou construit un algorithme sur des règles métiers de son client, pour ensuite le vendre ou le dévoiler à des concurrents - Un tiers fait un usage non autorisé des données, à des fins autres que celles prévues au contrat : amélioration de ses propres produits, utilisation frauduleuse de données bancaires, etc.
	Accès aux données	L'accès aux données doit être garanti (afin de valider leur qualité et leur conformité mais aussi pour leur portabilité), si possible avec complétude des informations sur les ensembles de données d'apprentissage, de test et de validation
Réversibilité	Le prestataire fournit dans certains cas une solution non réversible	
Dépendance au fournisseur (*vendor lock-in*)	Le prestataire est indispensable à la maintenance de l'algorithme	
	La pérennité de l'entreprise externe est incertaine	Le prestataire peut être racheté par un concurrent
	Les coûts financiers pour récupérer les résultats sont prohibitifs	
	Les coûts de maintenance sont prohibitifs	
	Perte possible de reproductibilité	Il faut s'assurer qu'une documentation et qu'un suivi soient faits de façon permanente
	L'accès à l'ensemble de l'environnement de conception et construction, code source inclus, n'est pas garanti	
Dépendance au logiciel	Les droits de propriété liés aux bibliothèques tierce-partie utilisées limitent la réutilisation par le client	
	Utilisation de composants d'IA génériques	- L'utilisation de produits embarquant des composants d'IA génériques ne permet pas généralement de juger de leur qualité et rend opaques les prédiction ou décisions résultantes - La qualité d'une solution généraliste est en général inférieure à celle d'une solution spécialisée
Perte d'information	Manque de disponibilité des équipes du prestataire	
	Transfert de connaissance défectueux	Même en récupérant le code et les données sources, l'équipe interne ne récupère pas la connaissance pour répliquer le modèle ou même juste le rafraîchir
	Défaut d'information plus général du client	
Failles de sécurité	Un risque de sécurité intervient si la SSI chez le prestataire n'est pas à la hauteur de celle du client	
Processus de développement défectueux	Une perte de contrôle sur l'expertise en *data science* et sur la rigueur du processus de développement est possible	Cas où la méthodologie de développement et de test du prestataire n'est pas alignée sur celle du client

Responsabilité	Transfert de responsabilité	En détection de fraude, un déplacement est opéré de la responsabilité du risque fraude au concepteur du modèle, au moins sur un plan contractuel et non juridique
Risque de souveraineté	Concentration sur des plateformes non européennes	
Difficultés dans la reproductibilité	Composants logiciels et services en Cloud	Les versions des composants logiciels ou des modèles peuvent changer fréquemment et sans prévenance, empêchant de rejouer des décisions et inhibant grandement la possibilité de revue indépendante

Note : les deux dernières lignes de ce tableau dénotent des risques spécifiques à l'utilisation de plateformes de type IAaaS ou MLaaS (c'est-à-dire des services outillés, clef-en-main de ML ou IA).

> 💡 De façon générale et en lien avec le principe au cœur du document de réflexion, un répondant souligne que l'externalisation rend plus difficile l'explication. Ainsi l'effet "boîte noire" peut être amplifié, la revue du modèle nécessite alors des méthodes de validation empirique (telles qu'ébauchées dans le document, ou plus classiquement par un stress des inputs sur les modèles internes).

À l'inverse un répondant fait valoir le bénéfice potentiel de l'externalisation : les solutions fournies peuvent être à la pointe des enjeux (nouveaux algorithmes, nouvelles méthodologies, réponse à des enjeux réglementaires, meilleure sécurité). Cela est dû à un investissement plus important et plus focalisé dans la recherche fondamentale, une meilleure compréhension des enjeux technologiques et une captation des meilleurs talents scientifiques.

Question 20 : Sécurité

Les types d'attaque contre le ML le plus couramment mentionnés dans les réponses sont le *data poisoning* (en apprentissage) et les attaques adversarielles ou par usurpation (en prédiction). Elles présentent les principaux risques car elles impactent la fiabilité du modèle, néanmoins leur plausibilité reste contingente à la sécurité du système d'information (SSI) au sens traditionnel.

Elles sont improbables en apprentissage, mais aussi en phase prédictive tant que les modèles sont uniquement accessibles en interne et que les systèmes d'acquisition de données de test sont contrôlés (par exemple un scan de chèque est exclusivement effectué avec un scanner appartenant à la banque, prévenant la manipulation précise des valeurs de pixel possibles dans une image scannée ou enregistrée).

De nombreux répondants s'accordent à estimer que la plupart des techniques d'attaques ne sont pas adaptées aux modèles d'IA, car il faudrait des milliers d'attaques coordonnées pour cartographier le comportement d'un modèle afin de l'induire en erreur (attaques adversarielles) ou de le reconstituer (attaques par inversion).

Pour certains répondants, l'introduction d'IA n'engendre pas de nouveau risque :

- Il faut détecter les intrusions dans un système d'IA comme dans tout SI: le problème racine est la présence d'un intrus dans le SI, indépendamment de son objectif (modifier les données d'entrée d'un modèle, le modèle lui-même ou les sorties).
- En LCB-FT ou en détection de fraude, le risque induit est la connaissance du fonctionnement des algorithmes (via la récupération du code source et des paramètres) et non une attaque contre les données ou le modèle.
- Le principe de base en SSI qui doit aussi être respecté pour l'IA consiste à assurer la traçabilité des opérations et l'auditabilité du système.

La confidentialité des données d'apprentissage du modèle doit être garantie en un sens tout particulier, qu'elles contiennent ou non des données à caractère personnel, pour éviter l'inférence possible du schéma de décision du modèle (attaque par inversion de modèle). Un exemple donné concerne un modèle de risque de crédit augmenté de données comportementales, la variable prédictive la plus importante devenant le nombre de "petits" retraits aux guichets automatiques bancaires sur une temporalité récente : le régulateur local a dû insister (à juste titre) sur la confidentialité de cette information, la seule connaissance des seuils d'impact de cette variable pouvant permettre d'influer sur l'obtention d'un prêt.

Enfin, un cas d'usage à part, présenté par un répondant, relève d'une faille de sécurité sans caractère malveillant : un algorithme de *Reinforcement Learning* est utilisé en aide au conseil en distribution de produits d'assurance. Le modèle s'auto-ajuste en fonction du retour d'expérience (*feedback)* du conseiller ou de l'assuré, il faut donc vérifier l'absence de dérive du modèle en cas de *feedback* erroné ou biaisé (sans intention de nuire, contrairement aux autres failles de sécurité listées ici).

Approche multifactorielle de l'évaluation

Question 21 : Évaluation analytique

Si la documentation des algorithmes et modèles de ML n'est pas standardisée, de nombreux répondants ont mis en place leurs propres formalismes, qui convergent autour des éléments suivants :

- documentation du code source (*versioning* compris) ;
- hypothèses de modélisation ;
- données utilisées pour l'apprentissage et la validation ;
- statistiques descriptives sur le périmètre de l'étude et les données d'apprentissage et de validation ;
- en fonction du cas d'usage, l'évaluateur doit pouvoir être autorisé à accéder aux données et soit à les transférer sur son environnement, soit à se connecter à l'environnement de développement ;
- description des variables utilisées ou créées et de leur traitement ;
- protocole de sélection de modèle parmi les challengers ;
- modèle (type et hyperparamètres) retenu ;
- choix du seuil de décision ou du calibrage ;
- mesure de la performance et vérification de la robustesse ;
- mécanismes explicatifs globaux du modèle (contribution et interactions entre variables) ;
- éléments d'explication locale des résultats (LIME, SHAP, etc.) ;
- traçabilité des échanges avec le métier justifiant un choix technique donné ;

L'hétérogénéité des fonctionnalités de documentation offertes par les plateformes de ML est notée au passage : un équilibre doit être trouvé entre facilité de documentation des plateformes de type glisser-déplacer et de celles permettant une utilisation avancée mais offrant moins de capacité d'auto-génération de documentation.

Un répondant souligne que pour un modèle soumis à validation indépendante ou audit externe, il convient de mettre à disposition tous les éléments permettant de rejouer le modèle et d'en mesurer les performances, ainsi que la documentation associée. Cela signifie donc que tous les éléments de la liste précédente doivent être fournis.

Question 22 : Évaluation empirique

Données de *benchmarking*

Allant à l'encontre du document de réflexion de l'ACPR qui suggère l'utilisation de jeu de données de *benchmarking* dans une situation générale d'audit, plusieurs répondants adressent des critiques concernant leur emploi en audit externe, tout en reconnaissant leur adéquation à un audit interne :

- questionnement de la pertinence d'indicateurs de performance utilisant des données synthétiques (plutôt qu'utiliser les jeux de test *out-of-time* et *out-of-sample* du client) ;

- questionnement sur l'existence de biais du test de *benchmarking* utilisant des données propres au superviseur (surtout si elles sont plus volumineuses et plus profondes historiquement que les données de l'établissement) : cela ne permettrait pas une comparaison juste, ni informative sur l'optimalité de la modélisation effectuée par l'établissement bancaire étant donné les éléments à sa disposition. L'objectif devrait plutôt être de déterminer si le modèle produit est optimal ou non vis-à-vis des données dont dispose l'établissement ;
- le schéma de données, mais aussi leur sémantique, sont des obstacles pratiques majeurs à la création de jeux de données de benchmarking

Mise en concurrence de modèles

Quant à l'utilisation de modèles *challengers* :

- En audit externe, elle risquerait tout comme les jeux de données de *benchmarking* de ne pas être informative[9].
- En audit interne, la faisabilité semble limitée par les contraintes d'allocation de ressources humaines, matérielles et de temps. De plus, la réglementation imposant des revues périodiques sur l'ensemble du cycle de vie des modèles, cela inclurait la revue de tous les modèles *challengers* mis en œuvre.
- Un répondant expérimenté dans l'IA sur des données non structurées indique que l'implémentation de modèles *challengers* est faisable sur des données tabulaires, mais l'approche par *benchmarking* est préférable pour des données textuelles
- Un autre répondant estime que l'implémentation de modèles challenger est illusoire dans un contexte où les ressources de l'auditeur sont limitées : il est improbable qu'un modèle développé plus rapidement (et sans forcément accès aux données du modèle audité) soit plus performant que celui développé en interne, typiquement sur plusieurs mois.
- Une autre difficulté, évoquée dans le document de réflexion, est reprise parmi les réponses : il n'existe pas d'outillage prédéfini (encore moins standardisé) permettant d'évaluer un modèle en dehors de l'environnement utilisé pour sa création. Étant donné la diversité des technologies et architectures utilisées, l'intégration serait nécessairement ardue et longue : en d'autres termes, la portabilité d'un modèle en dehors de son environnement est une question particulièrement complexe et ouverte.
- L'analyse de la cohérence (vis-à-vis de résultats attendus et dictés entre autres par les exigences réglementaires, avec un taux d'acceptation des écarts par rapport au modèle de référence) serait finalement moins complexe à mettre en en œuvre pour l'évaluation que des modèles challengers, qui impliquent souvent de passer du temps à expliquer les écarts de résultats.

Un répondant souligne une limite très concrète à l'audit de ML par le superviseur : dans un contexte *Big Data*, l'accès nécessaire à certaines ressources (clusters de calcul et de stockage) peut être contraint par les besoins en ressources IT. Ainsi, l'exécution en parallèle sur un serveur d'exploitation est généralement très limitée.

Question 23 : Méthodes explicatives

Méthodes utilisées

Les méthodes explicatives listées par le document de réflexion présentent un degré d'adoption très variable :

- toutes les méthodes explicatives pré-modélisation mentionnées sont utilisées par un ou plusieurs répondants ;
- aucune méthode explicative conjointe à la modélisation n'est utilisée ;
- parmi les méthodes explicatives post-modélisation mentionnées, les plus couramment citées sont les PDP (*Partial Dependency Plots*), LIME (dont plusieurs répondants remettent en question la fiabilité sur des modèles non-triviaux), SHAP et ses variantes, et les *Global Surrogate Models*.

[9] C'est-à-dire qu'on pourrait supposer que la performance est moindre que celle du modèle audité, sans pouvoir en tirer de conclusion.

Quelques méthodes sont en cours d'étude : celles basées sur l'extraction de règles logiques comme *Anchors* (Ribeiro, 2018), et une méthode d'explicative contrefactuelle, MACEM (*Model Agnostic Contrastive Explanation Method*) (Dhurandhar, 2019).

Les autres méthodes d'explication utilisées par les répondants sont :

- Les mécanismes basés sur l'attention (*Attention Maps* ou *Heatmaps*)
- ProtoDash : cf. Gurumoorthy, 2019
- ALE (*Accumulated Local Effects*) et ICE (*Individual Conditional Expectation*) : cf. Molnar, 2020
- BreakDown : cf. Staniak, 2018
- IG (*Integrated Gradients* : Sundararajan, 2017) et GIG (*Generalized Integrated Gradients* : Merrill, 2019)
- *Deep Taylor Decomposition* : cf. Montavona, 2017

Une association professionnelle représentant un groupe d'acteurs rapporte que relativement peu de méthodes d'explicabilité sont observées en production, elles sont plutôt appliquées en construction et principalement pour des travaux de connaissance client, et il s'agit souvent de LIME ou SHAP.

Acceptabilité d'une explication

Un répondant estime que pour les modèles les plus sophistiqués, il est parfois nécessaire de renoncer à expliquer (étymologiquement : déplier) le modèle et se contenter que les résultats soient justifiés (par un raisonnement humain) ou acceptés (car l'utilisateur final considère qu'ils sont les meilleurs ou les moins mauvais possibles, dans un contexte de « rationalité limitée »).

Les explications contrefactuelles sont jugées par un répondant comme non désirables dans le cadre des modèles de risque client en octroi de crédit, lorsqu'il s'agit de fournir des explications aux clients ou aux conseillers. Les raisons invoquées sont :

- la pratique de non-divulgation des méthodes de calcul des décisions d'octroi (afin d'éviter les risques de contournement et de falsification de données déclaratives) ;
- le faible niveau d'acceptabilité par le client si l'explication est centrée sur une ou deux variables (par exemple l'ancienneté client ou *tenure*) en comparaison d'une explication générique sur le fonctionnement de l'algorithme accompagnée du positionnement du client sur l'échelle min-max du score d'octroi ou de la probabilité de défaut.

Un répondant du secteur assurantiel précise que fournir une explication locale, fidèle et de nature technique, telle quelle à l'utilisateur métier ou à l'assuré est souvent insatisfaisant. En effet l'explication est rarement exploitable, parfois découplée du bon sens commun, voire incompatible avec la réglementation en matière de devoir de conseil. L'alternative préférable consiste à proposer une explication basée sur des moments de vie, le but étant d'allier puissance prédictive à des explications contextualisées et directement exploitables.

> 💡 Les méthodes explicatives peuvent être plus ou moins techniques/implicites. Une étude d'expérience utilisateurs a été menée par un répondant afin de comprendre les différents niveaux de besoin en terme d'explicabilité et interprétabilité d'un modèle de ML, puis d'identifier les méthodes explicatives les plus adaptées pour les différents profils d'utilisateurs.

Un répondant pointe le danger de méthodes explicatives efficaces : les explications *séduisent trop facilement l'opérateur humain*. Ainsi les explications les plus convaincantes sont celles qui "racontent une histoire" quitte à perdre en complétude ou en fidélité par rapport au modèle. On peut dès lors se demander si la meilleure explication est simplement l'explication la plus intelligible et la plus persuasive possible.

À noter que la littérature académique sur l'efficacité des méthodes d'explications auprès d'utilisateurs réels (et non d'informaticiens) n'est pas très fournie. On peut donc s'attendre à des évolutions importantes sur ce sujet dans les années à venir.

Enfin, un exemple est rapporté dans lequel l'enjeu d'explicabilité est d'ordre surtout pédagogique. Il s'agit d'un projet NLP ayant recours à un modèle hybride : réseau neuronal profond combiné à un module de *clustering* entraîné sur les erreurs de classification et intrinsèquement explicable. L'enjeu d'explicabilité est alors lié à l'appropriation par l'humain d'un nouveau processus confié à l'algorithme, où l'opérateur « voit » comment ses décisions passées améliorent l'efficacité du système.

Réglementation

Bien que le cadre réglementaire relatif à l'utilisation d'IA dans le secteur n'ait pas fait l'objet d'une question particulière, de nombreux acteurs ont exprimé leur position dans ce domaine. Les points suivants sont particulièrement notables.

De nombreux répondants pensent que la réglementation de l'IA doit s'inscrire dans un cadre juridique européen harmonisé, et que tous les acteurs concernés doivent être mis sur un pied d'égalité (« *level-playing field* »).

Certains répondants estiment qu'une certification ou « labélisation » des briques d'IA externes pourrait également être un avantage.

Est également noté l'intérêt de privilégier une économie de moyens, d'une part en capitalisant sur les cadres existants - lesquels recouvrent substantiellement les thèmes majeurs du document de réflexion (traitement adéquat des données, externalisation, etc.), d'autre part en évitant une multiplication normative qui serait préjudiciable à bien des égards.

Les répondants du secteur de l'assurance considèrent que de par sa complétude, le corpus réglementaire sectoriel est à même d'intégrer et d'encadrer les évolutions induites par l'IA sans nécessité de recourir à une réglementation spécifique. Ils préconisent toutefois une réflexion - dans la lignée de la présente consultation - sur les moyens de faciliter cette intégration et d'éviter les incertitudes réglementaires, sources d'insécurité juridique et financière.

Concernant le rôle du régulateur, un répondant estime que la mesure d'efficacité de l'IA doit prendre en compte ses apports économiques et sociaux, notamment dans un domaine tel que la LCB-FT où ces bénéfices sont externalisés. Ces bénéfices devraient ainsi être comparés à ceux des systèmes en place : si l'IA apporte un net avantage, l'autorité de régulation devrait inciter les acteurs à l'adopter ; si en revanche les gains sont plus modestes elle devrait rester technologiquement neutre afin que les acteurs établissent leur propre analyse de rentabilité.

Concernant les approches par obligation de résultats ou de moyens, un répondant rappelle que la plupart des algorithmes d'IA aujourd'hui utilisés et efficaces ont été trouvés par tâtonnements et essais-erreurs. La méthode est très différente de celle d'un régulateur qui décrit des moyens à mettre en œuvre, des contraintes à respecter, et des procédures à suivre. L'évolution des réglementations vers une obligation de résultats, si elle s'accompagne d'un allègement d'obligation de moyens, devrait bénéficier à l'innovation par l'IA, dans un esprit de gouvernance par les objectifs.

Considérations pour le superviseur

Certaines suggestions relatives aux bonnes pratiques de gouvernance de l'IA en finance, à la méthodologie d'évaluation des algorithmes, ou à l'évolution du cadre réglementaire ressortent des réponses à la consultation. Les plus saillantes sont résumées ici, en soulignant qu'elles reflètent les retours du marché et non les positions de l'ACPR.

Pour l'évaluation (selon une approche analytique entre autres) il serait utile de mettre en place au niveau des autorités compétentes un référentiel d'évaluation adapté à chaque catégorie d'organisation ou type d'algorithme d'IA (par activité, par taille, par type de données…)

Plusieurs répondants soulignent le principe de neutralité technologique de la réglementation, permettant de s'assurer que les systèmes d'IA respectent les principes et règles transversales déjà identifiés dans les différentes

réglementations portant sur la conception et l'usage de tout type de technologie, sans imposer de règles spécifiques à l'IA.

> 💡 Le principe de neutralité technologique leur paraît d'autant plus pertinent que les algorithmes d'IA sont le plus souvent embarqués ou encapsulés dans des systèmes englobants, plus complexes et déjà régulés. Par exemple, une interface de conseil en investissement encapsulant l'IA se doit de respecter la réglementation MiFID. Ainsi l'utilisation d'IA ne devrait pas en elle-même augmenter les exigences d'explicabilité, lesquelles doivent rester fondées sur les risques et les cas d'usage, et non sur la technologie.

D'autres répondants notent qu'une approche par les risques, transsectorielle et avec une portée extraterritoriale permet de limiter les distorsions de concurrence, d'autant plus que la frontière entre secteurs d'activité est quelquefois difficile à établir. En particulier, les intermédiaires et prestataires de services de paiement, qui ne sont pas régulés comme les établissements de crédit, devraient être soumis en matière d'IA aux mêmes conditions réglementaires que ces derniers.

> 💡 Un répondant se demande enfin si les bonnes pratiques de gouvernance, notamment en matière d'explicabilité, sont aussi applicables aux outils et modèles développés par le superviseur. Par exemple si lors d'une mission de contrôle sur place, un modèle *challenger* tel que décrit dans le document de réflexion est mis en place pour le filtrage de transactions (en Sécurité Financière) et détecte de nouveaux cas suspects, cela impose-t-il à l'autorité de contrôle une exigence d'explicabilité pour prouver que ces nouveaux résultats sont valides ?

Bibliographie

D'intéressantes publications, souvent très récentes et reflétant l'état de l'art, ont été citées par les répondants.

Valérie Beaudouin, Isabelle Bloch, David Bounie, Stéphan Clémençon, Florence d'Alché-Buc, et al. *Identifying the "Right" Level of Explanation in a Given Situation.* (2020)

Amit Dhurandhar, Tejaswini Pedapati, Avinash Balakrishnan, Pin-Yu Chen, Karthikeyan Shanmugam, Ruchir Puri. *Model Agnostic Contrastive Explanations for Structured Data.* arXiv:1906.00117 [cs.LG] (2019)

K. Gurumoorthy, A. Dhurandhar, G. Cecchi. *ProtoDash: Fast Interpretable Prototype Selection.* arXiv:1707.01212v4 [stat.ML] (2019)

Fabian Hinder, Barbara Hammer. *Counterfactual Explanations of Concept Drift.* (2020)

D. A. Melis, T. Jaakkola. *Towards robust interpretability with self-explaining neural networks.* In Advances in Neural Information Processing Systems (2018)

John Merrill, Geoff Ward, Sean Kamkar, Jay Budzik, Douglas Merrill. *Generalized Integrated Gradients: A practical method for explaining diverse ensembles.* arXiv:1909.01869 (2019)

Grégoire Montavona, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, Klaus-Robert Müllerad. *Explaining nonlinear classification decisions with deep Taylor decomposition.* Pattern Recognition, Volume 65 (2017)

Marco Tulio Ribeiro, Sameer Singh, Irvine. *Anchors: High-Precision Model-Agnostic Explanations.* AAAI (2018)

C. Rudin. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.* Nature Machine Intelligence (2019)

M. Staniak, P. Biecek. *Explanations of model predictions with live and breakDown packages* (2018)

Mukund Sundararajan, Ankur Taly, Qiqi Yan. *Axiomatic Attribution for Deep Networks.* arXiv:1703.01365 [cs.LG] (2017)

J. Tang, Jian Yin. *Developing an intelligent data discriminating system of anti-money laundering based on SVM.* 2005 International Conference on Machine Learning and Cybernetics (2005)

Mark Weber, Giacomo Domeniconi, Jie Chen, Daniel Karl I. Weidele, Claudio Bellei, Tom Robinson, Charles E. Leiserson. *Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics.* arXiv:1908.02591 (2019)