Décembre 2020

Gouvernance des algorithmes d'intelligence artificielle dans le secteur financier

Synthèse des réponses à la consultation

Auteur : Laurent Dupont

Pôle Fintech-Innovation, ACPR

Les 26 réponses écrites à cette consultation de l'ACPR proviennent d'établissements bancaires, de prestataires technologiques, de cabinets de conseil, d'associations professionnelles, d'instituts de recherche. Les réponses écrites couvrent, pour la plupart, l'ensemble des questions, et ont été complétées pour la présente synthèse par une dizaine d'échanges téléphoniques. Ces réponses confortent les principes énoncés dans le document de réflexion, et permettront de préciser les axes de travail futur du Pôle Fintech-Innovation.

Expérience et organisation en intelligence artificielle (IA) et en particulier en *machine learning* (ML)

Les acteurs du secteur ont des compétences internes en IA/ML, complétées par des initiatives transverses de formation et de sensibilisation au ML. Les experts IA sont regroupés dans des structures dédiées décentralisées auprès des métiers, ou de façon plus centralisée ou encore par entité. Les profils vont du monde universitaire à des praticiens expérimentés, en passant par la recherche appliquée.

Algorithmes et cas d'usage

Les algorithmes d'IA mis en œuvre par les répondants sont plus diversifiés que prévu : modèles de type *Gradient Boosted Trees* très couramment utilisés, *Deep Learning* plus marginal, méthodes non-supervisées (*clustering* inclus), et surtout de nombreuses méthodes standard ou ad-hoc d'assemblage de modèles.

Les catégories de cas d'usage le plus fréquemment citées sont la productivité interne (par exemple analyse automatique de documents), la relation client, des usages très spécifiques en aide à l'investissement, et plusieurs processus de la chaîne de valeur en assurance[1].

Un point d'attention est le nombre de cas d'usage en traitement du langage naturel (NLP), surtout pour les processus internes de gestion, tandis que le document de réflexion se focalisait sur les processus métier critiques.

Principe d'explicabilité

Un intérêt majeur et croissant pour l'explicabilité ressort des réponses, pour des motifs de performance, de conformité réglementaire, d'auditabilité, et de facilitation de l'adoption d'IA en interne.

[1] Pour une description plus détaillée des cas d'usage dans le secteur financier, on pourra se reporter au précédent document de réflexion de l'ACPR de 2018 : « Intelligence artificielle : enjeux pour le secteur financier ».

Les 4 niveaux constituent selon les répondants une graduation intéressante car qualitative, et ils sont généralement en phase avec les définitions proposées hormis pour le plus élevé (N4) considéré par certains comme inatteignable en pratique et relevant davantage de l'assurance-qualité que d'explicabilité. La frontière entre N2 et N3 fait parfois aussi l'objet de questions de clarification : à cet égard, une mise en situation de chaque niveau sur un cas d'usage donné et en lien avec les algorithmes envisageables est jugée utile.

Les répondants sont en accord avec les deux facteurs de choix du niveau d'explication (la criticité du cas d'usage et les destinataires de l'explication), certains en ajoutant d'autres comme le périmètre d'application du modèle ou la robustesse des explications produites.

Quant aux exemples pratiques de niveaux d'explication, les quelques propositions d'ajustement consistent à relever le niveau exigé : pour l'agent de conformité en LCB-FT, pour la validation et le contrôle interne des modèles bâlois, et pour le consommateur final (où N2 pourrait être exigible *a minima*).

L'absence d'exemples impliquant du NLP est de nouveau déplorée, alors même que l'utilité d'une explication sur un modèle de NLP est perçue comme limitée, le contrôle de l'algorithme se faisant par ses résultats. Également, les usages avancés de l'IA en LCB-FT (analyse de graphes, apprentissage non supervisé) méritent de figurer dans de futures études.

Enfin, l'articulation des exigences d'explicabilité avec d'une part les contraintes réglementaires, d'autre part les enjeux sociaux tels que la non-discrimination, devrait, selon certains répondants, être développée.

Principe de performance

Les répondants ont largement enrichi la liste de mesures techniques de performance de l'IA fournie à titre illustratif dans le document de réflexion, afin de couvrir un plus large spectre de problèmes (classification, régression, traitement du langage naturel) et de cas d'usage.

Est aussi soulignée la nécessité de distinguer les métriques destinées aux *data scientists* et visant à optimiser le modèle de celles destinées aux experts métier, devant leur être adaptés et facilement interprétables.

Parmi les nombreuses métriques fonctionnelles de performance, on peut distinguer les métriques d'efficacité normative (par exemple, en sécurité financière, la couverture des scénarios identifiés ou la prévention du risque d'adaptation des fraudeurs) et les métriques d'efficacité opérationnelle (temps de traitement et qualité des alertes, risque opérationnel lié à la décision concernant une alerte, etc.)

Principe de stabilité

Les sources de dérive le plus souvent citées concernent les données en entrée du modèle, son réentraînement, mais aussi les environnements technique, réglementaire (nouvelles règles de filtrage, nouveaux services proposés, nouveaux clients, etc.) et économique. Les risques associés sont une baisse de performance, l'introduction de biais, un risque financier.

Les techniques de remédiation des dérives de modèle proposées sont de nature classique : suivi continu d'indicateurs statistiques de cohérence, détection préventive de problèmes de généralisation, résolution de ces problèmes (par échantillonnage rigoureux, simplification du modèle, *Transfer Learning*, réentraînement non systématique et à fréquence limitée, etc.) et, comme palliatif, historicisation et réversibilité des modèles.

Principe de traitement adéquat des données

Peu de réponses sur ce point sont spécifiques à l'IA. Au plan organisationnel, les réponses préconisent généralement que les services juridiques et les équipes dédiées à la protection et à la sécurisation des données se concertent avec les data scientists pour vérifier la conformité du cas d'usage, la finalité des modèles, l'utilisation des données, les conséquences possibles du traitement et l'existence de garanties appropriées.

Une tension est aussi soulignée entre les exigences réglementaires sur la durée de rétention des données et le besoin de conserver les logs de décisions et les données utilisées afin de produire des explications individuelles.

Les répondants utilisent un panel de méthodes de **détection de biais** : *back-testing* à la validation, méthodes explicatives, revue manuelle, suivi d'indicateurs statistiques en production. Un flou est noté concernant la définition de métriques d'équité (*fairness*), appelant un éclairage par les autorités compétentes.

Si la **remédiation des biais de données** semble familière aux acteurs du secteur, celle des **biais de modèles** est peu maîtrisée, limitée aux variables disponibles, et même considérée comme superflue dans le cas de modèles prédictifs dont l'explicabilité est assurée.

Intégration dans les processus

Les répondants relèvent une très grande variété de critères d'**évaluation de l'intégration** d'IA : productivité interne, relation client, interactions humain-machine (incluant explicabilité et droit à la contestation des résultats), facilité de mise en production et de maintenance, réversibilité et moyens de remédiation des erreurs, conformité réglementaire. Une distinction est proposée entre deux scénarios-types : soit l'IA apporte de la précision mais pas un changement radical du processus métier, soit celui-ci est bouleversé par l'IA, induisant un risque de rejet par le métier ou d'intégration « forcée ».

Un désaccord émerge des réponses quant à l'**autonomie de l'IA**: les humains doivent-ils être incités à remettre en question les résultats et à conserver leur libre arbitre vis-à-vis des algorithmes, ou à l'inverse cette autonomie génère-t-elle des prises de risque supplémentaires (par exemple par forçage humain des décisions) ?

Les processus parallèles confiés à l'humain en vue d'évaluer l'IA en continu ne font pas non plus l'unanimité, en raison de leur surcoût financier et en ressources et de leur difficulté de mise en œuvre.

Selon les réponses, la **méthodologie de conception de l'IA** devrait évoluer dans la même direction générale que l'industrie logicielle, en particulier vers un meilleur contrôle qualité et une robustesse assurées par des principes de conception de type « *MLOps* ». Est toutefois notée une différence essentielle entre l'IA et le logiciel traditionnel : la conception d'IA et surtout de ML procède par essais-erreurs, partant de solutions concurrentes à implémenter en parallèle afin de jauger leur efficacité sur les processus de décision. Leur évaluation doit donc être itérative et procéder par une méthodologie expérimentale éprouvée plutôt que par preuve formelle.

Contrôle interne

Les réponses confortent l'idée que les **risques pour les processus métiers** restent globalement les mêmes avec l'IA que pour les modèles statistiques classiques : si le risque opérationnel est généralement amplifié, les modèles de ML rentrent néanmoins dans la gouvernance interne de gestion du risque de modèle et les processus de contrôle interne s'adaptent déjà à l'évolution des usages et des technologies. De nouveaux risques réputationnels et juridiques seraient toutefois engendrés par des décisions automatiques à caractère discriminatoire, soulevant la question de la **responsabilité dans les processus décisionnels** impliquant de l'IA.

Les méthodes de remédiation des risques liés à l'IA sont alignées avec les 4 principes de conception exposés dans le document de l'ACPR, auxquels s'ajoutent les bonnes pratiques usuelles (cartographie des risques de modèle, procédures de rétablissement du contrôle humain) et une implication accrue du département juridique en accompagnement et encadrement des projets d'IA.

Les répondants confirment l'importance de la **validation fonctionnelle** (proportionnée au niveau de criticité des processus en jeu) tout au long du cycle de vie de l'IA, requérant au passage le développement de compétences en IA parmi les directions de contrôle interne et les responsables d'entités de groupes bancaires ou assurantiels.

Une tension est soulignée entre le besoin d'auditabilité (totale explicabilité) des **modèles internes** dans un secteur réglementé et la promesse d'hypothèses minimalistes et d'auto-adaptation faite par l'IA. Les positions divergent sur les modalités d'utilisation du ML pour les modèles bâlois : pour certains, un algorithme prédictif de ML pourrait être déployé directement en production pour la modélisation des risques (sur un périmètre bien défini, avec des bénéfices démontrés, et des garanties offertes sur les 4 principes énoncés dans le document de réflexion de l'ACPR), pour d'autres, le *machine learning* ne pourrait être utilisé qu'indirectement pour l'amélioration de modèles bâlois (ajout de nouvelles variables explicatives, ajustement de règles métiers, etc.).

Les relations entre la **gestion du risque de modèle** (MRM ou *Model Risk Management,* imposant des procédures de contrôle incompatibles avec un cycle court de validation) et l'introduction d'IA (qui introduit quelques risques souvent absents du MRM comme la modification du comportement humain) apparaissent à certains répondants comme une lacune dans le document de réflexion. Est en revanche confirmée la proposition du document de l'ACPR selon laquelle l'utilisation de ML ne remet pas fondamentalement en question la **politique de changement de modèle**.

Les processus de **validation technique** et de monitoring continu sont selon les répondants globalement similaires aux modèles traditionnels. Quelques points d'attention supplémentaires sont toutefois relevés : complexité accrue des prétraitements de données, détection de biais, sélection des hyperparamètres, et la surveillance des 4 critères exposés dans le document de l'ACPR.

Sécurité et externalisation

Les répondants confirment et étendent le domaine de **risques liés à l'externalisation** d'IA, y incluant l'accès aux données et le risque de fuites, la dépendance au fournisseur (*vendor lock-in*) mais aussi au logiciel, le défaut d'information et la perte de compétences, les processus de développement et de sécurité défaillants chez le prestataire, et le transfert de responsabilité. Dans le cas du *cloud* s'y ajoutent le risque de souveraineté et de non-reproductibilité. L'externalisation, en amplifiant l'effet "boîte noire", rend aussi plus difficile l'explication de l'IA.

Les répondants font écho aux types d'**attaques contre le ML** mentionnés dans le document de l'ACPR, tout en les considérant peu plausibles dans un environnement technique de production typique du secteur financier. Les failles de sécurité habituelles sont jugées plus inquiétantes, telle l'intrusion dans un système informatique, ou encore la connaissance du fonctionnement d'un algorithme de détection de fraude à des fins de contournement.

Approche multifactorielle de l'évaluation

De nombreux répondants ont mis en place leurs propres méthodes d'**évaluation analytique**, centrées sur la documentation exhaustive des algorithmes mais aussi des données et du processus de développement de l'IA, couvrant là aussi les 4 principes détaillés dans le document de l'ACPR. Un processus d'évaluation idéal devrait pouvoir rejouer le modèle et mesurer ses performances, surtout s'il est soumis à validation indépendante ou audit externe.

L'utilisation de **données de *benchmarking*** est jugée adéquate en audit interne mais fait l'objet de nombreuses objections quand elle est envisagée pour un audit externe (possible existence de biais, questions liées au recours à des données synthétiques ou problèmes dus à l'ignorance de la sémantique des données métier).

Les positions sont plus ambivalentes sur les **modèles *challengers*** : leur utilisation est souvent considérée comme peu informative, très consommatrice des ressources (humaines et matérielles) de l'auditeur et des ressources informatiques de l'entité auditée, et limitée par une absence de standardisation – comme noté dans le document de réflexion. Une méthode par analyse de cohérence représente pour certains répondants une bonne alternative : moins complexe à mettre en œuvre, elle fournirait des résultats plus facilement exploitables que la mise en concurrence de modèles.

Les **méthodes explicatives** les plus couramment utilisées sont les méthodes pré-modélisation, ainsi que certaines des méthodes post-modélisation mentionnées par le document de l'ACPR, auxquelles s'ajoutent une grande variété de méthodes à l'état de l'art, souvent issues de travaux de recherche très récents. La plupart sont toutefois utilisées à titre expérimental, quelques-unes en construction de modèles ou pour des travaux de connaissance client, et très peu en production. Les répondants semblent réservés vis-à-vis des explications contrefactuelles promues par le document de réflexion, qui posent des défis de mise en œuvre (pratique de non-divulgation de méthodes de calcul, acceptabilité douteuse par le client).

Les explications contextualisées et concrètes, alignées avec le « bon sens commun », tout en étant conformes à la réglementation, sont préférées par de nombreux répondants aux explications locales techniquement plus fidèles au modèle et non transformées.

Réglementation

La réglementation de l'IA doit selon les répondants s'inscrire dans un cadre juridique européen harmonisé, assurant un « *level-playing field* » entre acteurs, éventuellement avec une certification des briques d'IA externes, et toujours selon une approche par les risques.

L'approche normative n'est pas jugée nécessaire, voire est parfois considérée comme préjudiciable, le corpus réglementaire sectoriel étant à même d'encadrer les évolutions induites par l'IA sans nécessité de recourir à une réglementation spécifique. Pour certains répondants, toutefois, des clarifications réglementaires permettraient de combler des zones d'insécurité juridique. Enfin, l'évolution des réglementations vers une obligation de résultats, tout en maintenant une neutralité technologique essentielle, devrait bénéficier à l'innovation par l'IA.