Impact de l'Intelligence Artificielle (IA) sur les bio-industries

27 janvier 2025

L'Équipe R&D de Qualees (**salma BARKAOUI** – Head of Data Sciences-, **Ivan Ignatiev** – CTO-, et moi-même– Directeur R&D-) a été invitée par l'équipe organisatrice du « Research Day » (Vendredi 24 janvier 2025) de l'École de Biologie Industrielle (EBI), composée notamment de **Delphine HERMOUET, PhD**, **Jad Eid**, et **David Jung**, pour participer à une table ronde intitulée :

« Impacts de l'intelligence artificielle sur les bio-industries »

Thèmes abordés

1. **IA comme accélérateur de la recherche**
2. **IA et optimisation des processus industriels / logistiques** (efficacité, qualité, traçabilité)

Intervenants clés

- **Mme Sophie HAMELIN Hamelin (L'Oréal)** – Transformation digitale, « chercheur augmenté »
- **M. Stephane Menio (Safran Landing Systems)** – Directeur R&D
- **M. Lionel Pelletier (Aktehom)** – Intégrité des données, intelligence réglementaire
- **M. Fabrice Ruiz (Clinsearch, administrateur EBI)** – Animateur de la table ronde

Cette diversité de profils de grande qualité a permis d'explorer différents points de vue sur les apports et limites de l'IA dans les bio-industries.

Nous n'aurons pas eu le temps d'aborder le second thème du fait des questions passionnées et passionnantes soulevées par le traitement du premier thème. Nous pourrons l'aborder dans un article ultérieur.

2. L'importance cruciale des données

2.1. « Garbage in, Garbage out »

Il est inconcevable d'aborder l'intelligence artificielle sans insister sur la qualité et la quantité des données nécessaires. Même les modèles les plus avancés (ex. GPT LLM) ne peuvent fournir des résultats pertinents que si les données sont fiables, nettoyées et représentatives.

L'Adage « Garbage in, Garbage out » quelque peu occulté par la puissance perçue des modèles de Generative Pre Trained Large Language Model (GPT-LLM) et l'omerta régnant sur les jeux de données ayant servi à leur entrainement reste plus que jamais valable à l'ère de l'Intelligence Artificielle.

2.2. Certification et réglementation

Dans des secteurs fortement réglementés comme la santé, la certification des algorithmes et la conformité (RGPD, IA Act) sont des enjeux majeurs. À ce jour, le processus d'entraînement des modèles (LLM compris) est souvent opaque :

- **Secret industriel**
- **Ombres sur la provenance** (risque de violation du droit d'auteur, « vol » de données, etc.)
- **Régulation européenne stricte vs. dérégulation américaine**

La **qualité des données d'entraînement** et leur audit sont des aspects souvent passés sous silence. Pourtant, il est indispensable d'évaluer la qualité intrinsèque des jeux de données pour certifier un modèle.

2.3. Domaine d'application et variabilité des données

Comme le souligne **salma BARKAOUI**, l'évaluation de la qualité dépend fortement du domaine d'application. Dans le médical, la collecte des données est moins standardisée qu'en industrie (capteurs IoT, etc.), et leur hétérogénéité plus marquée.

- **« Propre » ≠ immédiatement exploitable** : Il faut préparer (encodage, imputation, détection des biais, etc.) les données, même si un data owner les considère comme étant déjà « propres ». Le développement des Entrepôts de

Données de Santé (EDS) avec des points de service sous forme d'API devrait améliorer l'exploitabilité et l'interopérabilité des données.

- **Méthodologies d'évaluation** : De la même manière qu'il existe des benchmarks pour les algorithmes, il est nécessaire d'élaborer des méthodes de référence pour juger la qualité des datasets. Il serait également souhaitable que ces méthodes d'évaluation fassent consensus et soient établies comme des standards (pour permettre la comparaison des jeux de données).

2.4. Contrôle qualité : la donnée comme matière première

De la même manière qu'un industriel contrôle la qualité de sa matière première, il est impératif d'évaluer et de corriger la qualité des données avant tout entraînement d'un modèle. C'est essentiel pour **industrialiser** la production d'IA, notamment dans la conception de **jumeaux numériques** (ex. TweenMe by Qualees).

- **Pré-qualification** : Identifier si les données peuvent être améliorées par un prétraitement (standardisation, filtrage, correction).
- **Cas limites** : Mieux vaut renoncer à l'entraînement que de bâtir un modèle biaisé ou peu robuste.
- **Notamment dans le cas d'algorithmes de production de données synthétiques** qui vont produire les données d'entraînement d'autres IA et éventuellement réduire l'efficacité ou la pertinence de leur entrainement.

2.5. Impact écologique et énergétique

Enfin, Salma Barkaoui rappelle que l'entraînement d'un modèle IA dépend lourdement de la qualité des données sur le plan **énergétique**. Des données mal préparées se traduisent par un gaspillage de ressources (calcul, électricité, refroidissement), sans réelle valeur ajoutée.

Consommations (exemples pour l'entrainement des LLM)

- GPT-3 : ~1 300 MWh (eq. consommation annuelle de 240 foyers français)
- GPT-4 : ~3 000 MWh (eq. 500 foyers/an)
- BLOOM : ~433 MWh

Consommation d'eau : Liée au refroidissement des serveurs (ex. GPT-3 aurait utilisé ~700 m^3 d'eau). Liée également à la production d'électricité d'origine nucléaire (estimée à 3L par kWh produit).

Tendance : La puissance de calcul nécessaire pour entraîner les modèles les plus avancés est multipliée par 4 à 5 chaque année (sources Epoch AI).

Coûts financiers et disponibilité du matériel

- Difficultés pour les universités (même prestigieuses) de suivre la demande en GPU.
- Concentration de la ressource au profit des GAFAM (ex. 1,8 million de GPU chez Microsoft contre 300 chez Stanford).
- Cela pose un problème de "démocratisation" de l'accès à l'IA par la concentration des moyens dans les mains d'une oligarchie industrielle.

Enjeux d'approvisionnement en électricité

- De nouveaux centres de données gourmands (1 MW pour 1 000 GPU A100 de NVIDIA)
- Recours à la relance de centrales (ex. Three Mile Island) ou à des **SMR** (Small Modular Reactors).
- Empreinte hydrique importante (1 à 1,5 L d'eau par kWh pour le refroidissement des serveurs).

Approche Qualees

- Choix d'**IAs "compactes"** et spécialisées, nécessitant beaucoup moins de ressources qu'un LLM.
- Notre cluster Kubernetes consomme ~500 kWh/an en fonctionnement continu.

3. Spécificités du domaine médical : les données HDLSS

3.1. Problématique HDLSS

Les données médicales sont souvent **High Dimensionality, Low Sample Size** (HDLSS) :

- **Beaucoup de variables** (génétique, imagerie, biomarqueurs, dossiers cliniques)
- **Peu de lignes** (quelques milliers de patients tout au plus)

Cette « malédiction dimensionnelle » complexifie la mise au point de modèles performants.

3.2. Opposition avec les LLM

Contrairement aux **LLM** (qui s'appuient sur des volumes textuels massifs (plusieurs pétaoctets), souvent "unidimensionnels" car se limitant à des séquences de mots), les données médicales :

- **Sont multimodales** (texte, imagerie, biologie, etc.)
- Nécessitent **une expertise clinique** pour être prétraitées et standardisées.

3.3. Prétraitement des données textuelles médicales

- **Tokenisation spécialisée** (médicaments, codes SNOMED/ICD, etc.)
- **Normalisation des abréviations** (HTA -> Hypertension Artérielle)
- **Filtrage des données sensibles** (PHI anonymization)

La localisation (langue utilisée) et la diversité des formats (ex. CBC vs. NFS) compliquent encore davantage la tâche.

3.4. Prétraitement des données tabulaires

- **Imputation des données manquantes** (moyenne, médiane, kNN, MICE, etc.)
- **Normalisation / standardisation** (scores Z, min-max)
- **Réduction de dimensionnalité** (PCA, t-SNE)

Pour notre plateforme TweenMe, on utilise des approches multivariées (kNN, PCA) pour mieux gérer la haute dimensionnalité en médecine.

4. Cybersécurité et IA dans le domaine de la santé

Au-delà des mesures classiques (authentification forte, architecture zero trust, xDR, SIEM, SOC…), l'IA introduit de nouveaux risques :

1. **Prompt injection malveillant** : Forcer un LLM à divulguer des informations sensibles ou adopter un comportement non prévu.
2. **Violation de la confidentialité différentielle** : Extraire des données sensibles du jeu d'entraînement (ex. « bug beetle juice » sur ChatGPT).
3. **Pollution des données d'apprentissage** : Le « data poisoning » fausse le modèle et peut avoir des conséquences graves en diagnostic.
4. **Stratégies de piratage via IA** : Deep fakes en temps réel, phishing automatisé, etc.

5. Exemples d'IA en R&D : limites et réalités

5.1. Cas AlphaFold v3

- **Points forts** : Prédiction de la conformation tertiaire (voire quaternaire) de protéines, calcul des forces de liaison ligand/récepteur.
- **Limite majeure** : Ignorance des conditions physico-chimiques réelles (pH, phase aqueuse, T°, etc.), cruciales pour la purification ou la formulation.

5.2. Génération de séquences (ADN, protéines)

- **GenAI** : Souvent citée pour sa capacité à générer de nouvelles séquences.
- **Réalité** : Une simple macro Excel peut générer des séquences aléatoires ; la valeur ajoutée vient de la **prédiction fonctionnelle** et de la **faisabilité expérimentale**.
- **Chimie organique (small molecules)** : L'IA générative n'indique pas forcément la voie de synthèse ni le rendement des réactions chimiques (i.e. la faisabilité technico économique n'est pas abordée par l'IA générative).

5.3. Retour d'expérience étudiant (EBI)

- **Comparaison « wet lab » vs. in silico** : Cristallographie vs. outils de modélisation (dont AlphaFold v3).
- **Conclusion** : Les étudiants ont retenu un autre outil qu'AlphaFold, jugé trop éloigné des résultats expérimentaux.

Cette observation rappelle que même des IA avancées peuvent faire des erreurs patentes. Elles doivent être **validées** par des méthodes expérimentales et faire l'objet d'une analyse critique (éviter le biais de confirmation).

6. Conclusion : l'IA, un catalyseur à manier avec discernement

Plusieurs éléments clés ressortent de la table ronde :

1. **Qualité et traçabilité des données:** Sans données fiables, l'IA est vouée à générer des biais et des erreurs.
2. **Certification et régulation:** Les industries régulées exigent une robustesse et une transparence accrues, pas toujours garanties.
3. **Spécificités du secteur médical:** Données hétérogènes, faiblement volumineuses, nécessitant une expertise clinique.
4. **Cybersécurité:** Prompt injection, data poisoning, deep fakes : des menaces montantes.
5. **Limites des modèles:** Ex. AlphaFold v3 : ignore la complexité des conditions réelles. Génération de séquences : un simple aléatoire ne suffit pas, la valeur ajoutée est dans l'identification de la fonction et l'évaluation de la faisabilité.

L'IA doit donc être vue comme un **accélérateur** et un **outil d'aide**, non un substitut à l'expertise humaine. Son déploiement demande rigueur méthodologique, attention portée à la qualité et respect des contraintes industrielles, médicales et réglementaires.

Recommandations :

- **Renforcer la traçabilité et l'audit des données** (contrôle qualité, validation indépendante)
- **Établir des standards communs** (interopérabilité, harmonisation)
- **Former davantage les équipes** (fondamentaux data science, cybersécurité)
- **Confronter systématiquement la théorie à la pratique** (validations in vivo / in vitro)

En définitive, l'IA constitue un formidable levier d'innovation pour les bio-industries, à condition de respecter la **vigilance scientifique**, la **prudence réglementaire** et la **responsabilité environnementale**.