WHITE PAPER

L'IA VA T'ELLE TUER VOTRE ORGANISATION OU LA RÉINVENTER

www.reply.com

L'IA VA T'ELLE TUER NOTRE ORGANISATION OU LA RÉINVENTER

SOMMAIRE

RÉSUMER ÉXÉCUTIF

L’intelligence artificielle (IA) transforme profondément les entreprises en leur offrant des opportunités pour automatiser les tâches, optimiser les processus et améliorer la prise de décision. Ce livre blanc, dernier volet d’une trilogie, se concentre sur l’**utilisation éthique et responsable de l’IA, sa gouvernance et son rôle stratégique au sein des systèmesd’information et des métiers.**

Après avoir exploré les IA numériques et leurs usages, ainsi que l’ensemble des domaines de l’intelligence artificielle, cette dernière partie approfondit les enjeux liés à son déploiement, notamment la réglementation, la gouvernance et les risques associés. Elle illustre également les bonnes pratiques, les architectures adaptées et le cycle de vie des solutions IA, afin d’encadrer leur utilisation.

L’intégration maîtrisée de ces aspects permet de limiter les dérives et favorise une adoption plus responsable et conforme aux exigences de transparence et de régulation.

Ce guide pratique constitue une ressource clé pour toute organisation souhaitant exploiter pleinement le potentiel de l’IA dans une approche stratégique et durable.

INTRODUCTION

Dans un contexte où l'innovation technologique transforme les modèles économiques, l'intelligence artificielle (IA) s'impose comme un levier essentiel pour les entreprises cherchant à innover et à renforcer leur compétitivité. Bien plus qu'une simple réponse aux défis actuels, elle offre l'opportunité d'accompagner les métiers, d'optimiser les processus et d'adapter les systèmes d'information (SI) aux enjeux futurs.

Ce livre blanc est le troisième volet d'une série en trois parties, conçue pour offrir une vision claire et approfondie de l'IA et de son impact stratégique.

La première partie a posé les bases essentielles pour comprendre l'intelligence artificielle numérique. Elle explore les mécanismes du Machine Learning (ML) et Deep Learning (DL), en mettant en lumière leurs fonctionnements et leur rôle dans la transformation des entreprises.

La deuxième partie fut consacrée aux domaines d'application de l'IA. Elle analysera où et comment ces technologies sont mises en œuvre aujourd'hui et comment elles évolueront demain. Des exemples seront présentés dans des domaines clé tels que la classification des connaissances, l'industrie, la santé, la sécurité et la reconnaissance des émotions.

Cette troisième et dernière partie traite des enjeux liés à la régulation, à l'éthique et à la gestion des risques IA. Elle explore les cadres réglementaires, le cycle de vie des systèmes d'IA et les bonnes pratiques pour utiliser des solutions d'IA numériques de façon plus responsables et maîtrisées.

Cette série constitue une ressource essentielle pour comprendre, exploiter et encadrer l'intelligence artificielle avec succès.

« Tout deep learning repose sur le machine learning, et tout machine learning repose sur l'intelligence artificielle. Toutefois, l'intelligence artificielle ne repose pas toujours sur le machine learning.» – Google **Quelle est la différence entre le deep learning, le machine learning et l'intelligence artificielle ?**

L'Intelligence Artificielle porte sur des applications et des technologies nouvelles engendrant des risques IA. Dans cette dernière partie, S**ylvie Varesano, experte en architecture, gouvernance et intégration des systèmes de Machine Learning chez Wemanity Reply,** décrypte le non- déterminisme de l'IA numérique, la réglementation IA, et expose des pratiques pour une utilisation plus éthique et responsable.

À travers cette analyse, ce livre blanc vise à fournir une vision claire et accessible de l'IA appliquée, permettant aux organisations d'identifier des cas d'usage innovants et d'intégrer ces technologies efficacement dans leur stratégie.

POURQUOI RÉGULER L'IA

L'IA-numérique une solution non-déterministe?

"Sous domaine de l'intelligence artificielle, le Machine Learning (Apprentissage Automatique) comprend de nombreux algorithmes pour créer à partir des données des modèles dont les performances s'améliorent au fil de leur entraînement. Le Machine Learning est divisé en trois grandes catégories: l'apprentissage supervisé, l'apprentissage non supervisé, l'apprentissage par renforcement. Toutes les formes de Machine Learning s'appuient sur des statistiques et des probabilités qu'un événement ou une action ne produise ou non." Bercy numérique

L'IA numérique est également appelée IA connexionniste, repose sur une masse de données numériques et le principe de connexions neuronales, pour reproduire le résultat intelligent observé. Appuyé sur un grand ensemble de données, les Machines Learning appliquent des fonctions statistiques et probabilistes pour identifier les modèles de données.

La statistique permet d'identifier des corrélations entre les données collectées, l'échantillon représentatif des données que l'étude statistique veut analyser. Selon l'échantillon, les hypothèses posées, les statistiques peuvent donner un résultat ou son contraire : "Il existe de multiple biais menant à une erreur statistique : trop peu de données, une question mal posée ou des termes mal définis, des périmètres non comparables... Les écueils sont nombreux." Le Monde

Dans les mathématiques, le domaine des probabilités est la mesure dans l'espace 0:1 de la survenance d'un événement. Sous le domaine de la théorie de la mesure qui porte sur l'espace R, la probabilité mesure les phénomènes aléatoires qui semblent être dus au hasard, et s'opposent aux fonctions déterministes

Contrairement à la logique, à l'algèbre ou à l'analyse, les statistiques et les probabilités ne relèvent pas des mathématiques déterministes.L'IA numérique, qui repose sur ces disciplines, est par nature non déterministe. Son utilisation requiert donc une gouvernance rigoureuse ainsi qu'une gestion proactive des risques associés.

UNE UTILISATION ÉTHIQUE L'IA

En 2019, plusieurs entreprises, dont Bouygues, Microsoft, Orange, Air France-KLM et Enedis, ont créé Impact IA, un collectif dédié aux enjeux éthiques de l'intelligence artificielle. Ce groupe explore des thématiques clés telles que l'IA et le travail, son impact environnemental, la mixité et la diversité numérique, ainsi que son rôle dans l'industrie de la santé. En suivant de près les évolutions du secteur, Impact IA vise à identifier et à promouvoir des projets d'intelligence artificielle à impact positif.

Avec l'essor de l'intelligence artificielle, plusieurs groupes de réflexion s'intéressent aux enjeux éthiques et aux bonnes pratiques à adopter. Les chercheurs en Machine Learning travaillent notamment sur l'explicabilité des modèles, cherchant à mieux comprendre leurs décisions.

L'équipe de Frédéric Jurie explore, par exemple, des approches pour évaluer la qualité des images générées et analyser pourquoi un modèle a identifié une émotion plutôt qu'une autre. Cependant, le processus décisionnel du Machine Learning repose sur des milliers de paramètres interconnectés, sans correspondance sémantique évidente, rendant son interprétation complexe.

L'explicabilité des IA est un enjeu clé pour leur adoption à grande échelle. Dans les Systèmes d'Information des entreprises, chaque action sur les données doit être justifiée, ce qui impose actuellement deux approches : soit une validation systématique par un acteur interne avant toute utilisation du Machine Learning, soit une restriction à des domaines spécifiques ou à certains types de données. Cette exigence de transparence est essentielle pour garantir la confiance et la conformité des solutions IA.

Par exemple, la modélisation du raisonnement juridique implique l'intégration des notions déontiques de permission, obligation et interdiction dans l'approche éthique. La protection de la vie privée, quant à elle, repose sur plusieurs mécanismes, tels que l'anonymisation des données personnelles (k-anonymat, l-diversité, confidentialité différentielle), la prévention des biais d'apprentissage, ainsi que la recherche d'équité et d'explicabilité dans les processus de prise de décision.

La nécessité de justifier les résultats d'un modèle de Machine Learning est une particularité propre à l'IA numérique. En revanche, l'IA symbolique, basée sur des règles logiques, algébriques et analytiques, fournit systématiquement une explication de ses résultats, garantissant ainsi une explicabilité intrinsèque des algorithmes.

Chacune de ces approches présente des avantages et des limites. L'avenir de l'intelligence artificielle repose ainsi sur leur combinaison, donnant naissance à l'IA neuro-symbolique, qui vise à allier la puissance d'apprentissage du Machine Learning à la rigueur explicative de l'IA symbolique.

LA RÉGLEMENTATION

Le 19 octobre 2022, le Parlement européen a adopté le règlement sur les services numériques (Digital Service Act), visant à encadrer l'usage de l'IA.
Puis, le 13 juin 2024, l'Union européenne a adopté la loi IA Act, qui régit les produits IA utilisés sur son territoire et le 5 septembre 2024 le Conseil européen a publié une convention-cadre sur l'IA. Ces directives permettent aux entreprises d'intégrer les nouvelles solutions d'IA en cherchant à davantage maîtriser leurs risques tant sur le SI que les utilisateurs, par la mise en place d'une gouvernance des solutions IA.

L'IA Act sera en vigueur dès le 2 février 2025 pour les usages interdits :

- La manipulation,
- L'exclusion ou la discrimination,
- La classification sociale d'individus,
- La prédiction de criminalité,
- L'exploitation des reconnaissances faciales non ciblées,
- La détection d'émotions et son exploitation
- L'identification personnel dans les espaces publics

Puis, à partir du 2 août 2025, un cadre réglementaire spécifique encadrera les GPAI (systèmes d'IA à usage général), tels que les LLM (Large Language Models). Enfin, en 2026, les solutions d'IA dites "à haut risque" devront se conformer à des exigences strictes, impliquant notamment :

- une gestion des risques,
- une gouvernance des données,
- une documentation technique,
- un enregistrement automatiquement des événements,
- des instructions d'utilisation aux utilisateurs,
- une surveillance humaine,
- une conception du système d'IA pour atteindre les niveaux appropriés de précision, de robustesse et de cybersécurité,
- un système de gestion de la qualité.

En réponse, la norme ISO/IEC 42001 a été mise en œuvre et publiée en décembre 2023. L'ISO 42001 a été conçue pour structurer et promouvoir des usages plus éthiques, plus responsables et sécurisés de l'IA .

La certification ISO 42001 de votre solution IA (construction et/ou utilisation de cette solution), garanti votre conformité à cette réglementation.

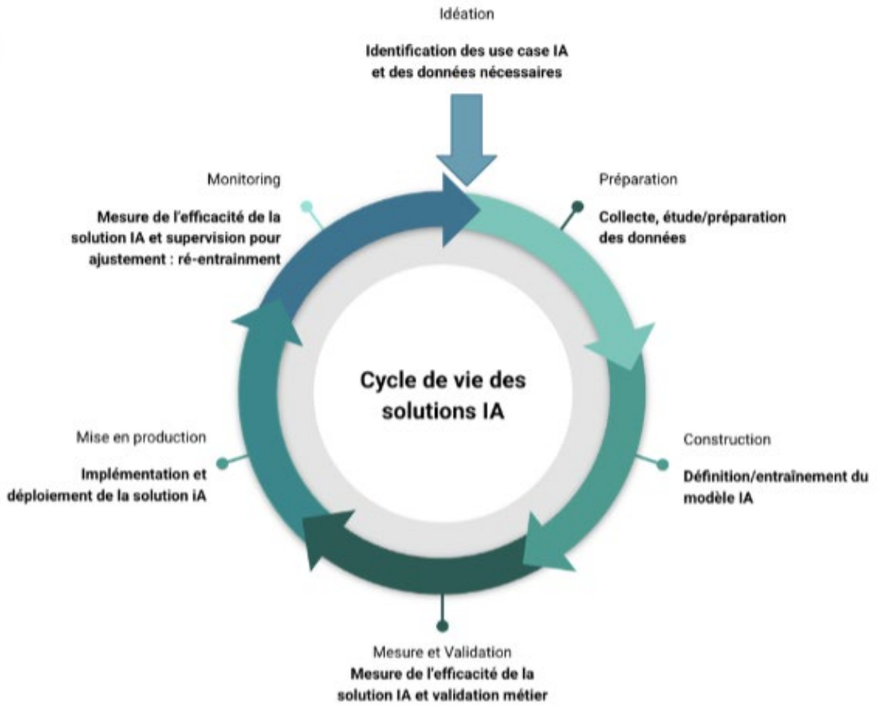
Pour mesurer votre conformité, différentes solutions voient le jour, comme au LNE (Laboratoire National de Métrologie et d'Essais) qui a développé une solution permettant d'évaluer les systèmes d'IA : évaluer les algorithmes et les systèmes embarquant de l'IA (robots d'aide à la personne, caméras intelligentes, robots d'intervention civils et militaires, etc.), caractériser leur fiabilité, sécuriser leur usage, et garantir leur caractère éthique (traitement équitable des données, juste information délivrée à l'utilisateur).

La qualification de systèmes intelligents offre aux acteurs de l'IA un outil d'aide à la décision pour évaluer les performances de ces systèmes IA et bénéficier d'analyses (justesse, robustesse, périmètre de fonctionnement, etc.). Cette gouvernance permet ainsi de réduire les **risques IA encourageant une utilisation responsable.**

LA GOUVERNANCE DES SOLUTIONS IA

La gouvernance des solutions IA doit permettre un suivi tout au long de leur cycle de vie. Une solution IA est expérimentée afin de définir la réponse au Use Case recherché avant que cette solution IA ne soit implémentée. Une IA- numérique n'étant pas déterministe, elles doivent être validées par l'utilisateur final.

Par exemple, si un client achète l'article suggéré par l'application, la prédiction est confirmée.



Pour chaque solution IA, le cycle de vie est ainsi défini :

- **Idéation :** identification du cas d'usage de l'IA et des données nécessaires à son exploitation.

- **Préparation :** collecte, nettoyage et analyse des données pour assurer leur qualité et leur pertinence.

- **Construction :** définition et entraînement du modèle IA, en tenant compte des objectifs métier et techniques.

- **Mesure et Validation :** évaluation de l'efficacité du modèle et validation par les utilisateurs métier avant son déploiement.

- **Mise en production :** développement et déploiement de la solution selon les normes de l'entreprise, sous forme d'API ou d'application pour une utilisation opérationnelle.

- **Monitoring :** suivi des performances de l'IA et de la satisfaction des utilisateurs, avec des ajustements réguliers (optimisation ou réentraînement du modèle) pour garantir son efficacité.

LA GESTION DES RISQUES IA

Comme pour tout système IT, le fournisseur de solutions IA doit assurer une gestion rigoureuse des risques liés à la conception de son intelligence artificielle. De leur côté, les entreprises utilisatrices doivent mettre en place un dispositif de gestion des risques associés à son exploitation. Tout incident identifié devra être pris en compte et corrigé dans le cadre du processus de gestion des risques. **La gestion des risques IA couvre à ce stade :**

- **Les usages :** stratégie IA et cas d'utilisation.

- **Les ressources :** compétences des personnes impliquées dans l'idéation, l'expérimentation et le développement.

- **Les données :** origine, qualité et diversité des données tant en construction qu'en utilisation.

- **Le Data Lab :** solution IT, algorithme et mode de construction utilisésen expérimentation.

- **La conception :** industrialisation et déploiement des solutions IA (construites ou simplement ajoutées au S.I.).

- **Maintenance :** supervision et maintien opérationnel des solutions IA.

- **Impacts métier** : changement dans le processus métier qui emploie la solution IA.

- **Utilisateur et utilisation :** connaissance de l'usage et de l'utilisation de la solution IA.

Les risques identifiés doivent être atténués tout au long du cycle de vie des solutions IA grâce à une gouvernance rigoureuse.

Le Data Lab du Crédit Agricole, en collaboration avec l'École Polytechnique, a développé un framework de gestion des risques liés aux LLM en production, reposant sur le Red Teaming. Ce dispositif permet d'identifier, atténuer et renforcer la sécurité face aux vulnérabilités des modèles. Présenté lors du FIIA 2025, le Crédit Agricole a souligné que la généralisation de l'usage des LLM entraîne une augmentation des attaques et des risques associés. En réponse, il a conçu un "LLM as Judge", un framework de contrôle des risques IA visant à garantir une utilisation plus éthique et responsable des modèles.

L'adoption croissante des LLM les expose inévitablement à des attaques. Déployer une solution IA implique des risques qui doivent être maîtrisés.

L'exemple de Lucie, immédiatement mise à l'épreuve et échouant face à ses limites, illustre l'impact négatif d'une mauvaise préparation sur la réputation d'un modèle. De plus, les hallucinations des LLM suscitent des inquiétudes quant à leur fiabilité.

Dans le domaine médical, Berger-Levrault, qui développe des LLM spécialisés, travaille à réduire ces erreurs grâce à l'architecture RAG et au fine-tuning. Lors du FIIA 2025, il a mis en avant une conclusion clé : la qualité des données prime sur leur quantité. Une mauvaise qualité de données prolonge l'apprentissage et amplifie les biais, compromettant la performance du modèle.

Comme toute solution IT, l'intelligence artificielle doit être gouvernée en fonction des risques qu'elle présente. Une approche proactive et structurée est essentielle pour garantir son adoption en toute sécurité, transparence et éthique.

DES SOLUTIONS ÉTHIQUES ET RESPONSABLES

Le principe d'éthique et responsable de l'IA, selon la réglementation, est pour l'instant défini :

- Sur l'origine et la diversité des données utilisées qui ne doit pas exclure des populations ou opinions.

- Sur la cible des utilisateurs qui se doit d'être le plus inclusif.

- Sur les algorithmes et la construction des modèles : plus une solution est explicable plus est réduit le risque algorithmique, à défaut un monitoring est nécessaire.

- Sur l'utilisation de la solution IA tant par une utilisation en connaissance de la solution (acculturation, formation) qu'une utilisation respectueuse des utilisateurs.

LES DONNÉES

Point névralgique des IA numériques, la bonne gestion des données permet la réduction du risque qu'elle portent sur la solution IA finale.

- L'identification et la collecte des données d'influence du cas d'utilisation, mais également une grande diversité des configurations de ces données et un grand volume.
- Traitemement, sécurisation et stockage des données d'entraînement dans le respect du RGPD.
- Revue du résultat prédit ou généré, par l'utilisateur pour éliminer le non-déterminisme du résultat.

Face à l'important besoin en volume de données, certains expérimentateurs en IA génèrent des données synthétiques pour entraîner leurs modèles. Ces données, bien que conçues pour pallier le manque d'informations réelles, soulèvent des problèmes éthiques liés à l'IA.

En effet, une IA générative entraînée sur des données synthétiques risque de ne pas refléter la diversité des situations du monde réel. Ces données ne sont qu'une approximation des configurations initiales, ce qui peut limiter la robustesse et la généralisation des modèles, augmentant ainsi le risque de biais et d'erreurs dans les décisions automatisées.

LA CONSTRUCTION D'UNE IA PRÉDICTIVE

Pour former une IA Prédictive, il faut identifier les données en entrée d'une tâche métier, les données qui influent le résultat de la tâche, et les étiquettes (associées à leur résultat) à prédire.

Dans l'exemple de la prédiction d'achat immobilier, les données qui influencent l'achat sont celles qui déterminent le type de bien immobilier qu'à acheter. Le revenu du foyer, mais également la composition du foyer, détermine si l'achat sera un F1 ou un F5. Par contre, la couleur des cheveux de l'acheteur n'a aucun impact sur l'achat effectué.

Dans cette activité métier, l'acteur assigné à la tâche utilise un ensemble de données pour produire un résultat. Ces mêmes données doivent être collectées par la solution IA afin d'entraîner un modèle et ainsi développer une IA prédictive adaptée à cette tâche spécifique.

L'entraînement du Machine Learning s'effectue au sein d'un Data Lab, un environnement IT dédié à l'expérimentation, permettant la validation des modèles. À partir de données étiquetées, le modèle est entraîné en s'appuyant sur des approches statistiques ou probabilistes.

Une fois que l'accuracy du modèle, c'est-à-dire la proportion de prédictions correctes par rapport aux résultats réels, atteint un niveau satisfaisant (généralement supérieur à 85 %), le modèle peut être validé et préparé pour une industrialisation.

LE FINE-TUNING DES IA GÉNÉRATIVES

Le réglage fin consiste à adapter un modèle pré-entraîné à des tâches ou à des cas d'utilisation spécifiques.C'est devenu une technique d'apprentissage profond fondamentale, en particulier dans le processus d'entraînement des modèles defondation utilisés pour l'IA générative." IBM

"Le fine-tuning consiste à poursuivre l'entraînement du modèle sur une tâche et/ouun domaine spécifique pour le spécialiser. "PFIA2024

Il est plus simple et économiquement avantageux d'optimiser un modèle de base pré-entraîné, déjà doté de connaissances générales adaptées à une tâche spécifique, plutôt que de développer un nouveau modèle à partir de zéro. Cette approche est particulièrement pertinente pour les modèles d'apprentissage profond, tels que les grands modèles de langage (LLM), qui reposent sur des millions, voire des milliards de paramètres.

En exploitant la préformation des modèles via l'apprentissage par transfert, le fine-tuning permet de diminuer considérablement les besoins en puissance de calcul coûteuse et en données étiquetées, rendant ainsi les grands modèles mieux adaptés à des cas d'utilisation spécifiques.

Par exemple, le fine-tuning permet d'ajuster un LLM pré-entraîné afin de modifier son ton conversationnel ou d'adapter le style d'illustration d'un modèle de génération d'images.

Il peut également être utilisé pour enrichir l'apprentissage d'un

modèle en intégrant des données spécifiques à une entreprise ou à un domaine métier. Cette adaptation permet ainsi à un LLM de répondre avec précision à des questions propres à un secteur d'activité, renforçant ainsi sa pertinence et son efficacité pour des usages spécialisés

Bien qu'il soit théoriquement possible d'entraîner un modèle à partir de zéro avec un ensemble de données spécifiques à une tâche, cette approche présente un risque élevé de surajustement. Dans ce cas, le modèle pourrait produire des résultats satisfaisants sur les données d'entraînement, mais manquerait de généralisation face à de nouvelles données, compromettant ainsi son efficacité.

Le fine-tuning permet d'atténuer ce risque en s'appuyant sur les poids d'un modèle pré-entraîné, auquel on applique un apprentissage supplémentaire sur un ensemble de données plus restreint.

Cet ensemble est spécifiquement conçu pour affiner les capacités du modèle en fonction des tâches et cas d'utilisation ciblés, garantissant ainsi une meilleure adaptation à son contexte d'application.

Les ensembles de données utilisés pour le fine-tuning apportent au modèle des connaissances spécifiques au domaine, au style, aux tâches ou aux cas d'utilisation visés, permettant d'adapter le modèle pré-entraîné à ces besoins particuliers.

Par exemple, un modèle de classification d'images capable d'identifier certaines espèces d'oiseaux peut être entraîné à reconnaître de nouvelles espèces grâce à des échantillons d'entraînement supplémentaires avec des étiquettes appropriées. De même, un LLM pré-entraîné sur le langage général peut être affiné pour des tâches de codage en utilisant un ensemble de données contenant des requêtes de programmation pertinentes et leurs extraits de code correspondants.

L'ARCHITECTURE RAG POUR LES LLM

Le fine-tuning nécessite un processus de réapprentissage souvent long et exigeant en données. La technologie RAG (Retrieval-Augmented Generation) repose sur l'intégration d'informations déjà bien structurées, tabulées et régulièrement mises à jour, ce qui rend son utilisation plus rapide pour des utilisations spécifiques.

Le RAG (Retrieval-Augmented Generation) est une architecture hybride qui combine un modèle de génération de langage (LLM) avec un système de récupération d'informations. Cette combinaison permet au modèle de puiser les données dans une base de données externe pour enrichir les réponses du LLM.

Dans un RAG, une configuration typique est composée de trois éléments :

1. Composant de Récupération (Retrieval): responsable de la recherche et de la sélection d'informations pertinentes à partir d'une base de données ou d'un ensemble de documents.

2. Composant de Génération (Generation): pour générer des réponses ou du contenu en utilisant les informations récupérées.

3. Composant d'Augmentation (Augmented) : pour améliorer, structurer laréponse selon l'attente de l'utilisateur.

RETRIVAL

Des algorithmes spécialisés en recherche d'informations eux-mêmes basés Sur l'IA peuvent être utilisés. Ces algorithmes peuvent inclure des moteurs de recherche textuels, des systèmes de classement utilisant l'apprentissage automatique, ou des réseaux de neurones spécialisés dans la récupération et dans le tri des informations pertinentes.

La demande de l'utilisateur est convertie en une représentation vectorielle et mise en correspondance avec les bases de données vectorielles.

Prenons l'exemple d'un chatbot intelligent conçu pour répondre aux questions des salariés sur le Comité d'Entreprise. Lorsqu'un Salarié demande : « Combien d'entrées au parc d'attractions puis-je obtenir cette année ? », le système extrait les informations pertinentes, telles que la politique de subventions, le tarif du parc et les données personnelles du salarié (situation actuelle, historique des subventions). Ces documents spécifiques sont ensuite fournis au LLM, car ils apportent une réponse directe et contextualisée à la question posée.

La pertinence des informations est évaluée à l'aide de calculs vectoriels et de représentations mathématiques, qui mesurent la correspondance entre la requête et les données disponibles. Bien que les algorithmes et moteurs de recherche automatisent ce processus de récupération, il reste essentiel de vérifier l'exactitude des informations, soit par une intervention humaine, soit à l'aide de filtres ou d'autres mécanismes de contrôle.

Le modèle RAG augmente la saisie (ou les instructions) de l'utilisateur en ajoutant les données récupérées pertinentes dans leur contexte.

GÉNÉRATION

Un LLM de génération peut être utilisé tel quel, sans nécessiter de réapprentissage, tant que sa performance reste satisfaisante pour l'application RAG visée. Lorsqu'il est déployé en interne dans une entreprise sans interaction avec des sources externes, sa performance demeure stable.

Cependant, sa mise à jour éventuelle doit être envisagée avec prudence, après une période définie, pour éviter l'introduction de biais ou la diffusion involontaire d'informations confidentielles.

AUGMENTED

Ce composant, souvent un modèle de langage avancé, est conçu pour synthétiser les informations et les présenter de manière cohérente et contextuelle. Cependant, les règles et cadres structurés doivent être validés avec soin par des experts, notamment à travers des tests couvrant tous les cas d'utilisation possibles.

Ce composant peut être des systèmes de recommandation pour fournir des suggestions personnalisées exploitant le texte, ou des systèmes fondés sur des règles fournissant un cadre structuré pour certaines tâches spécifiques, ou encore des IA plus complexes en charge de préparer les données.

Par ailleurs, les systèmes de recommandation doivent offrir des propositions adaptées au niveau de compétence des utilisateurs, qui doivent être formés à interagir efficacement avec l'IA, notamment via l'apprentissage des prompts.

L'approche RAG renforce les capacités des LLM en leur permettant d'intégrerdes informations spécifiques en temps réel, ce qui est particulièrement précieux pour répondre à des questions nécessitant des connaissances actualisées ou très précises.

LE MLOPS

Formée sur un ensemble de données, l'IA-numérique rencontre de nouvelles données lors de son utilisation, apportant des exemples inédits de cas d'utilisation. Le réentraînement permet alors d'intégrer ces nouveaux cas et d'ajuster le modèle pour maintenir sa précision.

Une solution d'IA Prédictive peut être re-entraîné en automatique via le monitoring de sa prédiction. Les prédictions confirmées ou non sont collectées pour mesurer en temps réel la perte de prédiction de la solution d'IA. Une solution d'IA Générative nécessite également un monitoring par les utilisateurs afin de le re-entraîner, affiner pour maintenir ou améliorer la solution IA.

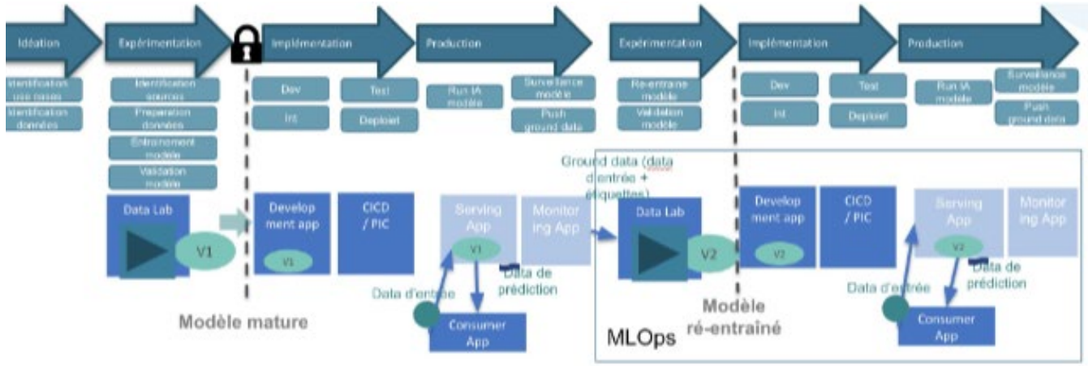
Le MLOps :

- En idéation : identification du Use Case , de la prédiction souhaitée

- En expérimentation (nouvelle phase propre aux solutions IA) : collecte et préparation des données pour entraîner le ML et former le modèle prédictif

- **En implémentation :** développement et déploiement suivant les règles de l'entreprise, sous forme d'API ou d'application.

- En production : utilisation par les acteurs métiers de l'application/API AI prédictive pour exécuter leur tâche. L'IA Prédictive doit être monitorée pour identifier les baisses de performance et déclencher son re- entraînement en expérimentation, lorsque nécessaire.



COMMENT REDUIRE LES COÛTS DES LLM ?

Un LLM, solutions la plus utilisée aujourd'hui, est extrêmement coûteuse tant en volume de données qu'en coût de calcul et de ce fait est loin d'être responsable écologiquement.

Pour pallier cette faiblesse des études sont réalisées. Loria (Laboratoire lorrain de recherche en informatique et ses applications) a fait le constat que plus le volume de données, la taille du modèle et le nombre de GPU (Graphic Processing Units) était important meilleur était le LLM formé.

Christophe Cerisara, chercheur au Loria, mène des recherches sur le défi des LLM en explorant des approches permettant de réduire leur taille tout en maintenant des performances optimales.

Ses modèles, appelés SmolLM, offrent des résultats comparables aux LLM classiques, mais avec un coût réduit, rendant leur utilisation plus accessible et efficiente.

La Scaling Law décrit comment les performances d'un réseau neuronal évoluent en fonction de la variation de ses paramètres clés. Christophe Cerisara met en évidence que Transformer, l'architecture neuronale introduite par Google en 2017, demeure la plus efficace pour l'apprentissage des modèles.

Dans ses travaux, il compare différentes approches visant à concevoir des "petits" LLM, en s'appuyant sur des modèles ayant bénéficié d'un pré- entraînement allégé. Cette recherche vise à répondre aux enjeux de responsabilité et d'optimisation liés aux LLM, en proposant des alternatives plus durables et adaptées aux contraintes de ressources.

Dans sa quête de la meilleure architecture, Christophe Cerisara constate que la quantisation ne permet pas l'entraînement des modèles et que la distillation, bien qu'efficace, demeure très consommatrice en données. En revanche, la compression par pruning semble offrir une solution prometteuse pour développer des IA plus responsables.

L'arrivée de DeepSeek, le nouveau LLM chinois, pourrait toutefois bouleverser ces résultats. Bien que peu coûteux en ressources GPU, DeepSeek reste extrêmement gourmand en données, soulevant de nouveaux défis en matière d'optimisation. Il ne fait aucun doute que, comme au Loria, d'autres chercheurs s'attelleront à trouver des solutions pour pallier cette faiblesse et améliorer l'efficacité des modèles IA de demain.

CONCLUSION

Après avoir exploré la troisième partie de cette trilogie de livres blancs, vous disposez désormais des clés pour répondre aux exigences réglementaires et adopter des solutions IA plus éthiques et responsables. Vous maîtrisez le cycle de vie des solutions IA ainsi que les typologies de risques, essentielles pour une intégration sécurisée de vos modèles. Il ne vous reste plus qu'à mettre en place une gouvernance IA efficace afin de mesurer et atténuer les risques, tout en optimisant vos solutions grâce à des techniques telles que le fine-tuning, le prompt engineering, le RAG, ou encore en combinant ces approches. Cette démarche vous permettra d'intégrer l'IA dans le respect de la réglementation en vigueur.

Glue Reply, vous accompagne à chaque étape du cycle de vie de vos solutions IA, de la définition de votre stratégie jusqu'à la mise en œuvre des architectures IA adaptées à vos besoins.

Que ce soit en phase d'utilisation ou de développement, les solutions IA nécessitent un encadrement rigoureux, incluant une gouvernance tout au long du cycle de vie, une gestion proactive des risques et une formation des parties prenantes, de l'idéation à l'adoption. Selon les cas d'usage et les technologies IA mises en place, les risques varient et exigent une approche sur-mesure.

Avec Glue Reply, nos architectes vous accompagnent dans l'identification des cas d'usage IA, leur intégration à vos systèmes, l'optimisation de leur exploitation ainsi que la mise en place d'une gouvernance IA responsable et sécurisée. En structurant vos initiatives IA avec une approche stratégique et adaptée aux enjeux réglementaires, vous pourrez exploiter pleinement le potentiel de l'IA tout en garantissant conformité, transparence et performance.

CONTACTEZ-NOUS

Découvrez comment adopter une IA éthique et performante dans votre entreprise !

M. Yvan GRAVIER,
Leader Agile IT chez Glue Reply : **y.gravier@reply.com**