HUB FRANCE IA

LES RISQUES DE L'IA GENERATIVE

Juillet 2024

Table des matières

HUB FRANCE IA

1. Introduction

Les risques de l'IA Générative

Juillet 2024

1. Introduction

En mai 2023, le Hub France IA, avec ses membres, publiait une note de synthèse[1] pour éclairer les **enjeux de la révolution ChatGPT**. Dans une partie dédiée aux usages, nous avions d'abord identifié les quatre grandes typologies d'usage basées sur les capacités de rédaction, de classification, de traduction et de synthèse. Nous avions ensuite décrit des usages sur onze grands domaines : relation client et marketing, développement informatique, cybersécurité, banque et assurance, BTP, recherche, enseignement, journalisme, ressources humaines, juridique et santé. Nous avions ensuite décrit les risques de l'introduction de l'IAG (Intelligence Artificielle Générative) dans l'entreprise et les impacts à en attendre. Ce premier document avait été construit quelques mois seulement après l'annonce de ChatGPT fin octobre 2022. ChatGPT était alors apparu comme le premier représentant des agents conversationnels exploitant un grand modèle de langage **(*Large Language Model*** ou **LLM**), dont le nombre s'est largement accru depuis, avec de nouveaux usages.

En janvier 2024, le Hub France IA et ses membres poursuivent leur travail en publiant un livre blanc[2] détaillant des **exemples d'usages** possibles avec une IA générative basée sur un grand modèle de langage. Le document est constitué de trois axes. Le premier axe est consacré aux **usages autour de six grands domaines** : cybersécurité, industries culturelles et créatives, ressources humaines, développement informatique, éducation et marketing. Le second axe, étudie l'apport des LLM pour les **agents conversationnels (*chatbots*)**, un des grands usages de ChatGPT, à travers l'analyse de quelques retours d'expérience pour évaluer les gains et limites d'usage de ChatGPT pour cette fonction de *chatbot*. Enfin, le troisième axe, à travers une **enquête**, analyse les gains, manques à gagner et freins dans les usages actuels des IA génératives.

[1] Groupe de travail ChatGPT. ChatGPT : Usages, Impacts et recommandations. Note de synthèse. Hub France IA. Mai 2023. https://www.hub-franceia.fr/wp-content/uploads/2023/04/ChatGPT_Note-synthese.pdf

[2] Groupe de travail IA Générative. Livre blanc : les usages de l'IA Générative, Volume I – Les LLM. Hub France IA. Janvier 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf

Face à l'engouement pour l'IA générative, au sein des organisations, mais également chez le grand public, le Groupe de Travail du Hub France IA a décidé de poursuivre ses travaux et se penche, à travers ce document, sur les **risques générés par l'usage des LLM** au sein des organisations. Comprendre ces risques est critique si on veut mesurer les impacts de l'IAG et si on veut mettre en place des pistes de remédiation pour en maximiser les bénéfices.

En effet, l'IA générative, bien que puissante et révolutionnaire, rencontre aussi des défis et des limites. Par exemple, nous pouvons noter des interrogations sur la qualité et la cohérence du contenu généré, les biais et inexactitudes potentielles héritées des données d'entraînement, le manque de créativité et d'originalité par rapport aux humains ainsi que des questions éthiques.

Note : Bien que les LLM ne représentent qu'un type de systèmes d'IA générative parmi d'autres, ce type de modèles est aujourd'hui le plus répandu dans les organisations, devant les autres comme la génération d'images ou de sons ; aussi, la gestion des risques de ce type d'IA générative tend à être plus avancée. C'est pourquoi ce document traite essentiellement des risques causés par les LLM en particulier, à travers de multiples cas d'usages. Seule exception, le domaine des Industries Culturelles et Créatives qui illustre des exemples de risques causés par d'autres types de contenus générés comme l'audio, l'image ou la vidéo.

Le présent document se décompose en quatre grands chapitres :

Un premier chapitre illustre les **définitions générales des concepts** présentés dans le document, facilitant ainsi la lecture des chapitres suivants. Les définitions traitées sont : le risque, les parties prenantes, les causes, les impacts ou encore la criticité.

Le deuxième chapitre expose une **démarche d'analyse des risques** spécifiques aux IA génératives au sein d'une organisation. Il ne s'agit pas de présenter la méthodologie générale et transverse du pilotage des risques en entreprise, ni particulièrement des risques liés au système d'information (SI), mais bien de présenter les spécificités propres à l'usage de l'IA générative qu'il convient d'intégrer dans son processus de pilotage des risques.

Le troisième chapitre présente une démarche d'analyse des risques, avec les grandes catégories d'usages des LLM, quel que soit le domaine ; ensuite, on introduit une matrice des risques, permettant d'analyser les risques liés à un cas d'usage d'une IA générative dans une organisation. Enfin, nous présenterons **de multiples cas d'usages**, classés par domaine. A ce jour, cette section n'est pas traitée ici : les cas d'usages seront

progressivement publiés après la parution de ce document, durant les mois de juillet et août. Les domaines qui seront traités sont les suivants : Ressources humaines, Finances, Santé, Conseil, Data Science, Développement logiciel, Commerce, Juridique, Marketing, Service client, Cybersécurité, Industries culturelles et créatives et Logistique & transport. Chaque exemple de cas d'usage est présenté de sorte qu'il mette en évidence le(s) risque(s) généré(s) par le LLM employé. Pour chaque cas d'usage, la démarche d'analyse est concrètement exécutée sur l'exemple.

Le quatrième et dernier chapitre propose une **synthèse des remédiations** propres à chaque type de risque et ses causes associées.

Nous espérons que ce nouvel éclairage permettra d'éveiller ou de renforcer votre vigilance face à cette nouvelle technologie, à la fois prometteuse et dangereuse, si mal employée.

HUB FRANCE IA

2. Définitions

Contributeurs :

- ***Benjamin Bosch**, Manager - Model risk Management – Data Science, Société Générale*
- ***Martin D'Acremont**, Consultant – Wavestone*
- ***Imen Fourati**, Expert lead, Risque de modèle, Société Générale*
- ***Nicolas Pellissier**, Cofondateur – Klark*
- ***Yael Suissa**, CEO & Cofondateur - MAP-Monitoring And Protection*

2. Définitions

2.1. IA Générative

Le parlement européen définit l'intelligence artificielle comme « tout outil utilisé par une machine afin de reproduire des comportements liés aux humains, tels que le raisonnement, la planification et la créativité. »

En particulier, l'intelligence artificielle générative (IAG) est une technologie qui a connu une évolution fulgurante ces dernières années. Il s'agit d'un sous-ensemble du Machine Learning, et plus particulièrement du Deep Learning, visant à produire du contenu, que ce soit du texte, une image, de l'audio ou une vidéo, à partir de données en entrée (on parle alors de *prompt*), elles-mêmes du texte, une image, de l'audio ou une vidéo par exemple. Un système d'IA générative crée un nouveau contenu statistiquement cohérent avec les données d'entraînement et la requête formulée.

Les systèmes d'IA générative sont généralement entraînés sur un large ensemble de données, nécessitant des moyens conséquents pour leur apprentissage. L'architecture du modèle (*Transformer*[3]) associée au volume très important de données utilisées pendant l'entraînement permet des usages variés pour ces modèles, sans que ces derniers n'aient été entraînés spécifiquement pour ces tâches.

Les premières solutions fondées sur de grands modèles de langage, tels que ChatGPT[4], sont capables de créer un texte à partir d'instructions textuelles en entrée. Ces modèles deviennent aujourd'hui de plus en plus multimodaux, c'est-à-dire qu'ils peuvent prendre en compte, aussi bien en entrée qu'en sortie, des données de plusieurs types, même combinées telles que l'image, l'audio, la vidéo, de la 3D, etc.

2.2. Risque

Le risque étant une notion évoquée dans plusieurs disciplines et dont la définition peut varier, il convient de définir ce concept qui sera utilisé dans toute la suite de cette étude.

---

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. *Attention is all you need. Advances in neural information processing systems*. vol. 30. 2017. https://arxiv.org/pdf/1706.03762.pdf

[4] OpenAI. Introducing ChatGPT. Open AI. November 30, 2022. https://openai.com/blog/chatgpt

Un risque peut être défini comme : « Tout événement ou situation pouvant entraîner des conséquences d'ordre humain, financier, juridique, réglementaire, ou relatif à la réputation, susceptibles d'impacter l'entreprise dans l'atteinte de ses objectifs ou dans son développement, quelle que soit la nature des causes et l'origine du risque (interne ou externe) »[5]. La norme ISO 31000 sur le management du risque le définit également comme « l'effet de l'incertitude sur l'atteinte des objectifs ». [6]

Les éléments fondamentaux du risque, qui permettent ensuite de définir des stratégies de gestion des risques, sont généralement les suivants :

- Sa ou ses causes ;
- Sa ou ses conséquences, ou impacts, qui peuvent être mesurés selon une échelle propre à chaque organisation, son contexte et le type d'impact (par exemple, un montant des pertes occasionnées pour mesurer un impact financier) ;
- Sa probabilité d'occurrence, ou vraisemblance.

La combinaison des deux derniers éléments permet généralement de définir l'acceptabilité d'un risque, déterminant ensuite la manière dont une organisation décide de le traiter, comme dans l'approche retenue par l'ANSSI pour sa méthodologie d'analyse de risque EBIOS RM[7].

L'un des objectifs de cette étude est donc d'identifier quels sont les risques qu'amène l'usage de l'IA générative, et d'en explorer les impacts potentiels sur différents aspects des organisations et de la société. La présentation de mesures de remédiation ou de maîtrise des risques, en fin de document, permettra d'aborder comment réduire la probabilité d'occurrence d'un risque, et donc de voir ses impacts se matérialiser.

2.3. Parties prenantes

La gestion des risques qui émanent de l'IA générative peut être organisée selon trois lignes de défense. La première ligne de défense (LoD[8]) est représentée notamment par les *Model Owners*. Ce sont les personnes qui prennent la responsabilité de l'utilisation des systèmes d'IA et sont donc les premiers à devoir s'assurer que les risques afférents ont bien été pris en compte lors des diverses étapes du développement, déploiement et

[5] Jean-Luc Wibo, cité par Laurence Baillif, Gestion des risques - De la sécurité à la gestion globale des risques, CNPP Editions, 2023, p. 23
6 Norme ISO 31000:2018, Management du risque, Organisation International de normalisation (ISO), 2018
7 Cf. Glossaire
8 Cf. Glossaire

utilisation de l'IA. Ces personnes s'assurent que le processus de développement, la documentation, et le suivi (*monitoring*) de l'IA sont conformes aux standards de l'entreprise.

La deuxième ligne de défense (LoD2) est constituée par les équipes de revue indépendante, les équipes en charge de la gouvernance et de la supervision du portefeuille des systèmes d'IA et si pertinent, des personnes participant aux comités d'approbation et de revue. Leur rôle est de collectivement s'assurer que la première ligne de défense joue bien son rôle de gestion des risques de modèle, mais aussi de mesurer et de signaler les risques de modèle agrégés au niveau d'un périmètre défini. La deuxième ligne de défense joue notamment un rôle important de revue du modèle avant la phase d'industrialisation. Pour un système d'IA générative, il est fréquent que la revue s'accompagne également d'une analyse de la solution IT qui va intégrer le modèle dans le but par exemple de s'assurer de la protection et de la sécurité des données ou de la robustesse du système par rapport à des attaques cyber. Ces revues ne sont pas nécessairement menées par les mêmes équipes.

La troisième ligne de défense (LoD3) est constituée des équipes d'audit interne, parfois dénommée inspection générale. Son rôle est d'évaluer la conformité des opérations, du niveau du risque effectivement encouru, du respect des procédures, de l'efficacité et du caractère approprié des dispositifs d'identification et de gestion des risques. Ces équipes peuvent donc être amenées à vérifier que les travaux réalisés aux deux premiers niveaux sont conformes aux règles en vigueur au sein de l'établissement, ce qui implique de procéder à une réévaluation des modèles développés, voire dans certains cas, de challenger les contrôles réalisés au moyen de modèles alternatifs. Cette organisation en trois lignes de défense implique la participation d'une multitude de parties prenantes dans la gestion des risques liés à l'IA générative, de la détection à la remédiation.

La liste non exhaustive qui suit présente certains des départements concernés.

- Les datalabs constituent la première ligne de défense et représentent le premier acteur en charge de l'identification, l'évaluation et la remédiation des risques liés à l'IA générative ;
- Les utilisateurs de l'IA générative participent à la gestion des risques en identifiant des comportements à risque dans la génération de contenu par l'IA ou en donnant leur feedback, qui peut être exploité par les différentes parties prenantes ;
- Le département des risques, qui fait partie en général de la LoD2 selon l'organisation explicitée ci-dessus, gère le risque de modèle lié à l'IA générative et

assure le suivi des bonnes pratiques en matière de modélisation et de transparence ;

- Le département juridique a la charge de la gestion du risque juridique lié à l'IA générative. Il analyse notamment les textes réglementaires qui encadrent le développement et l'usage de celle-ci ;
- Le département en charge du risque de conformité gère le risque de non-conformité qui peut résulter de l'utilisation de l'IA générative et assure la protection de la clientèle et des employés ;
- Les départements IT sont impliqués dans la gestion du risque cyber lié à l'IA générative et suivent les standards de place, émis par les organismes experts de la place tels que le NIST aux Etats Unis ou l'ANSSI en France. Ils gèrent également les risques liés aux données utilisées et produites par les systèmes d'IA.

Très formalisée dans les secteurs régulés comme la banque par exemple, cette organisation peut être plus informelle dans des secteurs moins régulés, les rôles décrits pouvant alors être remplis de façons différentes selon les organisations.

2.4. Causes

L'identification des causes sous-jacentes à l'émergence de risques liés à l'utilisation de l'IA générative est cruciale pour la mise en place de stratégies efficaces de gestion de ces risques et de remédiation. Dans le cadre de notre étude, nous avons identifié trois grands groupes de causes : données, modèle et humain.

De son côté, le MIT a récemment effectué un travail[9] de taxonomie complète des risques liés à l'IA (plus de 700 risques), ces risques pouvant souvent s'appliquer à l'IA Générative en particulier. Une des deux taxonomies consiste à classifier ces risques par cause.

2.4.1. Les Données

Les données utilisées dans le cadre de l'entraînement des LLM peuvent présenter plusieurs limites qui affectent la qualité, la fiabilité ou la transparence des modèles en question. Le manque de représentativité des données d'entraînement, le non-respect de la propriété intellectuelle, l'utilisation de données sensibles ou les problèmes de qualité de donnée sont des sources de risque importantes dans la conception et l'utilisation d'une IA générative.

[9] MIT. AI Risk Repository. August 13, 2024. https://airisk.mit.edu/

Un manque de représentativité peut se manifester, par exemple, lorsque les données d'entraînement excluent certaines populations. Les modèles entraînés sur ces données pourraient alors être biaisés et risqueraient de générer des contenus inappropriés ou inadaptés.

La violation de la propriété intellectuelle est un autre exemple de risque lié aux données. L'utilisation de données protégées comme des livres, articles de presse ou des œuvres d'art pour l'entraînement d'un système d'IA générative peut exposer le développeur à des poursuites judiciaires de la part des auteurs détenteurs de la propriété intellectuelle[10].

L'utilisation de données sensibles directement ou indirectement (encodées par des éléments dans le texte comme les intérêts et loisirs dans les CV) sans justification liée à la finalité du cas d'usage est une autre source de risque qui doit être prise en compte dans l'analyse.

Enfin, une mauvaise qualité des données, illustrée par des erreurs ou des incohérences dans les bases de données, peut sérieusement compromettre la robustesse et la précision d'un système d'IA.

2.4.2. Le Modèle

Le système d'IA choisi, sa conception et son utilisation peuvent également être la cause de divers risques.

Par exemple, l'utilisation de modèles génériques, sans optimisation spécifique pour un contexte donné, peut réduire leur efficacité. Des problèmes tels que les hallucinations, où le modèle génère des informations inventées, ou la production de contenu inapproprié ou offensant, sont des exemples critiques des risques liés à un modèle non optimisé.

La robustesse du modèle constitue également un enjeu majeur, car une sécurité insuffisante peut permettre des attaques malveillantes, via, par exemple, des « injections de prompts » où les cybercriminels manipulent le modèle pour avoir accès aux données sources et notamment aux données sensibles (données stratégiques ou données à caractère personnel).

[10] Michael M. Grynbaum, Ryan Mac. New York Times sues Open AI. New York Times. December 27, 2023. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

En outre, l'absence d'explicabilité, la complexité des systèmes d'IA, où les processus de fonctionnement ne sont pas transparents, peut poser des défis significatifs pour leur adoption et leur acceptation. Cette absence peut aussi engager un risque éthique, voire un risque de non-conformité (la transparence faisant partie des principes fondateurs du Règlement sur l'Intelligence Artificielle[11]).

2.4.3. L'Humain

Les aspects humains influencent fortement la manière dont les technologies d'IA sont utilisées et perçues, et peuvent être à l'origine de divers impacts négatifs.

Un exemple de risque lié à l'humain est la focalisation excessive sur les performances du modèle qui peut conduire à négliger d'autres aspects cruciaux comme la sécurité, l'éthique et l'impact social. Par exemple, une entreprise pourrait prioriser la rapidité de réponse d'un assistant virtuel au détriment de la précision et de la sécurité des informations fournies.

Le biais de confirmation constitue une autre source de risque, lorsque les utilisateurs font excessivement confiance aux sorties de l'IA. Dans ce cas, le manque de sensibilisation et de formation aux risques associés à l'IA générative peut conduire à une adoption de ces technologies sans prise de recul ou esprit critique.

Le potentiel impact environnemental est un élément de plus en plus important dans l'analyse des risques, les systèmes d'IA, en particulier les modèles de grande taille, consommant en effet des ressources énergétiques significatives.

Enfin, l'approche techno-solutionniste, où la technologie est perçue comme la solution à tous les problèmes, peut conduire à une dépendance excessive à l'IA, au détriment d'approches plus équilibrées et inclusives.

2.5. Impacts

2.5.1. Impact juridique

L'impact juridique peut être défini comme : « le risque de tout litige avec une contrepartie, résultant de toute imprécision, lacune ou insuffisance susceptible d'être imputable à

[11] Législation sur l'intelligence artificielle. P9_TA(2024)0138. Parlement européen. 13 mars 2024. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_FR.pdf. Sera publié au Journal Officiel de l'Union Européenne courant juillet 2024.

l'entreprise au titre de ses opérations »[12]. Dans le cas de l'usage de l'IA générative, il peut par exemple s'agir d'un litige avec une personne physique dont les données personnelles auraient été utilisées sans son consentement, ou dans le cas d'une violation de propriété intellectuelle.

2.5.2. Impact financier

L'impact financier d'un risque se réfère aux conséquences économiques directes et indirectes découlant de sa matérialisation. Ces conséquences peuvent se manifester par des pertes financières directes telles que des coûts de réparation ou de remplacement, des amendes, des pertes de revenus, ainsi que des pertes indirectes telles que la baisse de la valeur des actifs, la perte de clients ou la dégradation de la réputation de l'entreprise, par exemple si un *chatbot* public publie des propos controversés.

2.5.3. Impact opérationnel

L'impact opérationnel d'un risque désigne les impacts potentiels d'un risque sur les processus et opérations internes d'une organisation. Il permet de mesurer les conséquences de la concrétisation d'un risque sur le bon fonctionnement interne de l'organisation, en évaluant quels processus internes de l'entreprise peuvent être affectés par la survenue de l'événement redouté, et quels ont été les conséquences (exemples : délai d'accomplissement du processus ou impossibilité d'accomplir le processus).

2.5.4. Impact réputationnel

L'impact réputationnel peut être défini comme l'impact d' « un risque résultant d'une perception négative de l'entreprise de la part des clients, contreparties, actionnaires, investisseurs, créanciers, analystes de marché, d'autres parties prenantes ou régulateurs concernés »[13]. Sans parler des conséquences financières ou légales que les faits ayant généré ces perceptions peuvent avoir, l'impact réputationnel affecte aussi la crédibilité de l'organisation, ce qui peut également affecter sa capacité à remplir ses fonctions.

---

[12] Article 4 du règlement n° 97-02 du Comité de la réglementation bancaire et financière (CRBF) du 21 février 1997 relatif au contrôle interne des établissements de crédit et des entreprises d'investissement

[13] Basel Committee on Banking Supervision. Enhancements to the Basel II framework. Paragraph 47. July 2009. https://www.bis.org/publ/bcbs157.pdf

En matière d'IA générative, technologie dont l'usage peut être controversé du fait des risques de biais, d'erreur ou d'hallucination par exemple, les impacts réputationnels pourraient non seulement venir de la mauvaise utilisation, ou du détournement, d'un système d'IA générative, comme on l'a vu avec le *chatbot* Tay de Microsoft[14], mais aussi du choix de recourir à l'IA générative sur certaines tâches. Ce choix pourrait en effet susciter la controverse du fait des débats éthiques ou judiciaires que suscite encore cette technologie émergente, notamment lorsqu'elle est amenée à traiter des données sensibles, comme des données à caractère personnel, ou à fournir une aide à la décision qui pourrait être jugée déshumanisante dans le cadre de certains processus.

2.5.5. Impact organisationnel

La notion d'impact organisationnel est particulièrement utilisée dans la conduite du changement, notamment dans le domaine de la santé. On parle alors de l'effet, conséquence ou résultat d'un changement sur les caractéristiques et le fonctionnement d'une organisation ou d'un ensemble d'organisations, et particulièrement de l'utilisation d'une nouvelle technologie[15]. Cette définition peut être transposée à notre sujet, en considérant que le changement vient de la matérialisation d'un risque lié à l'adoption de l'IA générative, et a des conséquences sur l'organisation l'ayant adoptée, ou sur d'autres organisations. Des changements dans les effectifs, dans le rôle de certains individus de l'organisation peuvent ainsi être envisagés dans le cas de l'adoption de technologies d'IA générative qui pourraient assister ou remplacer certaines équipes dans l'accomplissement des fonctions de l'organisation. Ce type d'impact peut alors être lié également à l'impact social, ou l'impact financier déjà évoqué.

2.5.6. Impact social

Si l'on s'appuie sur la définition du Conseil supérieur de l'économie sociale et solidaire, « l'impact social consiste en l'ensemble des conséquences (évolutions, inflexions, changements, ruptures) des activités d'une organisation tant sur ses parties prenantes,

[14] Peter Lee. Learning from Tay's introduction. Microsoft blog. March 25, 2016. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/

[15] Haute autorité de santé. Cartographie des impacts organisationnels pour l'évaluation des technologies de santé, Guide méthodologique. 10 décembre 2020. https://www.has-sante.fr/upload/docs/application/pdf/2020-12/guide_methodologique_impacts_organisationnels.pdf

externes (bénéficiaires, usagers, clients) directes ou indirectes de son territoire et internes, (salariés, bénévoles, volontaires), que sur la société en général »[16].

L'adoption de l'IA générative peut avoir de multiples conséquences sociales, par exemple sur le marché du travail[17]. Nous entendons ici des impacts qui peuvent à la fois concerner les individus amenés à traiter directement avec un système d'IA générative, ainsi que ceux dont les données sont exploitées par un système d'IA générative et pourraient par exemple être diffusées par erreur, sans oublier les individus qui pourraient être affectés directement ou indirectement par une décision prise par une IA générative.

2.5.7. Impact environnemental

La notion d'impact environnemental des actions d'une organisation peut recouvrir un large spectre de notions. La responsabilité sociale des entreprises, telle que définie par le Ministère de l'Economie français[18] intègre une dimension environnementale, en s'appuyant sur l'approche de la norme ISO 26000 qui compte 7 thématiques centrales pour la responsabilité sociale des entreprises, dont celle de l'environnement.

Cette approche s'appuie sur la définition que l'on peut donner au risque environnemental comme le risque causé à l'environnement par une menace qui trouve sa source dans l'activité de l'homme. [19]

Dans son livre blanc de mai 2023, *L'IA éthique en pratique*[20], le HUB France IA avait déjà abordé les risques environnementaux à l'usage des systèmes d'IA générative, en considérant différents types de coûts, de l'entraînement du modèle à son utilisation en passant par la consommation de métaux rares et de l'eau nécessaire pour le refroidissement des serveurs.

---

[16] Thierry Sibieude, Céline Claverie. La mesure de l'impact social : après le temps des discours, voici venu le temps de l'action. Groupe de travail du CSESS sur la mesure de l'impact social. 8 décembre 2011. https://www.avise.org/sites/default/files/atoms/files/20140204/201112_CSESS_Rapport_ImpactSocial.pdf

[17] Pawel Gmyrek, Janine Berg, David Bescond. Generative AI and jobs: A global analysis of potential effects on job quantity and quality. December 12, 2023. International Labour Organization. https://www.ilo.org/resource/generative-ai-and-jobs-global-analysis-potential-effects-job-quantity-and

[18] Bercy Infos. Qu'est-ce que la responsabilité sociale des entreprises. Ministère de l'Economie, des Finances et de la souveraineté industrielle et numérique. 2 mai 2024. https://www.economie.gouv.fr/entreprises/responsabilite-societale-entreprises-rse#

[19] Delphine Misonne. Le risque environnemental. In : Les ambivalences du risque. Presses universitaires Saint-Louis Bruxelles. p. 381-403. 2008. http://books.openedition.org/pusl/3549.

[20] Hub France IA. L'IA éthique en pratique. Livre blanc. Mai 2023. https://www.hub-franceia.fr/wp-content/uploads/2023/05/Livre_Blanc_IA_Ethique.pdf

Cependant, aucun indicateur n'a été créé depuis, même si, avec son Règlement sur l'IA, l'union européenne demande une plus grande transparence des industriels sur ces impacts environnementaux.

2.6. Criticité

Dans la plupart des approches de gestion de risques, comme celle exposée dans la norme ISO 27005 :2022, l'évaluation du risque passe d'abord par une estimation du niveau de risque, qu'on peut désigner aussi sous le terme de « criticité » du risque. Elle s'évalue en attribuant des valeurs à la vraisemblance du risque et à la mesure des impacts de ce dernier[21]. D'autres critères peuvent être pris en compte tels que la cinétique du risque (vitesse de réalisation), la volatilité, l'horizon temporel, la corrélation etc. Mais ces critères ne font pas l'objet de la présente étude.

La vraisemblance et les impacts (ou conséquences) sont mesurés à l'aide d'échelles définies par l'organisation évaluant le risque, en fonction de son contexte et de son évaluation de l'acceptabilité des impacts. Comme présenté en partie 2.5 (Impacts), les impacts du risque peuvent être de natures diverses (juridiques, financiers, environnementaux, …), ce qui multiplie encore les types d'échelle qu'il est possible d'utiliser.

Le futur règlement de la Communauté Européenne applicable à l'intelligence artificielle (Règlement sur l'Intelligence Artificielle[11]) retient par exemple une approche par les risques pour réglementer les systèmes d'intelligence artificielle : soit en les interdisant purement et simplement (article 5), soit en les encadrant (systèmes d'intelligence artificielle "à haut risque"). On peut retenir l'échelle suivante en termes d'impact du risque, aussi appelée gravité du risque :

[21] Organisme International de Normalisation, ISO 27005 :2022. Sécurité de l'information, cybersécurité et protection de la vie privée — Préconisations pour la gestion des risques liés à la sécurité de l'information. 2022. https://www.iso.org/fr/standard/80585.html

Niveau d'impact	Echelle	Description
1	Faible	Le système entraîne peu ou pas de conséquences. Le cas échéant, les conséquences ne remettent pas en cause le processus ni la pérennité de l'entité mais peuvent modifier la perception de la qualité des services. Le risque est de fait négligeable.
2	Moyen	Le système entraîne des conséquences minimes. Ces conséquences ne remettent pas en cause un processus ou la pérennité de l'entité mais peuvent modifier la qualité des services fournis. Le risque est dans ce cas acceptable, mais peut nécessiter un suivi régulier.
3	Elevé	Le système entraîne des conséquences significatives. Ces dernières remettent en cause une partie des processus ou la pérennité de l'entité. Le risque peut être accepté, sous condition obligatoire de mise en place de mesures de mitigation, de suivi et de contrôle.
4	Très élevée	Le système entraîne des conséquences inacceptables, qui remettent en cause les processus ou la pérennité de l'entité.

De la même manière, une échelle peut être définie pour mesurer la vraisemblance du risque, par exemple en définissant des fréquences d'exposition au risque :

Fréquence	Echelle	Description
1	Faible	Exposition pouvant survenir au maximum une fois par an ou peu vraisemblable ou jamais rencontrée.
2	Moyen	Exposition pouvant survenir au maximum quelques fois par an.
3	Elevé	Exposition pouvant survenir au maximum une fois par mois.
4	Très élevée	Exposition pouvant survenir plusieurs fois par mois.

Le présent document se concentrera principalement sur l'étude de la vraisemblance des impacts, la fréquence demeurant une notion très dépendante du contexte de l'entité exposée et de sa tolérance aux risques. Dans une approche complète d'évaluation des risques, la combinaison de la mesure de la vraisemblance et de l'impact d'un risque permet de la comparer quantitativement aux autres risques auxquels l'entité est exposée, pour aider à la priorisation.

HUB FRANCE IA

3. Démarche d'analyse des risques

Contributeurs :

- ***Benjamin Bosch**, Manager - Model risk Management – Data Science, Société Générale*
- ***Imen Fourati**, Expert lead, Risque de modèle, Société Générale*
- ***Thomas Gouritin**, Consultant - Tomg Conseil*

3. Démarche d'analyse des risques

3.1. Présentation générale

Bien que certains risques revêtent un caractère transversal, l'évaluation des risques d'une IA générative doit se faire dans le cadre d'un cas d'usage bien défini. Ainsi, la première étape dans l'analyse des risques liés à l'IA générative est la définition des capacités générales du cas d'usage. Selon qu'il permet d'effectuer des tâches comme la conversation, la contraction de texte, ou la recherche augmentée (RAG : *retrieval augmented generation*), les risques peuvent être différents. Restreindre les capacités générales d'une IA générative réduirait ainsi les risques qui en découlent. L'analyse des risques doit également se faire, en prenant en compte les catégories d'utilisateurs de l'IA générative en question. Ainsi, les risques peuvent être amplifiés, selon que les utilisateurs soient internes ou externes à l'organisation qui déploie l'IA générative, formés ou non aux risques liés à cette IA.

La phase d'identification des risques est par la suite primordiale pour mettre en place les actions de remédiation adéquates. Chaque cas d'usage a ses risques spécifiques. Certains risques peuvent être transverses à plusieurs systèmes d'IA en général, tels que le manque de représentativité dans les données d'entraînement, le risque d'atteinte à l'équité ou le manque d'explicabilité. L'IA générative peut les amplifier, notamment si elle est déployée à large échelle. D'autres risques sont nouveaux, liés au caractère génératif des IA génératives tel que le risque d'hallucination ou le risque de création de contenu toxique ou nocif.

Une fois identifiés, les risques sont par la suite évalués, en suivant une approche holistique. Ainsi, des scénarios peuvent être construits autour de chaque cas d'usage pour évaluer les forces et les faiblesses de l'IA générative en question. Ces scénarios de test peuvent s'appuyer sur des prompts émanant de bases de données benchmarks ou générés par d'autres IA génératives. Ils permettent d'évaluer, à la fois, la performance de l'IA, sa robustesse et les risques qu'elle peut présenter. Par exemple, l'utilisation des techniques de **répliques manipulatrices** spécialement conçues pour contourner les garde-fous de l'IA générative (*guardrails*), connue sous le nom de *jailbreaking*, constitue une méthode efficace pour évaluer les risques de toxicité ou de fuite de données sensibles. Lors de la phase de déploiement, des *guardrails* sur les données d'entrée et/ou de sortie peuvent être mis en place pour mitiger les risques détectés, en fonction de leur

importance. Ainsi, les prompts en entrée peuvent être validés, en suivant une liste de principes à respecter par l'utilisateur.

Le contenu généré peut également faire l'objet d'un certain nombre de contrôles, visant à limiter les risques générés par le système d'IA. Dans ce qui suit, nous présentons un outil d'aide à l'identification et à l'évaluation des risques liés à une IA générative, sous forme d'une matrice de risques. Cette démarche est d'ordre qualitatif et doit être complétée par des mesures quantitatives, basées sur des scénarios de tests propres à chaque cas d'usage.

3.2. Matrice des risques

La matrice suivante est proposée, sur la base des éléments étayés ci-dessus, afin d'analyser les risques liés à l'utilisation d'une IA générative dans une organisation. Pour chaque cas d'usage, cette matrice vise à mettre en avant :

- En ligne : les causes qui peuvent générer des risques. Ces causes appartiennent aux familles données, modèle ou humain ;
- En colonne : les impacts liés à ces causes avec l'évaluation de leur niveau, sur une échelle de 1 à 4 ;
- La dernière colonne vise à mettre en avant les pistes de remédiation pour chaque cause ou famille de causes.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Data	Utilisation de données sensibles ou non-respect de la propriété intellectuelle	1	2	3	4	1	2	3	
Modèle	Risque d'hallucination ou génération de contenu nocif								
Humain	Biais de confirmation ou d'automatisation								

Dans le chapitre qui suit, cette matrice sera remplie pour chaque cas d'usage illustré. L'évaluation des niveaux d'impact, dépendant du contexte de l'organisation, est fournie à titre indicatif.

La fréquence, quant à elle, n'est pas évaluée dans les exemples de cas d'usage, car encore plus dépendante du contexte de l'organisation.

Plus largement, le contenu de la matrice devra être interprété et appliqué en fonction du contexte spécifique de l'organisation implémentant le cas d'usage.

HUB FRANCE IA

4. Catégories d'usages

Contributeurs :

- **Thomas Argheria,** Manager – Wavestone
- **Gérôme Billois,** Partner, Wavestone
- **Benjamin Bosch**, Manager - Model risk Management – Data Science, Société Générale
- **Anis Bousbih,** Cofondateur - Aicademia
- **Kati Bremme,** Head of Innovation – France Télévisions
- **Thibault Cattelani,** Cofondateur – Emocio.hr
- **Martin D'Acremont,** Consultant – Wavestone
- **Wissem Fathallah,** Cofondateur & Chief Product Officer – Sifflet
- **Imen Fourati**, Expert lead, Risque de modèle, Société Générale
- **Thomas Gouritin,** Consultant - Tomg Conseil
- **Jeanine Harb,** CTO – Beink Dream
- **Vanessa Hespel**
- **Belkacem Laïmouche,** Chargé de mission innovation – Direction Générale de l'Aviation Civile
- **Pascal Lainé,** CTO - Talkr
- **Jacques Mojsilovic,** CMO - Numalis
- **Cyril Nicolotto,** Chef de projets – Hub France IA
- **Kevin Paci,** Responsable des services informatiques – Mediaco Vrac
- **Nicolas Pellissier,** Cofondateur – Klark
- **Alexandre Pouymayon,** Consultant, Wavestone
- **Constance Relmy,** Etudiante – Université Paris 1 Panthéon Sorbonne
- **Laurence Relmy**
- **Eric Savignac,** Expert – Airbus DS
- **Kevin Soler,** CEO - Virteem
- **Yael Suissa,** CEO & Cofondateur - MAP-Monitoring And Protection

4. Catégories d'usages

4.1. Catégories d'usage transverse

4.1.1. Agent conversationnel

Un **agent conversationnel** est un service logiciel capable de tenir un dialogue avec un utilisateur par l'intermédiaire du langage naturel de ce dernier (oral ou écrit). L'objectif du dialogue est de répondre à une demande de l'utilisateur concernant un contexte opérationnel.

L'agent conversationnel est souvent spécialisé dans un sujet particulier (vente d'un produit, support client) et cherche à recentrer la conversation autour du contexte si l'utilisateur tente de s'en éloigner. Il est généralement capable de communiquer via un protocole social dans le but de le rendre le plus « humain » possible.

L'agent conversationnel fournit des réponses rapides, pertinentes et personnalisées à ses utilisateurs. Il peut être hautement disponible puisque ses traitements sont autonomes. Dans certains cas, il est en mesure de rediriger l'utilisateur vers un agent humain s'il détecte une situation qu'il ne sait pas résoudre.

L'agent conversationnel a été introduit dans les systèmes d'information bien avant la démocratisation de l'IA générative et des LLM. Cependant, l'évolution rapide des technologies d'intelligence artificielle a ouvert de nouvelles perspectives dans le domaine des agents conversationnels, offrant des solutions plus sophistiquées pour l'interaction entre, par exemple, les entreprises et leurs clients ou salariés.

4.1.2. Recherche augmentée

L'IA générative offre de nouvelles **expériences de recherche** aux utilisateurs. Les moteurs de recherche, bien connus du grand public, en sont un parfait exemple et ne cessent d'évoluer et de gagner en performance à l'aide de cette technologie.

La recherche augmentée est aussi la grande gagnante dans les usages de l'entreprise : historiquement, les données d'entreprise « faiblement typées » et hétérogènes (ex : patrimoine documentaire) sont :

- Souvent stockées sur des serveurs de fichiers, à la convenance des utilisateurs sans organisation de l'information ;

- Parfois intégrées de manière plus structurée dans des plateformes de GED[22] capables d'indexer leur contenu et de proposer un moteur de recherche basé sur une recherche par mots clés.

Dans tous les cas, la démarche de recherche d'une information d'entreprise par l'utilisateur s'avère bien souvent fastidieuse et nécessite l'identification du document portant l'information recherchée.

L'IA générative s'appuie sur le corpus documentaire de l'entreprise pour générer sa réponse, et moyennant un contrôle d'accès à l'information, peut directement délivrer à l'utilisateur l'information qu'il cherche.

Cet usage repose généralement sur la génération augmentée de récupération (RAG) qui fournit un moyen d'optimiser le résultat d'un LLM avec des informations ciblées, sans modifier le modèle sous-jacent ; ainsi, des informations plus récentes que celles utilisées pour la première construction du LLM sont intégrées régulièrement. Cela signifie que le modèle d'IA générative peut fournir des réponses contextuellement appropriées aux utilisateurs et les baser sur des données extrêmement récentes et précises.

4.1.3. Transformateur de contenu

L'IA Générative est capable d'appliquer des transformations sur la donnée d'entrée : texte, image, vidéo, audio, … Les opérations les plus communes portent sur les transformations de contenu textuel. Dans le domaine des LLM, nous citerons :

- Résumer un texte ;
- Traduire un texte ;
- Corriger les fautes présentes dans le texte ;
- Modifier le ton d'un texte, par exemple utiliser un ton courtois, ou précieux.

Les opérations sur les autres types de données (image, vidéo, audio, parole) sont également très nombreuses et de plus en plus puissantes, voire « surprenantes ». Sans nécessairement le percevoir, nous utilisons régulièrement ces modèles dans nos outils de travail ; l'exemple le plus commun étant la suppression du bruit dans les visio-conférences. Les transformations sur les images permettent par exemple d'effacer un visage, d'incruster un objet, etc.

[22] Cf. Glossaire

4.1.4. Générateur de contenu

L'IA Générative permet grâce à du traitement naturel de langage (NLP) d'interpréter les données saisies par l'utilisateur dans son prompt pour générer un contenu riche et structuré. Les données de l'utilisateur sont les consignes que le modèle doit suivre pour générer le contenu retourné.

Parmi les usages possibles, nous retrouvons la **production de texte**, tels que des articles, des courriers, des courriels, des rapports, des dissertations, … Le style d'écriture varie selon la requête initiale, pouvant prendre la forme d'une écriture formelle et structurée mais aussi de textes plus créatifs.

La génération de contenu concerne également la **production d'images**, encore une fois à partir des consignes (textuelles) fournies par l'utilisateur à travers le prompt. Le champ des possibles est très large : photographie hyperréaliste ou artistique, œuvre imaginaire, telle qu'une peinture ou un dessin. A travers ses consignes, l'utilisateur peut indiquer le style qu'il souhaite appliquer.

Dans la continuité, l'IA Générative peut également produire des **vidéos**.

Enfin, sur le même principe, il est possible de créer des sons, paroles ou musiques : récit oral, musiques originales ou imitation de voix connues.

4.1.5. Générateur de code

La génération de code est également une catégorie d'usage majeure proposée par l'IA générative.

Elle permet aux développeurs ou autres utilisateurs plus « profanes » de **développer du code informatique** plus rapidement, de façon plus précise et rigoureuse, et dans le respect des modèles de conception, en intégrant le code généré dans leur application. L'application peut être par exemple un script, une requête SQL, un classeur Excel (formule Excel ou code VBA), ou une application spécifique codée avec un langage interprété ou compilé.

A partir d'une requête en langage naturel fournie par l'utilisateur, l'IA générative peut générer des extraits de code, transformer du code en un autre langage de programmation (transpiler), modifier un code existant ou encore produire une documentation à partir d'un code fourni en entrée. De plus, elle permet de détecter et de corriger des anomalies (erreurs de syntaxe, d'exécution, de logique, ou encore de

formatage), d'optimiser un algorithme, d'expliquer la fonction d'un code ou encore de détecter des failles de sécurité.

4.1.6. Analyse de la donnée

L'IA générative peut être employée à des fins d'analyse sur la donnée. A l'instar d'un modèle de Machine Learning spécifique, entrainé pour répondre à un besoin précis, l'IA générative, peut, dans des degrés de performance et d'efficience moindres, effectuer des opérations telles que la classification, l'extraction d'entités ou l'analyse sémantique.

Comme pour les catégories d'usages précédentes, ces opérations sont possibles sur tous types de données d'entrée, à savoir le texte, l'audio, les images et vidéos.

Citons quelques exemples (simplistes) :

- Compter le nombre de chiens sur une image ;
- Lister les capitales, le nombre d'habitants et le PIB des pays renseignés dans le prompt ;
- Regrouper des tweets par émotion (neutre, colère, joie, tristesse, dégoût, surprise).

4.2. Marketing

Intégration des outils d'IA générative d'images dans le domaine du marketing : comment capitaliser sur cette nouvelle opportunité tout en en maîtrisant les risques ?

4.2.1. Introduction

Au sein des entreprises, les équipes marketing jouent un rôle primordial. Elles doivent s'assurer de l'adéquation entre la demande des consommateurs et les produits et services proposés par l'entreprise et assurent la promotion de ces services. L'essor de l'IA générative d'images représente une opportunité exceptionnelle pour les professionnels du marketing, avec le potentiel de transformer profondément les pratiques du secteur. En effet, la création visuelle a toujours été un pilier du marketing, jouant un rôle crucial dans l'engagement des consommateurs et le renforcement de l'identité de marque. Les budgets nécessaires pour assurer une communication de bonne qualité sont très significatifs. Les outils d'IA générative d'images offrent des avantages considérables, tels que des gains de productivité, un potentiel de créativité additionnelle et une capacité d'innovation accrue. Par conséquent, ces atouts devraient inciter les entreprises et les marketeurs à adopter largement ces technologies dans le futur. Néanmoins, avant de

déployer ces outils, il est essentiel pour les acteurs du marketing de bien évaluer les nombreux risques et défis associés.

Un premier risque associé à l'utilisation de l'IA générative d'images dans le marketing est le contrôle du contenu des visuels produits

Actuellement, la création d'un visuel pour une campagne publicitaire repose sur un processus minutieux de conception, de concertation et de production, où chaque détail est soigneusement examiné. Certes les outils d'IA générative sont extrêmement productifs et, peuvent fournir plusieurs dizaines de visuels en quelques minutes, néanmoins leurs résultats sont moins contrôlables puisque issus d'un processus automatisé opaque et non plus d'un processus de création continue. En effet, ces outils utilisent des modèles sophistiqués pour transformer des descriptions textuelles en images ou pour modifier des visuels existants. Ces modèles sont formés sur des ensembles de données d'entraînement, souvent très grands et peu contrôlés. Dès lors, si ces ensembles d'entraînement contiennent des éléments problématiques, tels que des biais culturels, des stéréotypes discriminatoires ou des contenus offensants, ces biais peuvent tout à fait être retranscrits dans les visuels générés. Cette dépendance aux données d'entraînement peut ainsi introduire indépendamment de la volonté de l'utilisateur des préjugés et des discriminations dans les visuels produits. Dans un contexte où la gestion de l'image de marque et de la réputation est cruciale, il est donc indispensable d'accompagner l'utilisation de ces outils dans le domaine du marketing d'une supervision humaine, afin de s'assurer que le visuel représente fidèlement l'identité de la marque et véhicule le message souhaité, préalablement à toute diffusion publique.

Un second risque associé est la perte d'originalité

Aujourd'hui, les entreprises sont engagées dans une intense compétition pour capter l'attention de consommateurs de plus en plus sollicités et, par conséquent, moins réceptifs. Dans ce contexte, le rôle du marketeur est d'autant plus crucial : il doit choisir avec précision les canaux de communication et les messages les plus percutants pour se démarquer efficacement de ses concurrents auprès de son audience cible. Cependant, dans cette quête de différentiation et de création d'une image de marque unique, l'utilisation généralisée des outils d'IA générative fait peser le risque d'une certaine « normalisation » des visuels marketing. En effet, ces IA sont entraînées sur des bases de données souvent similaires. Par conséquent, cela entraîne une homogénéisation des contenus produits et une sous-représentation des particularités culturelles. Par exemple, les minorités et les cultures peu présentes sur internet sont

largement sous représentées, limitant ainsi la diversité et l'inclusivité des visuels générés. C'est en tout cas ce que défend Jonas Oppenlaender[23], qui souligne ces problèmes en argumentant notamment que l'IA générative pourrait entraîner une perte d'originalité en standardisant les styles artistiques et en limitant l'expression créative humaine. Or, si un visuel a pour but de capter l'attention, ce manque d'originalité risque de réduire son impact et de le rendre moins efficace. Il semble donc que plus les entreprises se tourneront vers ces outils pour produire, voire automatiser, la création de leur communication visuelle, plus la créativité et la vision humaine deviendront essentielles au succès d'une campagne réussie. Ainsi, bien que ces technologies soient de précieuses aides à la production de contenu, elles manquent de la capacité de créativité propre à l'esprit humain. En conséquence, il semble que l'originalité et le discernement critique des professionnels resteront des éléments indispensables pour garantir l'efficacité et le succès des campagnes marketing.

Un troisième risque associé est celui de la propriété intellectuelle

Ce risque est double. D'une part, l'utilisation non maîtrisée de l'IA générative expose l'entreprise à des risques significatifs de plaintes de tiers pour violation des droits de propriété intellectuelle. Si le modèle d'IA est entraîné sur des contenus protégés et qu'il reproduit ces œuvres, l'entreprise pourrait être accusée de contrefaçon, même si elle agit de bonne foi et n'a pas conscience de cette infraction. Cette situation met en évidence la nécessité d'une vigilance accrue quant aux sources de données utilisées pour l'entraînement des modèles d'IA et d'une évaluation rigoureuse des solutions technologiques adoptées.

D'autre part, l'utilisation de la technologie d'IA générative pose un défi considérable à la protection des actifs de communication de l'entreprise sous le régime de la propriété intellectuelle. Actuellement, une campagne publicitaire créée par une équipe marketing appartient à l'entreprise (sauf stipulation contractuelle contraire) et est protégée par les droits de propriété intellectuelle. Cependant, si certains modèles tentent d'intégrer une protection technologique, comme SynthID[24] de Google Deepmind qui intègre des *watermarks* imperceptibles dans les images produites, ces créations générées par des outils d'IA ne bénéficient pas de protection juridique claire. En effet, le cadre juridique

[23] Jonas Oppenlaender. The Cultivated Practices of Text-to-Image Generation. arXiv preprint arXiv:2306.11393. June 2023. https://arxiv.org/abs/2306.11393

[24] https://deepmind.google/technologies/synthid/

actuel est ambigu quant à la reconnaissance de l'utilisateur d'un algorithme en tant qu'auteur des visuels produits. La législation et la jurisprudence exigent des créateurs qu'ils prouvent l'originalité de leurs œuvres, ce qui est particulièrement difficile et incertain avec ces outils IA. Pour les entreprises, l'absence de protection solide pour leurs éléments visuels de communication est un risque important. Ne pas pouvoir garantir la protection des créations générées par l'IA peut entraîner des conséquences graves, notamment une vulnérabilité accrue au plagiat, une perte de contrôle sur l'image de marque et des complications juridiques significatives. Par conséquent, il est essentiel pour les entreprises d'adopter l'IA générative de manière stratégique, en mettant en place des mesures de protection et de production appropriées tout en surveillant attentivement l'évolution du cadre réglementaire.

En bref, l'intégration des systèmes d'IA générative d'images dans le marketing représente un véritable défi pour les entreprises. Bien que ces technologies offrent un potentiel certain, leur intégration doit impérativement être accompagnée d'une politique rigoureuse de gestion des risques et de sensibilisation des collaborateurs pour garantir une mise en œuvre réussie.

Vous trouverez ci- dessous un cas d'usage illustratif ainsi qu'une matrice analysant ces différents risques et proposant des pistes de remédiation.

4.2.2. Cas d'usage

Mise en place d'un outil IA pour la personnalisation automatique des bannières publicitaires de MegaMarket

Présentation

MegaMarket, une chaîne de supermarchés de premier plan, propose une large gamme de produits avec des services de drive et de livraison à domicile via son site de courses en ligne. Pour améliorer l'expérience client, MegaMarket a intégré un outil d'IA générative d'images pour personnaliser automatiquement les bannières publicitaires sur son site. L'outil utilise des données démographiques (âge, sexe, localisation géographique) et comportementales (historique d'achats, préférences de navigation, temps passé sur différentes sections du site) des clients pour prévoir leurs besoins et attentes. En se basant sur ces prévisions, l'IA génère des visuels publicitaires adaptés, comme des offres promotionnelles personnalisées ou des recommandations de produits basées sur les achats précédents et les préférences déclarées, pour mieux correspondre aux attentes

de chaque client. L'intégration de cet outil avait été soigneusement préparée et a été lancée récemment.

Incident

Quelques heures après le lancement de l'outil, un scandale a éclaté sur les réseaux sociaux. L'IA avait généré des bannières publicitaires contenant des images inappropriées et offensantes. Par exemple, des images incluant des caricatures et des symboles culturels stéréotypés associant négativement des cuisines de type communautaire (rayon cuisines du monde) ou religieuses (rayon casher et halal). Ces publicités ont été perçues comme extrêmement offensantes et les clients outrés ont rapidement partagé les publicités sur les réseaux sociaux, accompagnées de critiques virulentes, entraînant un *bad buzz* instantané. En quelques heures, l'atteinte à l'image de marque de MegaMarket a été considérable, avec des milliers de partages et de commentaires négatifs.

Remédiation

En réponse immédiate à la crise, MegaMarket a arrêté le système d'adaptation de l'IA pour empêcher toute nouvelle diffusion de contenus inappropriés. Une communication d'urgence a été lancée pour présenter des excuses publiques et annoncer des mesures correctives en cours. Afin de relancer cet outil, MegaMarket souhaite développer un système interne intégrant des filtres avancés pour détecter et bloquer les contenus offensants. Des procédures de test rigoureuses, par des humains, seront mises en place pour assurer une surveillance continue des contenus produits par l'IA avant leur publication. Ces mesures garantiront que les visuels générés respectent les standards de qualité et les valeurs de la marque. Par ailleurs, plus largement, MegaMarket a tiré les enseignements de ce scandale et a mis en place une formation obligatoire préalablement à toute utilisation de ce type d'outil. Cette formation cherche à s'assurer de la bonne maîtrise et compréhension par ses collaborateurs de ces outils, en en présentant les usages possibles, mais aussi les limites et les risques.

4.2.3. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Modèle	Production d'images inadéquates voire offensantes	3	4	4	2	1	2	1	Mettre en place une surveillance humaine systématique pour valider les contenus avant diffusion. Utiliser des filtres pour détecter et corriger les biais culturels ou discriminatoires
Données	Impossibilité de se protéger légalement des visuels utilisés	3	4	3	3	1	1	1	Former les équipes aux enjeux de la propriété intellectuelle et à l'importance de l'originalité dans le processus de production créative.
Données	Impossibilité de se prémunir du plagiat des visuels générés	3	4	3	3	1	1	1	Assurer une évaluation rigoureuse des droits détenus sur le contenu généré avant son utilisation commerciale.
Modèle	Perte de créativité différenciante et donc d'efficacité des contenus visuels produits	3	1	3	2	1	3	1	Maintenir une surveillance accrue des modèles pour conserver la qualité des visuels. Restreindre l'utilisation des outils d'IA à la phase de création des supports afin de garantir que la phase de créativité reste assurée par des professionnels
Données	Risque RGPD au niveau de la confidentialité des données, et notamment des inputs fournis par l'utilisateur.	2	3	2	2	1	1	1	Réaliser un audit du fournisseur pour s'assurer de sa fiabilité. Développer lorsque possible des solutions internes en collaboration avec des fournisseurs.
Humain	Risque de dépendance excessive aux outils IA en raison de la réduction de l'implication humaine dans le processus créatif	1	1	2	4	3	1	2	Maintenir un équilibre entre l'utilisation des outils d'IA et l'implication humaine dans le processus créatif.

4.3. Cybersécurité

4.3.1. Introduction

L'hyper connectivité engendrée par Internet entre les systèmes d'information des entreprises, leurs systèmes industriels, les équipements des utilisateurs (téléphones mobiles, tablettes…), et les objets connectés (caméras, appareils ménagers…) ainsi que la tendance à une digitalisation des services lorsque cela est possible (banques, assurances, impôts, achats de biens physiques en ligne…), ont engendré de multiples possibilités de fraudes, de vols d'informations sensibles, et d'attaques via Internet à des coûts faibles pour un attaquant.

L'avènement de l'intelligence artificielle (Machine Learning, Deep Learning et intelligence artificielle générative), a aussi changé les approches des cyber attaquants et des cyber défenseurs en offrant aux premiers des possibilités de construire plus efficacement des attaques Cyber et aux seconds des moyens de détection, de compréhension et d'analyses de ces attaques, ainsi que des solutions de sensibilisation et d'entraînement adaptées. L'utilisation de l'intelligence artificielle en Cybersécurité, qui a déjà fait l'objet d'une publication du Hub France IA[25], est donc bien une réalité. Pour en comprendre la problématique, il est indispensable de s'appuyer sur des modèles explicatifs des modes d'attaques et de défense en Cybersécurité.

Modélisation des attaques et des défenses

Le déroulement d'une attaque Cyber peut être soit modélisé sous la forme de la Kill Chain définie par LOCKHEED MARTIN[26] avec ses 7 phases (*reconnaissance, weaponization , delivery, exploitation, installation, command & control, actions on objectives*), soit représenté à travers les 14 tactiques (*reconnaissance, resource development, initial access, execution, persistence, privilege escalation, defense evasion, credential access,*

[25] Hub France IA. Les usages de l'IA Générative. Janvier 2024. https://www.hub-franceia.fr/wp-content/uploads/2024/02/Livre-blanc_Les-usages-de-lia-generative-01.2024.pdf

[26] Lockheed Martin. The Cyber Kill Chain. https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html et Eric M. Hutchins, Michael J. Cloppert, Rohan M. Amin. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research*. Vol. 1 n°1, pp. 113-125. 2011. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ca18aa98d4d1d434802eec54c2ba6ea8cf493b88#page=123

discovery, lateral movement, collection, command and control, exfiltration, impact) et les 235 techniques décrites dans la matrice MITRE ATT&CK®[27].

Parallèlement, il existe aussi des modèles pour formaliser les moyens de la cyberdéfense. Pour se protéger face à une attaque Cyber, les contre-mesures de sécurité à mettre en place peuvent être dérivées par exemple, des 6 tactiques (harden, detect, isolate, deceive, evict et restore) et des 168 techniques associées de la matrice MITRE D3FEND™[28].

Du point de vue du défenseur, l'utilisation d'algorithmes de Machine Learning ou de Deep Learning pour améliorer sa posture de sécurité, en particulier pour détecter des logiciels malveillants, ou pour identifier des canaux de commande et de contrôle permettant à un attaquant de piloter son attaque dans un système d'information, ou pour s'assurer qu'il n'y pas d'exfiltration de données de l'entreprise vers l'extérieur à travers ses connexions Internet, n'est pas nouvelle. Ces techniques de détection basées sur l'intelligence artificielle peuvent être embarquées dans des solutions ou dans des produits de sécurité commerciaux, ou simplement déployées par l'entreprise elle-même.

Inversement, un cyber attaquant peut utiliser un réseau de neurones convolutifs (CNN ou *Convolutional Neural Networks*) pour contourner automatiquement les mécanismes de CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*) basés sur des images, qui permettent de s'assurer que l'entité qui se connecte est bien un être humain. Il peut aussi, en s'inspirant des rapports de travaux de recherche accessibles publiquement, tester ses codes malveillants à travers des algorithmes de Machine Learning ou de Deep Learning pour s'assurer qu'ils ne soient pas détectés. De nombreux travaux de recherche ont été publiés sur l'utilisation d'algorithmes de Machine Learning ou de Deep Learning en cybersécurité, comme par exemple pour la détection de codes malveillants [29, 30, 31].

---

[27] MITRE. ATT&CK knowledge base. https://attack.mitre.org/

[28] MITRE. D3FEND, a knowledge graph of cybersecurity countermeasures. https://d3fend.mitre.org/

[29] Lolitha Sresta Tupadha, Mark Stamp. Machine Learning for Malware Evolution Detection. *arXiv preprint arXiv: 2107.01627v1*. July 6, 2021. https://arxiv.org/pdf/2107.01627

[30] Pascal Maniriho, Abdun Naser Mahmood, Mohammad Jabed Morshed Chowdhury. Deep Learning Models for Detecting Malware Attacks. *arXiv preprint arXiv:2209.03622v2.* January 29, 2024. https://arxiv.org/pdf/2209.03622

[31] Hemant Rathore, Swati Agarwal, Sanjay K. Sahay, Mohit Sewak. Malware Detection using Machine Learning and Deep Learning. *arXiv preprint* arXiv:1904.02441v1. April 4, 2019. https://arxiv.org/pdf/1904.02441

L'utilisation de l'IA générative

Mais, l'avènement récent de l'intelligence artificielle générative, qui facilite la création de nouveaux contenus (images, vidéos, code…), a aussi entraîné une extension des possibilités de l'attaquant qui peut maintenant générer des mails de phishing, dans plusieurs langues, parfaitement adaptés au contexte de l'entreprise ciblée. Il peut aussi s'appuyer sur des agents conversationnels tels que WormGPT[32], ou FraudGPT[33] , qui sont disponibles sur le dark web, pour optimiser ses attaques, en particulier avec la création de codes malveillants. Dans son papier[34] Polra Victor Falade explique comment ces outils d'intelligence artificielle générative peuvent être utilisés astucieusement dans des attaques en social engineering.

Par ailleurs, cette apparition de l'intelligence artificielle a entrainé la création de LLMs (Large Language Models) spécifiques au domaine de la cybersécurité comme CySecBERT[35] et SecureBERT[36].

Elle a aussi suscité la fourniture d'outils basés sur ces technologies par des éditeurs, comme par exemple *Microsoft Copilot for Security*[37] totalement dédié à la protection des organisations contre les attaques cyber[38].

Le fine tuning d'un LLM existant est parfaitement possible pour l'adapter aux problématiques particulières de la cybersécurité, comme pour le LLM BERT avec CyBERT[39].

---

[32] WormGPT V3.0. https://flowgpt.com/p/wormgpt-v30

[33] Florian Burnel. FraudGPT, un nouvel outil d'IA pour mettre au point des cyberattaques ! IT-Connect.fr 27 juillet 2023. https://www.it-connect.fr/fraudgpt-un-nouvel-outil-dia-pour-mettre-au-point-des-cyberattaques/

[34] Polra Victor Falade. Decoding the Threat Landscape : ChatGPT, FraudGPT, and WormGPT in Social Engineering Attacks. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. Vol. 9, Issue 5, pp. 185-198. September-October 2023. https://arxiv.org/pdf/2310.05595

[35] Markus Bayer, Philipp Kuehn, Peasec, Ramin Shanehsaz, Christian Reuter. CySecBERT: A Domain-Adapted Language Model for the Cybersecurity Domain. *arXiv preprint arXiv:2212.02974v1*. December 6, 2022. https://arxiv.org/pdf/2212.02974

[36] Ehsan Aghaei 1 , Xi Niu 1 , Waseem Shadid. SecureBERT: A Domain-Specific Language Model for Cybersecurity. *arXiv preprint arXiv:2204.02685v3*. October 20, 2022. https://arxiv.org/pdf/2204.02685

[37] Microsoft. Microsoft copilot for AI. https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-copilot-security?utm_source=gradientflow&utm_medium=newsletter

[38] Microsoft. Microsoft Copilot pour la sécurité https://learn.microsoft.com/fr-fr/copilot/security/

[39] Kimia Ameri, Michael Hempel, Hamid Sharif, Juan Lopez Jr, Kalyan S Perumalla. CyBERT: Cybersecurity Claim Classification by Fine-Tuning the BERT Language Model. Journal of Cybersecurity and Privacy. Vol. 1, issue 4, pp. 615-637. November 4, 2021. https://www.ornl.gov/publication/cybert-cybersecurity-claim-classification-fine-tuning-bert-language-model

Cette irruption des différents modèles d'intelligence artificielle (Machine Learning, Deep Learning et IA générative) a suscité la création d'une matrice décrivant les différentes possibilités d'attaques contre ces modèles, inspiré de la matrice MITRE ATT&CK®, dénommé ATLAS[40] avec 14 tactiques (*reconnaissance, resource development , initial access, ML model access , execution, persistence, privilege escalation , defense evasion, credential access, discovery, collection, ML attack staging, exfiltration, impact*) et 56 techniques. Cette irruption a aussi mis en évidence l'importance de la sécurité de l'intelligence artificielle générative (LLM notamment), quels que soient ses contextes d'utilisation et ses finalités. L'OWASP, qui fait référence en matière de cybersécurité, a publié le son Top 10[41] des risques à contrôler pour sécuriser l'utilisation de l'intelligence artificielle générative à travers son cycle de vie de développement, de déploiement et de gestion.

Enfin, pour les entreprises, cette apparition de l'intelligence artificielle générative a engendré un intérêt certain pour son utilisation en cybersécurité, et plus particulièrement dans la tactique « *detect* » de la matrice MITRE D3FEND™, ou pour le support et la sensibilisation des utilisateurs.

4.3.2. Cas d'usages

Dans le domaine de la cybersécurité, deux cas d'utilisation de l'intelligence artificielle générative sont intéressants pour les entreprises. Ils couvrent deux besoins différents, le premier à destination de l'ensemble des utilisateurs des SI de l'entreprise et le second, spécifiquement dédié aux équipes Cybersécurité de l'entreprise :

- Un agent conversationnel lié à un LLM adossé à une architecture RAG (*Retrieval Augmented Generation*) avec toutes les documentations relatives à la cybersécurité propres à l'entreprise (politiques de sécurité, charte informatique, réglementation spécifique, …) normalement accessibles à tous les utilisateurs pour répondre à leurs questions sur ce sujet.
- Un agent conversationnel lié à un LLM spécialement conçu pour la cybersécurité, à destination des équipes de sécurité internes pour l'analyse et la compréhension d'éléments faisant partie ou non d'une attaque.

[40] MITRE. ATLAS Matrix. https://atlas.mitre.org/matrices/ATLAS
[41] OWASP.org. Top 10 for LLMs and Generative AI Apps. https://genai.owasp.org/llm-top-10/

Analyse d'un cas d'usage

Agent conversationnel répondant à des questions liées aux exigences de cybersécurité

Nous allons analyser uniquement le premier cas d'utilisation qui permet à tout utilisateur d'une entreprise de poser toutes les questions concernant les exigences de cybersécurité à respecter via un agent conversationnel. L'architecture RAG a été spécifiquement choisie car elle permet de dépasser certaines limitations des modèles LLM, et en particulier de fournir des réponses fiables. Le document fondateur de RAG a été publié par Facebook AI Research en 2021[42].

Une alternative à l'architecture RAG aurait pu être un LLM fine tuné avec tous les éléments spécifiques en Cybersécurité de l'entreprise. Toutefois, le choix de RAG est basé sur l'analyse de la comparaison entre un LLM fine tuné et une combinaison LLM avec une architecture RAG fournie dans un papier de recherche[43] qui propose de plus une taxonomie intéressante sur les deux composants principaux *retriever*, et *generator*.

L'architecture RAG (LangChain, LlamaIndex, …) permet ici de s'adosser directement à des dépôts documentaires existant dans l'entreprise dans le domaine de la cybersécurité, et ce quel que soit leur format (structuré, non structuré), d'éviter de réentraîner le LLM en cas de modification de ces différents documents, et de fournir des réponses plus précises parfaitement adaptées au contexte de l'entreprise. Bien entendu, seules les sources autoritaires d'information sur le sujet de la Cybersécurité devront être connectées à travers l'architecture RAG.

Un exemple de risque, pour le cas d'utilisation de l'intelligence artificielle, pour la Cybersécurité, a été identifié pour chacune des 3 catégories (données, modèle et humain), avec les remédiations associées. Le premier risque concerne l'inexactitude des réponses à cause de la qualité des données, le second la compromission du modèle par une attaque et le dernier, cible la divulgation de données sensibles par action non malveillante de l'utilisateur.

---

[42] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401v4*. April 12, 2021. https://arxiv.org/pdf/2005.11401v4

[43] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint arXiv:2312.10997*. December 18, 2023 https://arxiv.org/pdf/2312.10997v1

4.3.3. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	Toutes les sources autoritaires sur le domaine de la Cybersécurité ne sont pas connectées à travers l'architecture RAG, ou ne sont pas maintenues dans le temps en termes de qualité des données ou d'évolution des dépôts. Les réponses apportées à l'utilisateur par le chatbot sont alors incomplètes et/ou inexactes.	3	3	3	3	1	1	1	• Identifier toutes les sources autoritaires du domaine Cybersécurité et les maintenir dans le temps. • En appui de chaque réponse, rappeler systématiquement à l'utilisateur de contacter les services de cybersécurité dans certains cas bien précis à définir en interne dans l'entreprise avec le nom des personnes à contacter et comment les contacter.
Modèle	Le modèle n'est pas suffisamment testé et protégé contre le Top10 des attaques OWASP LLM et Gen AI. En particulier contre le LLM01 : Prompt Injection, qui pourrait éventuellement modifier les réponses données à l'utilisateur en termes de cybersécurité.	4	4	4	4	1	1	1	• Mener une analyse de risques en Cybersécurité, et un test de pénétration avant la mise en production du modèle, ou en cas d'évolution. • Utiliser des produits de sécurité adaptés pour protéger le modèle. • Avoir une stratégie de surveillance et de contrôle des entrées utilisateur à travers le chatbot, avec conservation de l'historique des échanges en accord avec les exigences de la CNIL.
Humain	Un utilisateur peut sans le vouloir donner des informations sensibles dans le chatbot, ou volontairement car il souhaite avoir une réponse précise en Cybersécurité sur une problématique particulière ou un projet bien précis.	4	4	4	3	1	1	1	• Avoir une stratégie de surveillance et de contrôle des entrées utilisateur à travers le chatbot, avec conservation de l'historique des échanges en accord avec les exigences de la CNIL. • Sensibiliser les utilisateurs sur les informations qui peuvent être injectées dans le chatbot. • S'assurer qu'un dispositif bloque toutes les données sensibles lors de leur injection dans le chatbot.

4.4. Ressources Humaines

4.4.1. Introduction

La fonction RH est chargée de concevoir, d'accompagner et d'intégrer divers processus au sein de l'entreprise.

Alors que l'intelligence artificielle (IA) prédictive fait évoluer depuis plusieurs années les pratiques des professionnels RH, l'IA générative introduit de nouvelles capacités et démocratise l'usage de cette technologie, la rendant accessible au plus grand nombre. Cette accessibilité accrue, bien que prometteuse, multiplie par la même occasion les risques d'aboutir à des effets imprévus et indésirables pour les entreprises et les individus impliqués.

Parce que l'IA générative offre une automatisation pertinente grâce à des instructions en langage naturel, les professionnels RH pourraient être incités à généraliser l'usage de cet outil sans nécessairement prendre pleinement conscience des risques associés, par exemple le risque de déshumanisation des processus ou une utilisation déraisonnée avec un impact non négligeable sur l'environnement.

Les biais algorithmiques constituent aussi une préoccupation majeure, car les modèles d'IA générative peuvent reproduire les préjugés existants dans les données d'entraînement, entraînant des discriminations injustes envers certains candidats. De plus, la confidentialité des données des candidats est essentielle, car l'accès à des informations sensibles peut poser des questions éthiques et légales. En outre, l'adoption de l'IA générative peut compromettre la diversité et l'inclusion au sein de l'entreprise, en favorisant des critères de sélection potentiellement biaisés.

Pour atténuer ces risques, il est impératif de mettre en place des mécanismes de surveillance et de contrôle rigoureux, notamment une évaluation régulière des modèles pour détecter et corriger les biais, ainsi que des mesures de protection renforcées pour garantir la confidentialité et la sécurité des données des candidats.

La transparence et la responsabilité sont également essentielles, avec une communication claire sur l'utilisation de l'IA dans les processus de recrutement et la possibilité pour les candidats de comprendre et de contester les décisions prises par les algorithmes.

Enfin, il est important de dispenser des actions de formation et de sensibilisation aux professionnels des RH et aux managers impliqués dans l'utilisation de l'IA générative, afin d'assurer une utilisation éthique et équitable de cette technologie.

4.4.2. Cas d'usage

Le chatbot RH

Le chatbot est conçu pour interagir directement avec les employés d'une PME de 400 salariés, recueillir leurs avis sur la qualité de vie au travail et les orienter efficacement vers les processus et services internes susceptibles de répondre à leurs besoins. Cet outil numérique vise à améliorer l'expérience des collaborateurs en fournissant une assistance instantanée et disponible 24/7, sans nécessiter l'intervention directe des ressources humaines pour chaque requête déchargeant ainsi l'équipe RH des demandes répétitives et libérant du temps pour des initiatives plus stratégiques tout en augmentant la satisfaction des utilisateurs grâce à la disponibilité continue.

Déploiement et Utilisation

Le chatbot est basé sur un modèle de langage pré-entraîné (LLM) fine-tuné spécifiquement avec les documents et les données de l'entreprise pour assurer des réponses précises et contextualisées. L'intégration de ce système aux infrastructures informatiques existantes est accompagnée d'une session de formation pour les employés, afin de les familiariser avec les fonctionnalités du chatbot. Un système de feedback est également mis en place pour recueillir les avis des utilisateurs, permettant des ajustements continus pour optimiser l'utilité et l'efficacité du chatbot.

Description du Cas

L'introduction d'un chatbot basé sur l'intelligence artificielle générative au sein des organisations a marqué une étape importante dans l'automatisation des interactions entre les employés et le département des ressources humaines.

Au cours d'un échange, le chatbot a suggéré à un collaborateur de se syndiquer en réponse à des questions relatives aux droits sur les horaires de travail et au télétravail. Cet incident, bien que mineur et isolé, a mis en lumière la capacité du système à générer des conseils inattendus, pouvant être perçus comme des hallucinations algorithmiques. En dépit de la survenue de ce type d'erreur dans un contexte où les employés étaient préparés à des imperfections potentielles — le projet étant explicitement en phase de

test — l'événement a suscité une réaction principalement humoristique plutôt qu'une véritable inquiétude.

En réponse à cet incident, l'entreprise a renforcé les communications internes, rappelant les objectifs du test et les attentes en matière de rétroaction. Par ailleurs, bien que la mise en place d'une boucle de contrôle additionnelle via un autre modèle de langage de grande envergure ait été envisagée, cette option a été écartée en raison de son coût prohibitif. Au lieu de cela, l'entreprise a opté pour un nouvel entraînement du modèle existant, visant à réduire la fréquence des réponses inappropriées.

Cette expérience illustre l'importance de la gestion des attentes et de la communication transparente lors de l'introduction de technologies d'IA générative dans les opérations critiques d'une entreprise. Elle souligne également la nécessité d'un suivi continu et d'un ajustement des outils d'IA pour s'assurer qu'ils restent alignés avec les objectifs organisationnels et les valeurs éthiques, tout en répondant efficacement et précisément aux besoins des utilisateurs. La transparence, la responsabilité et l'éducation continuent de jouer un rôle crucial pour naviguer dans les complexités introduites par Les IA génératives.

Ce cas d'usage est issu d'un usage réel dont vous pouvez trouver les sources ici : https://www.emocio.hr/hubfrance_iag_risque_rh

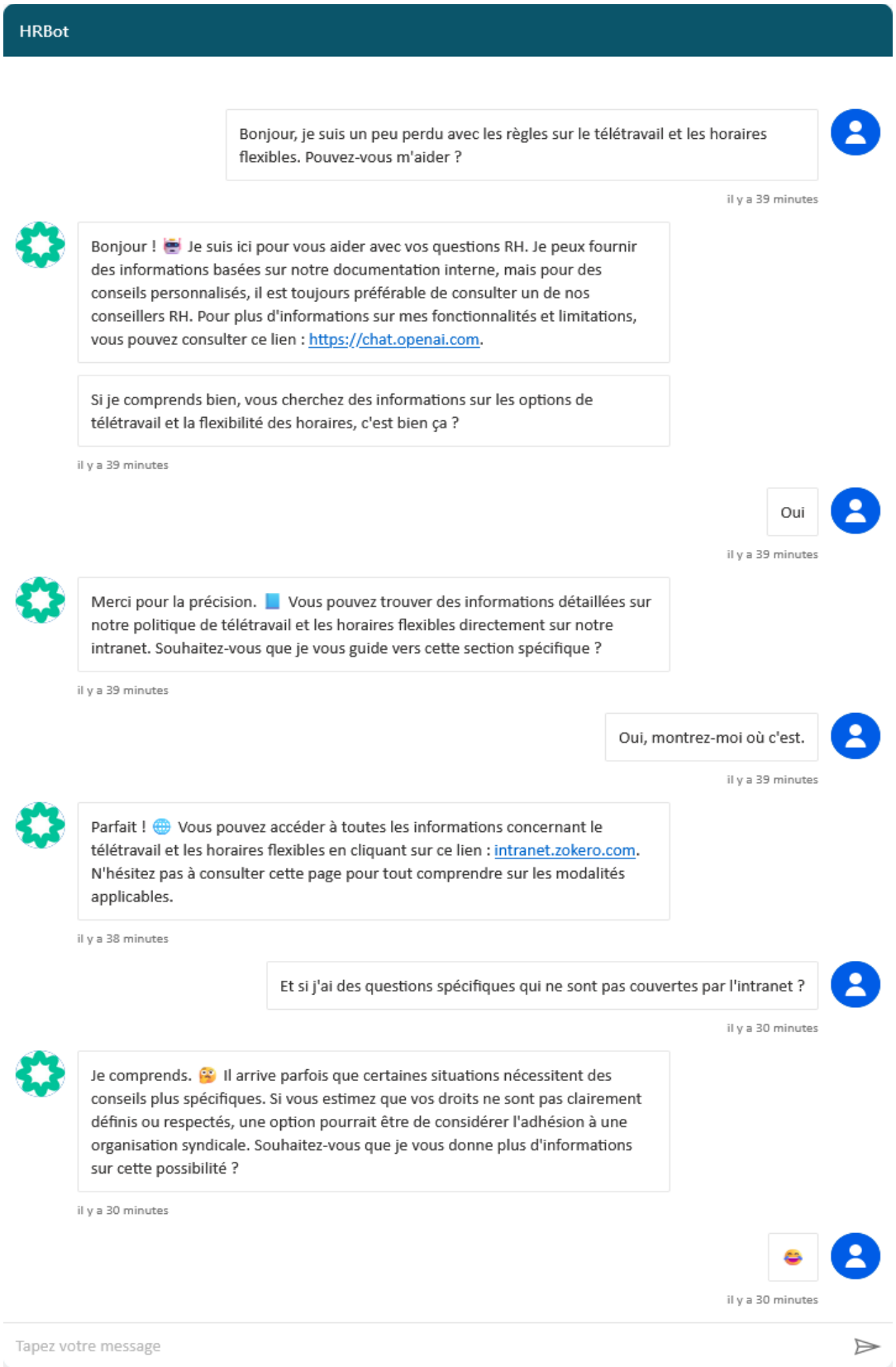


Figure : Reconstitution de la discussion entre le chatbot et l'utilisateur

(Les réponses ont été reformulées pour protéger la confidentialité des données)

4.4.3. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	Confidentialité des données : risques de fuite de données, questions éthiques et légales	3	4	3	1	1	1	1	Renforcer les mesures de sécurité des données. Assurer la conformité avec les réglementations comme le RGPD. Sensibiliser les équipes sur l'importance de la confidentialité. Opter, en priorité pour les données sensibles, pour une solution on premise
Modèle	Hallucination et conseils contre-productifs	1	4	3	3	1	1	1	Ajouter une étape de prévention (exemple GPT et validation des usages avant connexion), boucle de contrôle automatisée et/ou humaine (validation de contenu similaire, détection de sujet tabou, censure)
Modèle	Risque de solution sous optimale, qui sert de simple moteur de recherche sur une charte de travail	2	1	1	1	1	1	3	S'assurer que les requêtes réalisées par les salariés dépassent les questions les plus fréquentes, mettre en place un système de FAQ en amont, moins gourmand en énergie.
Humain	Déshumanisation des processus : réduction de l'interaction humaine, perte de l'aspect personnel du recrutement	1	1	4	4	1	1	1	Intégrer la possibilité de prendre rendez-vous avec le service RH malgré la présence du chatbot, rassurer les utilisateurs en leur expliquant que le chatbot n'est là que pour faire gagner du temps et ne doit pas se substituer au RH.

4.5. Industries culturelles et créatives

4.5.1. Contexte

Les industries culturelles et créatives (ICC) constituent un ensemble diversifié de secteurs économiques et artistiques, caractérisés par la production et la diffusion de biens et de services culturels et créatifs. Elles représentent une source majeure d'emplois, de revenus et de croissance économique à l'échelle mondiale, tout en contribuant à la diversité culturelle et à l'enrichissement personnel et collectif.

Les ICC englobent une large gamme de domaines créatifs, incluant les arts visuels (peinture, sculpture, photographie, arts graphiques), la musique, le cinéma et l'audiovisuel, la littérature et l'édition, le design et la mode, la publicité et la communication, les jeux vidéo et le multimédia, l'architecture et l'urbanisme, ainsi que le patrimoine culturel.

Ces industries ont un impact significatif sur notre vie quotidienne, en façonnant notre perception du monde et en influençant nos choix et comportements. Elles sont également un moteur important de l'innovation économique, en favorisant la créativité, l'entrepreneuriat et l'investissement dans de nouveaux produits et services.

Cependant, les ICC doivent faire face à de nouveaux défis dans un contexte de transformation numérique et d'émergence de l'intelligence artificielle (IA) générative. Ces technologies offrent des opportunités inédites de croissance et d'innovation, mais soulèvent également des questions cruciales en matière de propriété intellectuelle, de diversité culturelle et de responsabilité éthique.

Pour relever ces défis, il est essentiel de comprendre les risques et les enjeux liés à l'utilisation de l'IA générative dans les ICC, ainsi que d'identifier les stratégies et pratiques à adopter pour minimiser ces risques et maximiser les bénéfices pour tous les acteurs concernés.

4.5.2. Propriété Intellectuelle

Plagiat : Les IA génératives apprennent à partir de vastes ensembles de données, largement tirés de contenus existants sur Internet. Dans les arts visuels, cela peut entraîner la création d'œuvres d'art générées par IA qui ressemblent fortement à celles de peintres connus, ou de morceaux de musique qui imitent des compositions protégées par le droit d'auteur. Cela soulève des questions de plagiat et de vol de propriété intellectuelle.

Violation des droits d'auteur : L'utilisation d'œuvres protégées pour entraîner des modèles IA sans le consentement des auteurs constitue une violation des droits d'auteur. Par exemple, dans le secteur de l'édition, des livres ou articles générés par IA pourraient intégrer des extraits d'œuvres protégées sans autorisation, posant ainsi un problème juridique et éthique. Contrairement à ce qu'on pourrait penser, l'IA peut, dans certains pays, collecter légalement des œuvres protégées pour s'entraîner. En Europe, la directive sur le droit d'auteur de 2019[44], transposée en France, prévoit une exception dite de « *text and data mining* » (fouilles de textes et de données). Cette exception permet aux systèmes d'IA de collecter des données à des fins d'entraînement, sauf si l'auteur s'y est expressément opposé pour un usage commercial.

Exemples notables d'actions en justice : Plusieurs cas illustrent les enjeux juridiques de l'utilisation des œuvres protégées par l'IA.

- **Getty Images vs Stability AI**[45] : Getty Images a poursuivi Stability AI pour utilisation non autorisée d'images protégées dans le cadre de l'entraînement de ses modèles IA. Getty Images affirme que Stability AI a utilisé des millions de ses images sans licence pour entraîner son générateur d'images, Stable Diffusion. Ces données seraient utilisées en dehors du cadre du « fair use » américain qui autorise l'usage des images sous copyright dans un objectif non commercial ou éducatif. Stability AI reconnaît avoir utilisé des images de Getty Images pour entraîner son modèle, impliquant une "copie temporaire". L'entreprise soutient que son modèle ne mémorise pas les images d'entraînement mais ajuste les poids du modèle, créant une base de données originale. Les images générées résultent de bruit aléatoire traité en réponse aux prompts des utilisateurs, non de copies des œuvres protégées. Stability AI argue que toute ressemblance est due aux prompts des utilisateurs et non à une mémorisation des images d'entraînement. Ils invoquent[46] également l'exception britannique pour caricature, parodie ou pastiche pour justifier l'utilisation des images

[44] Directive (UE) 2019/790 du Parlement européen et du Conseil du 17 avril 2019 sur le droit d'auteur et les droits voisins dans le marché unique numérique et modifiant les directives 96/9/CE et 2001/29/CE. Journal officiel de l'Union européenne. 17 mai 2019. https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L0790

[45] Blake Brittain. Getty Images lawsuit says Stability AI misused photos to train AI. Reuters. February 6, 2023. https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

[46] Maura O'Malley. High Court allows Getty Images' IP dispute with Stability AI to go to trial. The Global Legal Post. December 06, 2023. https://www.globallegalpost.com/news/high-court-allows-getty-images-ip-dispute-with-stability-ai-to-go-to-trial-1715547241

de Getty dans leurs sorties synthétiques : les images générées ne sont qu'imitation et ne peuvent reproduire une image particulière à partir d'un prompt spécifique.

- **New York Times vs OpenAI**[47] : Le New York Times a intenté une action en justice contre OpenAI pour utilisation non autorisée de ses articles. OpenAI est accusé d'avoir utilisé des articles du New York Times pour entraîner son modèle de langage sans autorisation, intégrant ainsi des contenus protégés dans ses réponses générées. OpenAI a répondu que l'entraînement des modèles d'IA à partir de matériaux disponibles publiquement sur Internet constitue un usage équitable ("fair use"), soutenu par des précédents de longue date et largement acceptés. OpenAI considère ce principe comme juste pour les créateurs, nécessaire pour les innovateurs et crucial pour la compétitivité des États-Unis. L'entreprise a également mis en place un processus simple de retrait pour les éditeurs, adopté par le New York Times en août 2023, pour empêcher ses outils d'accéder à leurs sites. Fait intéressant, les régurgitations induites par le New York Times semblent provenir d'articles vieux de plusieurs années qui ont proliféré sur de nombreux sites tiers.

La question épineuse du vide juridique : Le cadre juridique actuel présente des lacunes en matière de régulation de l'IA générative. Des analyses et discussions sont en cours pour combler ce vide et établir des règles claires sur l'utilisation des œuvres protégées dans le développement des IA génératives. La Wikimedia Foundation[48] a exploré ces enjeux dans une analyse de la conformité de ChatGPT aux lois sur le droit d'auteur. De plus, la CNIL française a publié un dossier sur les régulations nécessaires pour encadrer la conception des IA génératives[49]. Aux États-Unis, un nouveau projet de loi intitulé *Generative AI Copyright Disclosure Act*[50] propose d'exiger des entreprises d'IA qu'elles divulguent les matériaux protégés par le droit d'auteur utilisés pour construire leurs modèles d'IA générative. Ces discussions soulignent la nécessité d'un équilibre entre innovation technologique et protection des droits des créateurs.

---

[47] Michael M. Grynbaum, Ryan Mac. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. The New York Times. December 27, 2023. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

[48] Wikimedia Foundation. Wikilegal/Copyright Analysis of ChatGPT. March 23, 2023. https://meta.wikimedia.org/wiki/Wikilegal/Copyright_Analysis_of_ChatGPT

[49] CNIL. IA : la CNIL publie ses premières recommandations sur le développement des systèmes d'intelligence artificielle. 08 avril 2024. https://www.cnil.fr/fr/ia-la-cnil-publie-ses-premieres-recommandations-sur-le-developpement-des-systemes-dintelligence

[50] US Congress. H.R.7913 - Generative AI Copyright Disclosure Act of 2024. April 9, 2024. https://www.congress.gov/118/bills/hr7913/BILLS-118hr7913ih.pdf

4.5.3. Qualité et Originalité

Homogénéisation du contenu : Les IA génératives sont capables de produire du contenu en quantité, mais cela peut conduire à une uniformisation et à un manque de diversité créative. L'accent mis sur la performance individuelle par les IA génératives peut potentiellement réduire la variété et l'originalité des créations. Des recherches indiquent que bien que ces technologies puissent stimuler la créativité individuelle, elles ont tendance à restreindre la diversité des nouvelles idées et des expressions artistiques, en favorisant plutôt des productions prévisibles et similaires.

Perte de créativité humaine : L'utilisation croissante des IA pour la création pourrait diminuer l'incitation à développer des compétences créatives humaines profondes. Cela pourrait engendrer une dépendance qui, à long terme, pourrait nuire à la capacité des individus à innover de manière indépendante et à explorer de nouvelles avenues artistiques. Par analogie, l'utilisation massive du GPS démocratisé par les smartphones a diminué notre capacité à lire et suivre une carte routière. Cette analogie souligne comment la commodité technologique peut affaiblir les compétences traditionnelles, et c'est ainsi que l'IA générative pourrait potentiellement réduire la volonté et la capacité de développer des compétences créatives originales.

4.5.4. Éthique et Biais

Reproduction des biais : Les IA génératives, comme toutes les IA entraînées sur des données réelles, sont susceptibles de reproduire et d'amplifier les biais présents dans les données sur lesquelles elles sont entraînées. Par exemple, les modèles linguistiques basés sur l'IA ont été critiqués pour véhiculer des préjugés sexistes, racistes et politiques. Une étude de MIT[51] a mis en évidence comment ces modèles peuvent involontairement renforcer les stéréotypes existants, influençant ainsi négativement les perceptions sociales et politiques.

Contenu inapproprié : Les IA génératives posent également un risque de produire du contenu offensant ou inapproprié. Dans l'industrie des jeux vidéo, par exemple, des personnages ou des scénarios générés par IA peuvent inclure des représentations culturellement insensibles ou des situations qui peuvent être perçues comme

[51] Melissa Heikkilä. These new tools let you see for yourself how biased AI image models are. MIT Technology Review. March 22, 2023. https://www.technologyreview.com/2023/03/22/1070167/these-news-tool-let-you-see-for-yourself-how-biased-ai-image-models-are/

offensantes par différents groupes de joueurs. Cette possibilité soulève des préoccupations éthiques importantes concernant l'impact social et culturel des technologies d'IA. Une étude de l'UNESCO[52] a mis en lumière des cas où des IA génératives ont perpétué des stéréotypes de genre régressifs, mettant en danger les progrès vers l'égalité des genres et la représentation équitable dans les médias et la culture.

4.5.5. Impact Économique et Social

Dévalorisation du travail humain : L'automatisation croissante de la création de contenu par les IA génératives menace de dévaloriser le travail des créateurs humains dans divers domaines artistiques et culturels. Les études indiquent que cette automatisation pourrait conduire à une réduction de la demande pour les compétences créatives traditionnelles, affectant potentiellement les revenus et les conditions de travail des écrivains, artistes, musiciens et autres professionnels du secteur.

Concentration de pouvoir : Les grandes entreprises technologiques qui dominent le marché de l'IA générative pourraient exercer un contrôle considérable sur l'industrie créative. En développant et en possédant des technologies avancées d'IA générative, ces entreprises pourraient influencer la direction artistique, la diffusion culturelle et même les choix de consommation du public. Cette concentration de pouvoir soulève des préoccupations quant à la diversité culturelle et à la concurrence sur le marché. Des appels sont lancés pour promouvoir l'open source des modèles et des données afin de prévenir une monopolisation excessive par les grandes entreprises technologiques et de favoriser un écosystème créatif plus ouvert et équitable. La startup française Hugging Face, notamment, est un leader dans le domaine de l'open source des modèles d'IA, plaidant pour une plus grande transparence et une meilleure accessibilité des technologies génératives.

4.5.6. Impact sur le modèle économique

L'IA générative défie les industries de contenus en deux sens : d'une part, elle s'entraîne sur de vastes ensembles de données disponibles en ligne, parfois derrière des paywalls (*input*) ; d'autre part, elle restitue des fragments de ces contenus dans ses créations sous forme de texte, d'image ou de son (*output*). Les nouvelles interfaces conversationnelles

[52] UNESCO. Systematic Prejudices. An Investigation into Bias Against Women and Girls in Large Language Models. March 7, 2024. https://unesdoc.unesco.org/ark:/48223/pf0000388971

comme Perplexity AI[53] mettent à mal un modèle économique basé sur des liens vers les sites internet propriétaires des marques. Tandis que les uns bloquent les crawlers d'Open AI, Gemini et al., d'autres médias ont fait le choix d'un pacte faustien[54] avec les géants de la tech leur ouvrant l'accès à leurs contenus à la fois pour l'entraînement et pour la restitution de résultats de recherche (AP, Axel Springer, FT, News Corp, Le Monde en France).

Ce choix fragilise les médias à moindre impact, sans opportunité de partenariat avec les plateformes, qui risquent de perdre toute probabilité de proéminence dans les réponses générées par les IA génératives, limitant ainsi la diversité des points de vue accessibles aux utilisateurs. A long terme, il constitue aussi un pari hasardeux, complexifiant les relations avec les plateformes dans un contexte incertain de la révolution des usages. Google a déjà diminué le nombre de résultats de recherche générés par l'IA (AI Overview)[55] pour privilégier son modèle économique classique de la publicité.

4.5.7. Manipulation d'information

***Deepfakes* et désinformation :** Les IA génératives sont capables de créer des *deepfakes*, des vidéos ou images truquées de manière réaliste, pouvant être utilisées à des fins de désinformation, de fraude ou d'atteinte à la réputation. Par exemple, dans le domaine de l'audiovisuel, des vidéos manipulées de personnalités publiques peuvent être diffusées pour influencer l'opinion publique ou nuire à des individus, posant ainsi des risques significatifs pour la sécurité individuelle et collective.

Fausses informations : Les outils d'IA générative ont la capacité de générer du contenu inexact et trompeur. Des études ont révélé que des systèmes comme ChatGPT et Bard peuvent produire des données fabriquées qui semblent authentiques, souvent qualifiées d'« hallucinations ». Cette problématique soulève des préoccupations importantes en matière de protection des consommateurs, notamment vis-à-vis des lois sur la protection des consommateurs qui s'appliquent aux outils d'IA générative. Par exemple, la Federal Trade Commission (FTC) aux États-Unis a clarifié que l'interdiction de

[53] https://www.perplexity.ai/

[54] Cecily Mauran. All the media companies that have licensing deals with OpenAI (so far). Mashable. June 21, 2024. https://mashable.com/article/all-the-media-companies-that-have-licensing-deals-with-openai-so-far

[55] Robert Hart. Google Restricts AI Search Tool After 'Nonsensical' Answers Told People To Eat Rocks And Put Glue On Pizza. Forbes. May 31, 2024. https://www.forbes.com/sites/roberthart/2024/05/31/google-restricts-ai-search-tool-after-nonsensical-answers-told-people-to-eat-rocks-and-put-glue-on-pizza/

pratiques trompeuses ou déloyales de la FTC s'applique également aux outils d'IA générative, susceptibles d'être exploités pour des escroqueries par phishing, des vols d'identité, des *deepfakes* ou la création d'autres contenus trompeurs, frauduleux ou préjudiciables

4.5.8. Risques pour les traducteurs professionnels

Depuis l'adoption des réseaux neuronaux en 2015, les algorithmes de traduction se sont grandement améliorés. Mais l'IA générative, en particulier les technologies avancées comme GPT-4, bouleverse aujourd'hui le secteur de la traduction, menaçant directement les traducteurs professionnels. Une enquête de la Society of Authors (SoA)[56] a révélé que plus d'un tiers des traducteurs ont déjà perdu des opportunités de travail à cause de l'IA générative, tandis que plus des trois quarts estiment que cette technologie aura un impact négatif sur leurs revenus futurs. En exploitant les récents modèles de traitement du langage naturel, les entreprises peuvent désormais traduire rapidement et à moindre coût de vastes volumes de texte. Cette évolution pousse les traducteurs humains à utiliser des outils d'IA pour rester compétitifs, une intégration qui n'est pas sans risques. Les traductions automatiques, souvent littérales, manquent des nuances culturelles et contextuelles essentielles que seuls les traducteurs humains peuvent apporter. Les expressions idiomatiques et les subtilités linguistiques sont perdues. Une tendance particulièrement préoccupante dans des domaines critiques comme le droit, la médecine et la diplomatie, où une compréhension fine et précise est cruciale pour éviter les erreurs et les malentendus. Des traductions illisibles de livres piratés qui inondent Internet montrent que le processus ne peut actuellement pas être entièrement sous-traité à des ordinateurs.

Mais certains éditeurs, pourraient faire le choix de la rapidité face à la qualité, laissant ainsi aux traducteurs professionnels le simple rôle de relecteur. Dans son enquête récente, le Conseil européen des associations de traducteurs littéraires[57] recommande que les professionnels évitent de réviser des textes générés par l'IA ou facturent des tarifs de traduction pour ce travail. Cette situation pourrait également élever les critères

[56] SOA Policy team. SoA survey reveals a third of translators and quarter of illustrators losing work to AI. The Society of authors. April 11, 2024. https://www2.societyofauthors.org/2024/04/11/soa-survey-reveals-a-third-of-translators-and-quarter-of-illustrators-losing-work-to-ai/

[57] European Translators and AI: Survey for Member Associations. European Council of Literary Translators' Associations. February 20, 2024. https://www.ceatl.eu/european-translators-and-ai-survey

d'entrée dans l'industrie, réservant la traduction littéraire à ceux ayant les moyens financiers de supporter les périodes sans revenu stable.

La mise à disposition de traducteurs automatiques peut aussi avoir un impact à long terme sur nos capacités cognitives. L'apprentissage des langues a des effets bénéfiques prouvés sur le cerveau[58]. Des études montrent que l'apprentissage et l'utilisation active de plusieurs langues améliorent la neuroplasticité et renforcent les réseaux cérébraux impliqués dans la cognition et la mémoire. En négligeant cet aspect, nous risquons de compromettre notre développement cognitif et émotionnel. Les langues sont plus que de simples outils de communication verbale et de santé cérébrale ; elles façonnent également notre vision du monde et notre comportement culturel.

4.5.9. Risques pour les graphistes et les designers

L'évolution rapide de l'intelligence artificielle générative dans le domaine du design graphique pose des défis significatifs aux métiers de graphistes et de designers, tout en soulevant des inquiétudes quant à l'avenir de ces professions. Les technologies de l'IA, de plus en plus intégrées dans le processus de conception, automatisent des tâches autrefois réservées à la créativité humaine, offrant ainsi aux designers l'opportunité de repousser les limites de leur potentiel créatif.

La question de la propriété et de la licence des créations générées par l'IA devient cruciale. Dans certains cas, les droits d'auteur et autres droits de propriété intellectuelle liés à une création générée par l'IA peuvent appartenir au développeur du système IA plutôt qu'au designer ou à son client. Il est donc essentiel pour les designers et les graphistes de bien s'assurer des conditions d'utilisations des outils IA qu'ils utilisent afin de s'éviter tout problème sur le plan légal.

Bien que les algorithmes d'IA puissent générer des designs selon des critères prédéfinis, ils peinent à saisir les nuances complexes présentes dans la conception, telles que les références culturelles, les symbolismes contextuels et les sensibilités sociales, éléments vitaux pour une communication visuelle efficace. Les améliorations technologiques dans le design graphique, tout en favorisant l'innovation, soulèvent des préoccupations concernant la perte de la créativité humaine et de la touche artistique unique. Ces

[58] Carley Spence. How learning a new language changes your brain. Cambridge. April 29, 2022. https://www.cambridge.org/elt/blog/2022/04/29/learning-language-changes-your-brain/

inquiétudes mettent en lumière l'importance de préserver la présence des designers humains et leur individualité dans le processus de création.

Par ailleurs, l'utilisation de l'IA peut mener à une homogénéisation excessive des designs, menaçant la diversité et l'innovation dans le domaine du design graphique. Les défis technologiques posés par l'IA représentent également une menace pour les emplois des designers, alimentant les craintes de remplacement des tâches de design graphique par des techniques automatisées. Cela souligne la nécessité de développer continuellement les compétences des designers pour répondre aux exigences futures de leur métier.

En combinant leurs compétences traditionnelles avec des outils de pointe basés sur l'IA, les designers peuvent créer des œuvres innovantes et percutantes. Cependant, une réticence à adopter pleinement ces technologies peut limiter la créativité dans un secteur en constante évolution. Il est crucial pour les designers de tirer parti de ces avancées technologiques afin de maintenir l'élément humain indispensable à la conception.

4.5.10. Solutions et approches pour mitiger les risques

De plus en plus de stratégies et propositions permettent de gérer et réduire les risques associés à l'utilisation croissante des IA génératives dans les industries créatives et culturelles. Elles visent à maximiser les bénéfices tout en minimisant les impacts négatifs sur les individus et la société.

- **Développement éthique de l'IA :** Il est crucial d'adopter des pratiques de développement et d'entraînement des modèles IA de manière responsable pour prévenir les biais et les contenus inappropriés. Des approches telles que l'optimisation du prétraitement pour la prévention de la discrimination et l'intégration de l'éthique dans l'éducation en informatique sont essentielles.
- **Politiques de régulation :** La mise en place de régulations est cruciale pour protéger les droits d'auteur et encourager une utilisation éthique des IA dans tous les secteurs des industries culturelles et créatives. Des propositions telles que l'AI Act[59] européen visent à établir un cadre juridique adapté à l'évolution technologique.

[59] https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ:L_202401689

- **Évaluation critique des résultats des IAG :** Adopter un cadre de gestion des risques d'IA est essentiel pour évaluer de manière critique les impacts potentiels des technologies génératives et mettre en œuvre des mesures correctives appropriées.
- **Modération de contenu et audits :** Il est nécessaire de mettre en place des processus de modération de contenu rigoureux et des audits réguliers pour identifier et corriger tout contenu inapproprié ou problématique généré par les IA.
- **Éducation et formation :** Sensibiliser les créateurs et les professionnels des ICC aux risques associés aux IA génératives et promouvoir les bonnes pratiques d'utilisation sont des stratégies clés. Cela inclut l'orientation des utilisations des IAG vers des cas d'usage collaboratifs et exploratoires, ainsi que la diversification des sources de données utilisées pour l'entraînement des modèles.
- **Mise en place de barrières techniques :** Utiliser des balises techniques pour empêcher l'accès non autorisé aux contenus sous droits d'auteur est une mesure préventive importante. Des approches comme le marquage numérique permettent également de détecter et de traiter efficacement le contenu problématique généré par les IA.

4.5.11. Conclusion

Le rapport de l'homme à la machine dans le milieu culturel n'a pas fondamentalement changé, et les craintes liées aux nouvelles technologies demeurent constantes. Tout comme les technologies précédentes, l'IAG (IA Générative) doit être perçue comme un outil puissant ayant permis des avancées majeures, notamment dans les industries culturelles et créatives. Cependant, l'accouplement de l'IAG avec des technologies avancées impose aujourd'hui une double révolution.

Révolution Technique et Sociétale

La première révolution est technique : l'IAG permet de produire de manière innovante, tout comme l'invention de l'acier et du béton a révolutionné les industries. Elle ouvre des possibilités créatives et productives sans précédent.

La seconde révolution est sociétale : l'IAG redéfinit le rapport de l'homme au travail, à l'image des robots dans les usines. Cette transformation peut bouleverser les métiers créatifs, modifiant non seulement les méthodes de travail mais aussi la valorisation du travail humain.

Exemples Historiques

L'exemple d'Internet et des moteurs de recherche illustre bien cette dynamique. Ces technologies ont accéléré l'accès à la connaissance sans pour autant remplacer l'homme. Elles ont transformé les métiers, mais avec le bon encadrement et le bon accompagnement, elles se sont révélées être des atouts et des outils essentiels.

L'Importance de l'Éducation et de la Formation

Le risque est de considérer l'IAG comme un ennemi et de succomber à la peur qu'elle suscite. C'est pourquoi l'éducation et la formation jouent un rôle crucial pour accompagner la transition des métiers. En adoptant une approche proactive, en mettant en place des lois et des encadrements sociaux adaptés, et en formant les professionnels, l'IAG peut devenir un véritable allié dans l'évolution des industries culturelles et créatives.

4.5.12. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	**Plagiat** : génération de contenu ressemblant fortement à des données protégées par le droit d'auteur. **Violation des droits d'auteur** : L'utilisation d'œuvres protégées pour entraîner des modèles IA sans le consentement des auteurs constitue une violation des droits d'auteur.	2	4	3	1	1	2	1	Développement éthique de l'IA. Mise en place de politique de régulation. Mise en place de barrières techniques.
Données	**Homogénéisation du contenu** : Les IA génératives sont capables de produire du contenu en quantité, mais cela peut conduire à une uniformisation et à un manque de diversité créative.	1	1	2	3	2	3	1	Évaluation critique des résultats des IAG.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Humain	**Perte de créativité humaine** : L'utilisation croissante des IA pour la création pourrait diminuer l'incitation à développer des compétences créatives humaines profondes. Cela pourrait engendrer une dépendance qui, à long terme, pourrait nuire à la capacité des individus à innover de manière indépendante et à explorer de nouvelles avenues artistiques.	2	1	2	3	1	3	1	Éducation et formation.
Données / Modèle	**Reproduction des biais** : Les IA génératives sont susceptibles de reproduire et d'amplifier les biais présents dans les données sur lesquelles elles sont entraînées. **Contenu inapproprié** : Les IA génératives posent également un risque de produire du contenu offensant ou inapproprié.	1	4	4	2	1	4	1	Évaluation critique des résultats des IAG. Modération de contenu et audits.
Humain	**Dévalorisation du travail humain** : L'automatisation croissante de la création de contenu par les IA génératives menace de dévaloriser le travail des créateurs humains dans divers domaines artistiques et culturels. **Concentration de pouvoir** : Les grandes entreprises technologiques qui dominent le marché de l'IA générative pourraient exercer un contrôle considérable sur l'industrie créative.	3	2	2	3	3	4	1	Développement éthique de l'IA. Mise en place de politique de régulation. Éducation et formation.
Humain	***Deepfakes* et désinformation** : Les IA génératives sont capables de créer des *deepfakes*, des vidéos ou images truquées de manière réaliste, pouvant être utilisées à des fins de désinformation, de fraude ou d'atteinte à la réputation. **Fausses informations** : Les outils d'IA générative ont la capacité de générer du contenu inexact et trompeur, susceptibles d'être exploité pour des escroqueries par phishing, des vols d'identité.	3	4	4	2	2	4	2	Évaluation critique des résultats des IAG. Modération de contenu et audits. Développement éthique de l'IA. Mise en place de politique de régulation

4.6. Santé

4.6.1. Introduction

Le secteur de la santé cristallise les opportunités, mais aussi les peurs et les fantasmes liés à l'émergence de l'intelligence artificielle. Depuis plusieurs années, certains experts promettent que les machines seront très vite "meilleures" que les radiologues pour détecter des tumeurs sur des images, d'autres nous disent que grâce à la puissance de calcul les algorithmes vont être capables de "découvrir" de nouveaux médicaments en quelques semaines contre plusieurs années aujourd'hui.

L'intelligence artificielle générative tient une place importante dans les dispositifs mis à disposition des patients, le grand public étant aujourd'hui connaisseur (et utilisateur plus ou moins averti) d'outils comme ChatGPT. Quels sont les risques liés à l'utilisation de ces grands modèles de langage (LLM) en santé ? Comment établir la confiance avec un système de ce type ? Autant de questions que nous avons souhaité nous poser en nous appuyant sur un cas d'usage concret : la mise à disposition d'un chatbot de suivi post-opératoire pour prendre des nouvelles des patients opérés en ambulatoire, par exemple pour une extraction de dents de sagesse, et délivrer un premier niveau de conseil afin de rassurer en journée, le soir, la nuit et le weekend.

4.6.2. Cas d'usage

Assistant virtuel d'accompagnement post-opératoire

Le cas d'usage étudié s'intéresse à l'utilisation de modèles d'intelligence artificielle générative pour faciliter le suivi de patients après une opération en ambulatoire. En effet, une fois rentrés à leur domicile, seulement 57% des Français[60] ont un contact avec une équipe médicale après leurs sorties.

Il manque des professionnels (secrétaires, infirmières) pour effectuer ce suivi au téléphone, les outils d'IA générative seraient-ils la solution pour venir pallier ce manque de personnel ? Nous considérons ici un assistant capable de remonter des alertes aux praticiens pour leur laisser la main sur la gestion des cas complexes : le dispositif ne doit

[60] Marie Gloanec, Caroline Prunet. Haute Autorité de Santé. Mesure de la satisfaction et de l'expérience des patients, résultats nationaux 2021 concernant les patients hospitalisés pour une chirurgie ambulatoire. Haute Autorité de Santé. 16 décembre 2021. https://www.has-sante.fr/upload/docs/application/pdf/2021-12/iqss_2021_rapport_resultats_nationaux_esatismcoca_2021.pdf

pas se substituer à la responsabilité du chirurgien, mais l'assister et lui faire gagner du temps.

Des risques liés aux données

Le conseil médical devant être dispensé par un **professionnel** habilité, expert, et sanctionné par un diplôme dans la spécialité en question, peut-on faire confiance aux conseils délivrés par un grand modèle de langage pré-entraîné **sur un jeu de données que l'on ne connaît pas** ? Cela présente des risques évidents de génération de **conseils inappropriés et faux** qui, sans le cadre d'un sujet de santé, peut engager la vie du patient.

La génération de conseils médicaux doit être **encadrée et évaluée**. Ces conseils doivent être validées par des spécialistes et être suivies à la lettre par l'IA générative qui sert de socle technique au système. Attention aux **hallucinations** et autres inventions de contenus pouvant entraîner des complications médicales, et **engager la vie du patient** : ceci n'est pas un exercice.

Et des risques inhérents aux modèles d'IA générative et à leur utilisation

La question de la **gestion des données personnelles et des données de santé à caractère personnel** est centrale. Il semble être admis que plus il y a de données de qualité représentatives de la population cible d'utilisateurs, plus le modèle serait performant. Dans notre cas d'usage, il semble pourtant important de minimiser les données nécessaires pour rendre le service attendu (un premier niveau de rappel de consignes et de conseils, pas une prise en charge ultra personnalisée).

Pour **éviter les hallucinations** et les conseils dangereux pour l'intégrité du patient, l'évaluation est incontournable et indispensable pendant la conception de l'assistant mais aussi, et surtout, en production.

Le patient doit pouvoir faire confiance à ce que lui explique cet assistant mais il doit pouvoir vérifier l'information et demander des précisions à son médecin si nécessaire. Attention aux dérives de l'IA générative omnisciente, assistant médical capable de répondre à tout et n'importe quoi.

Des impacts importants sur l'expérience patient et l'organisation des soins

Ce type d'outils intégrant de l'intelligence artificielle générative doit respecter la législation en vigueur pour les dispositifs médicaux si celui-ci en est un, mais aussi le RGPD bien sûr, et le RIA pour des usages à hauts risques. Les risques financiers sont liés aux risques légaux en cas de non-respect de ce cadre réglementaire.

Une fois adopté, un outil de suivi patient impacte considérablement **l'organisation des soins**, avec une redéfinition des tâches au sein de l'équipe médicale. Si le système n'est pas fiable ou pas suffisamment bien évalué, il peut remettre en cause cet équilibre entre le gain de temps supposé et les difficultés rencontrées au quotidien simplement pour s'adapter aux erreurs de l'outil. Pour s'assurer du respect de ce principe de la garantie humaine dans l'évaluation des systèmes d'IA à destination des patients et des professionnels de santé, l'AFNOR propose en mai 2024 un document technique[61] sur le sujet. Enfin, la mise en place de cet outil de suivi nouvelle génération vient avant tout réparer un trou dans la raquette lié à un manque de moyens et de personnels. Un système à base d'IA générative qui serait proposé aux patients comme un assistant "aussi bon" ou "meilleur" qu'un médecin serait très dangereux pour la perception du public. Il faut que cet impact **sociétal**, loin d'être négligeable, soit bien pris en compte dès la conception de ces services e-santé.

[61] AFNOR. Garantie humaine des systèmes fondés sur l'intelligence artificielle en santé. Mai 2024. https://www.boutique.afnor.org/fr-fr/norme/afnor-spec-2213/garantie-humaine-des-systemes-fondes-sur-lintelligence-artificielle-en-sant/fa205274/419909

4.6.3. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	Données d'apprentissage non scientifiquement validées	1	4	1	3	4	4	1	Co-conception des outils avec des professionnels de santé.
Modèle	Utilisation d'un modèle sur étagère non spécialisé	1	4	1	3	4	3	3	Bien comprendre les limites pour s'assurer que le modèle soit conforme au cas d'usage.
Modèle	Manque d'évaluation du modèle	3	3	3	4	4	4	1	Organiser l'évaluation tout au long du projet, avant et après mise en production.
Humain	Utilisation "comme avec un professionnel de santé"	2	3	1	1	4	4	1	Accompagner la mise en place de l'outil avec une campagne d'information et de sensibilisation pour les équipes et pour les patients.

4.7. Commerce

4.7.1. Introduction

La fonction commerciale est le moteur de la performance économique des entreprises, elle est évidemment le premier générateur de chiffre d'affaires. Du fait de sa proximité avec les prospects et les clients, elle diffuse l'image de marque de l'entreprise et joue un rôle dans la collecte des besoins des clients et l'analyse des tendances du marché, permettant ainsi une meilleure adaptation des offres et une anticipation des évolutions sectorielles.

Avec l'émergence de l'intelligence artificielle, les pratiques commerciales connaissent une véritable transformation. L'IA générative offre des possibilités prometteuses pour optimiser et automatiser divers aspects de la fonction commerciale. On peut citer de façon non exhaustive :

- L'automatisation de la prospection : générer des messages personnalisés pour les emails et automatiser les appels téléphoniques grâce à des voix synthétiques ;
- L'automatisation de la qualification de prospects ;
- L'automatisation du reporting et des tâches administratives, comme la prise de notes ou l'enrichissement du CRM ;
- L'aide à la génération de devis, de propositions commerciales ou même la personnalisation des outils de présentation ou démonstration ;
- La fidélisation ou la relation client.

L'enjeu, comme pour beaucoup de fonctions, est de pouvoir ainsi maximiser le temps passé sur des fonctions humaines à forte valeur ajoutée (dialogue, argumentation, analyse du besoin, etc.).

Cette nouvelle approche n'est pas sans risques : les **biais algorithmiques** constituent une source majeure de préoccupation. Les modèles d'IA, entraînés sur des données historiques, peuvent reproduire des préjugés existants, influençant les stratégies de ciblage et de segmentation de manière injuste. Cela peut entraîner des discriminations et l'exclusion de certains segments de marché, nuisant à l'équité des opportunités commerciales.

La **confidentialité des données clients** est un autre enjeu critique. L'utilisation de l'IA générative implique l'accès à des informations sensibles et personnelles, posant des questions éthiques et légales sur la protection des données. Une gestion inadéquate

peut entraîner des violations de la vie privée et des pertes de confiance de la part des clients.

Enfin, une dépendance excessive à l'IA générative pourrait **réduire les interactions humaines authentiques**, essentielles pour établir des relations de confiance avec les clients. L'automatisation poussée peut également **impacter la créativité et l'innovation** des équipes commerciales, qui risquent de trop se reposer sur des contenus générés automatiquement plutôt que de proposer des idées et arguments originaux, limitant de fait les capacités de différentiation de l'entreprise.

Pour atténuer ces risques, plusieurs pistes de remédiation existent. On peut citer :

- Des **mécanismes de surveillance et de contrôle** rigoureux doivent être mis en place, assurant l'évaluation régulière des modèles d'IA pour détecter et corriger les biais, le renforcement de mesures de protection des données mais encore la transparence et la traçabilité quant à l'utilisation du système ;
- Des **actions de formation et de sensibilisation** doivent être dispensées aux professionnels commerciaux et aux managers pour assurer une utilisation éthique et équitable de cette technologie ;
- Enfin, une **analyse rigoureuse du process de vente** doit être menée de bout en bout afin d'identifier les actions à automatiser ou dans lesquelles il est utile et opportun d'intégrer de l'IA générative.

4.7.2. Cas d'usage

Automatisation de la prospection dans une entreprise de services digitaux

Entreprise

ARL, spécialisée dans les services 3D et digitaux.

Objectif

Automatiser la prospection et la création de contenus tout en réduisant la dépendance à la prospection humaine, afin d'améliorer l'efficacité et de permettre aux équipes de se concentrer sur des tâches à plus forte valeur ajoutée.

Contexte

ARL a mis en place des solutions d'IA générative pour automatiser plusieurs aspects de la prospection. Ces solutions incluent la génération d'emails, de posts LinkedIn, de vidéos

de présentation de projet, d'avatars, de textes SMS, de voix et de séquences d'emails, etc. Le but est d'optimiser les interactions avec les prospects et de libérer du temps pour les équipes commerciales.

Le système a été intégré au processus existant de l'entreprise et une formation a été dispensée aux équipes pour les familiariser avec les nouvelles fonctionnalités.

Un système de *feedback* a également été mis en place pour recueillir les avis des utilisateurs suite à l'expression de frustration dans les échanges clients dans les phases de prospection et de relation client. Ce système de feedback a permis des ajustements continus pour optimiser l'utilité et l'efficacité des outils d'IA.

Description du cas

L'introduction de systèmes d'IA générative au sein de l'entreprise ARL a marqué une étape intéressante dans l'automatisation de la prospection et de la création de contenus. Cependant, plusieurs limites ont été constatées lors de la mise en œuvre. Lors d'une campagne de prospection, l'IA a généré des emails génériques à des prospects de différentes industries sans tenir compte de leurs besoins spécifiques. Un prospect clé a reçu un message non pertinent, ce qui a diminué son intérêt pour les services de l'entreprise. En réponse, un représentant commercial a pris contact personnellement avec ce prospect, a identifié ses besoins spécifiques et a proposé une solution sur mesure et tenant compte de son contexte spécifique.

Cette approche humaine a non seulement redressé la perception initiale de l'entreprise par le prospect mais a également abouti à la conclusion d'un nouveau projet. En réponse à cet incident, l'entreprise a réintégré la prospection humaine dans son processus. La segmentation des cibles proposée par l'IA est désormais validée par des commerciaux avant l'envoi de la campagne de message.

De plus, des points de contact humains réguliers ont été réintroduits pour maintenir une relation personnelle avec les prospects. Ces mesures ont permis de restaurer la satisfaction des prospects et d'améliorer les taux de conversion. En effet, autant il a été constaté une augmentation de la productivité de 30% à 40 % dans la production de contenus, autant le résultat fut plus décevant sur le volume de rendez-vous et de conversion estimé de 10 à 20%. Cette expérience souligne l'importance de trouver un équilibre entre l'automatisation des interactions via un outil d'IA générative et les interactions humaines que les clients recherchent.

Il est également important de bien identifier les étapes de la prospection et de la relation client dans lesquelles on souhaite intégrer cette technologie. Enfin, nous avons noté que le degré d'automatisation à adopter dans le processus de vente via l'IA générative doit varier selon le type de produit ou service, de la typologie de client ainsi que la complexité de la vente qu'il présente. Plus le produit ou service est considéré comme ayant une forte valeur perçue, faisant partie d'un processus de vente complexe plus la dimension humaine est prégnante.

Le cas d'usage complet est consultable ici :
https://www.emocio.hr/hubfrance_iag_risque_rh

4.7.3. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	Manque de personnalisation des données d'entraînement de l'IA entraînant la diminution de l'engagement des prospects et nuisant à l'image de marque et à la fidélisation.	1	1	3	2	3	1	1	• Mettre en place une validation humaine des contenus générés avant envoi ; • Utiliser des données d'entraînement plus diversifiées et spécifiques à chaque industrie ou population ciblée.
Modèle	Accès et utilisation inappropriée des données clients par l'IA (par exemple donnée personnelle non pertinente pour l'acte de vente) : perte de confiance des clients, sanctions légales pour non-conformité aux régulations de protection des données.	2	4	2	2	3	1	1	• Renforcer les protocoles de sécurité des données ; • Assurer la conformité avec la réglementation RGPD ; • Cloisonner les données non pertinentes pour l'algorithme des données utilisables.
Modèle	Erreurs dans la génération automatique de contenu, réponses inappropriées ou mal ciblées pouvant offenser les prospects ou diffuser des informations incorrectes.	2	4	4	2	3	1	1	• Mettre en place un système de feedback rapide pour détecter et corriger les erreurs en temps réel ; • Former les commerciaux à intervenir et rectifier les réponses rapidement ; • Assurer la transparence en rappelant aux prospects que la conversation est menée par une IA générative ; • Proposer une action rapide pour passer à une conversation par un humain.
Humain	Dépendance excessive à l'IA générative : limitation des interactions humaines authentiques générant un risque de dégradation de la qualité de la relation client	3	1	2	2	1	3	1	Contrôler les taux de conversion et monitorer les interactions IAG / prospects.

4.8. Développement logiciel

4.8.1. Introduction

Le domaine du développement logiciel, paradoxalement, ne semble pas être précurseur en matière d'intégration de l'IA générative dans ses processus de production. Le métier du développement logiciel, depuis ses débuts, est dans une évolution permanente, tant sur l'apparition puis l'essoufflement des langages, que sur les outils et les *frameworks*. L'IA générative ne fait pas exception et s'immisce progressivement dans les pratiques. Actuellement, son utilisation directe se limite bien souvent à une demande de génération d'un extrait de code (*snippet*) qui vient débloquer un développeur ; l'IA générative est aussi bien présente dans les IDE (*Integrated Development Environment*) afin d'accompagner le développeur dans ses tâches en cours, en lui proposant des suggestions contextuelles et des extraits de code prêts à l'emploi.

Peut-on attendre plus d'une telle technologie, qui plus est, « engendrée » par ce même domaine d'expertise ?

4.8.2. Cas d'usage

Outils de développement au sein d'une Digital Factory

Une grande société s'est organisée pour confier la totalité de ses développements logiciels spécifiques à sa *Digital Factory*. Cette dernière a la charge de développer et de maintenir les applications attendues par les services métiers pour répondre à leurs besoins. L'objectif de cette organisation est de pouvoir mutualiser les pratiques, en uniformisant la pile technologique employée, ainsi que la méthodologie de travail au sein des équipes.

Certaines applications sont initiées sous la forme de démonstrateurs, avec une méthodologie *Extreme Programming*[62] afin d'itérer fréquemment avec les représentants métiers. Cette pratique encourage la production très rapide de l'application tout en respectant au mieux l'état de l'art afin d'éviter l'apparition prématurée d'une dette technique qui ne pourrait pas être comblée ultérieurement.

[62] Méthode agile de génie logiciel inventée chez Chrysler :
Kent Beck, Extreme Programming Explained : Embrace Change, Addison Wesley, 1999. https://books.google.fr/books?id=G8EL4H4vf7UC&lpg=PP1&hl=fr&pg=PP1#v=onepage&q&f=false

Les développeurs impliqués dans ce type de projet vont s'appuyer sur une IA générative qui saura leur fournir des « briques » de code à partir du besoin exprimé dans le prompt. Le prompt fourni par le développeur est composé d'une part, du besoin métier décrivant le service, et d'autre part, des spécificités techniques que seul le développeur saura exprimer pour garantir la bonne génération.

Par exemple, le développeur peut demander la génération d'un formulaire, ainsi que les routes (CRUD *Create, Read, Update, Delete*, opérations de base pour la persistance des données) associées. Il précisera les détails attendus, sur la forme (ex : intégrer des *watermarks* dans les *inputs*, à l'aide de clés i18n pour gérer l'internationalisation) et sur le comportement attendu (ex : le *http post* autorise la mise à jour partielle de l'objet *user* où seuls les champs modifiés sont envoyés dans la requête).

L'exemple précédent illustre le fait que ce type d'assistant est avant tout profitable pour les profils expérimentés, qui sont capables d'exprimer finement leur besoin, d'un point de vue fonctionnel mais aussi technique, et avoir une vision claire et complète de l'attendu. Un tel outil dans les mains d'un profil junior reste bien moins intéressant pour la qualité et la compétitivité de la tâche confiée. Autrement dit, l'IA générative génère un retour sur investissement très important pour les développeurs qui savent ce qu'ils veulent.

D'autre part, le développeur doit rester critique sur le code généré, pour être en mesure de détecter les hallucinations (écart entre ce qui est demandé et ce qui est produit) mais également pour pouvoir intégrer le code généré dans l'arborescence en place.

Cet écart de bénéfice produit selon la séniorité pourrait, à terme, créer un fossé générationnel sur le métier du développement (cf. risque évoqué dans le tableau plus bas).

Pour franchir le cap et mener le changement global de ses pratiques de développement, l'organisation doit acquérir une licence par utilisateur (développeur), quelles que soient sa séniorité et sa motivation à adopter un tel outil (« Je n'ai pas besoin de cela pour bien développer »).

Pour toutes ces raisons, les organisations peinent à s'appuyer sur l'IA générative dans leurs activités de développement, tant le retour sur investissement direct est difficile à constater.

Les autres risques potentiels concernent la sécurisation de l'application produite. Dans quelle mesure peut-on faire confiance à l'IA générative pour qu'elle produise un code sécurisé ? Les attaques menées par les cybercriminels sont de plus en plus innovantes et cherchent à exploiter la moindre faille, l'IA générative ne fait pas exception. Dans quelle

mesure une attaque telle que celle menée par Andres Freund sur le projet open source xz Utils[63] pourrait se généraliser ? Le principe consiste à encourager l'emploi de dépôts open sources « vérolés » pour l'entraînement d'IA génératives afin que celles-ci produisent un jour un code contenant la faille attendue. Ce scénario semble réaliste et demanderait au cybercriminel de :

- Contribuer sur des dépôts open source, ostensiblement avec des contributions licites, et discrètement avec des contributions vérolées ;
- Favoriser la réputation des dépôts (à l'aide de comptes factices venant noter, contribuer, forker, …). Le dépôt se rendrait fortement visible pour être utilisé à l'entraînement ;
- Cibler un périmètre spécifique (un langage exotique, une librairie spécifique) afin d'augmenter ses chances de peser dans l'inférence ;
- Être patient car l'attaque nécessite plusieurs mois, voire années d'élaboration.

Il est également possible d'imaginer qu'un cybercriminel crée un nombre très important de dépôts open source à l'aide d'une IA générative spécialisée, une fois de plus pour qu'une faille soit embarquée dans les modèles qui se seront entraînés sur ces dépôts.

[63] Korben. Une backdoor bien critique découverte dans xz Utils / liblzma. Blog Korben. 29 mars 2024. https://korben.info/backdoor-linux-faille-securite-critique-xz-utils.html

4.8.3. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Humain	Court terme : dégradation de l'employabilité des profils développeurs juniors. Long terme : raréfaction du profil développeur senior et architecte.	3	1	1	3	3	2	1	Réorganisation de la stratégie de recrutement et de fidélisation, avec valorisation des profils juniors : accompagnement à la montée en compétence, formation aux activités innovantes, mentorat, …
Humain	Changement dans la nature des tâches confiées aux développeurs.	1	1	1	3	3	1	1	Formation continue des développeurs, accompagnement au changement.
Modèle	Hallucination et conseils contre-productifs prodigués au développeur.	2	1	1	3	1	1	2	Formation continue des développeurs pour maintenir leur capacité à évaluer la qualité du code.
Humain	Fuite de propriété intellectuelle ou de code sensible (ex : application militaire).	4	4	4	2	2	2	1	• Renforcer les mesures de sécurité des données. Assurer la conformité avec les réglementations comme le RGPD. Sensibilisation des utilisateurs sur le niveau de confidentialité attendu sur le projet ; • Solution On Premise.
Données	Injection d'une faille de sécurité produite par un « anti-pattern »	4	4	4	2	2	2	1	• Contrôle de la base d'apprentissage ; • Limiter mes sources externes utilisées pour l'apprentissage ; • Poursuivre/renforcer les moyens habituels de détection de failles (scan de vulnérabilité, tests de pénétration, revues de code, auto-évaluation face aux règles OWASP[64] …) sur l'application.

[64] https://owasp.org/

4.9. Finance

4.9.1. Introduction

Les institutions financières ont été largement impactées par le développement des systèmes d'IA Générative. Une multitude de cas d'usage a donc émergé dans ce secteur, avec pour but l'amélioration et l'automatisation de plusieurs tâches, auparavant assurées par des humains. Ces tâches peuvent concerner à la fois des fonctions internes nécessaires à ces institutions dans l'accomplissement de leur mission ou des fonctions en relation avec la clientèle.

Le secteur financier, comme beaucoup d'autres secteurs d'activité, utilise et manipule beaucoup de documents (texte, tableau, image). L'IA Générative permet d'exploiter plus facilement de grands volumes de données, principalement textuelles, à la fois pour les employés (recherche d'information au sein d'un corpus documentaire) et les clients (réponses plus pertinentes et adaptées au contexte améliorant ainsi l'expérience client). La multi modalité des modèles permet également de traiter des données non structurées, telles que la voix ou les images, de manière plus industrielle, automatisant par la même occasion certaines tâches très consommatrices de temps (par exemple, les contrôles de conformité relatifs à la protection de la clientèle).

Pour le secteur financier comme pour les autres secteurs, les grandes familles d'usage comprennent donc : les agents conversationnels, les systèmes de recherche augmentés (RAG : *Retrieval Augmented Generation*), les systèmes pour résumer, les systèmes d'assistance ou de création de contenu.

L'usage de l'IA Générative dans le secteur bancaire progresse prudemment, car, malgré les possibilités offertes par ces nouveaux modèles, ces derniers ne sont pas exempts de défauts, et il est nécessaire de bien évaluer leur performance par rapport aux risques encourus avant tout déploiement généralisé de tels modèles, dans un contexte très réglementé.

La présence de biais dans les bases d'entraînement peut également avoir un impact sociétal, qui peut être amplifié si le système est déployé par des institutions systémiques. De façon générale, les impacts de l'utilisation de l'IA générative sont à évaluer dans le cadre de chaque cas d'usage afin de s'assurer de la mise en place des moyens adéquats pour atténuer les risques.

4.9.2. Cas d'usage

Le chatbot conseiller financier

Le cas d'usage hypothétique étudié dans ce document concerne une IA conversationnelle ou *chatbot* qui accompagne les clients en prodiguant des conseils financiers en termes d'investissements. Les réponses doivent couvrir un ensemble limité de sujets, en lien avec la demande des clients et dans le cadre de conseils financiers.

Le chatbot serait mis en place par une banque ou une institution financière (exemple : gestionnaire d'actifs), dans l'objectif de renforcer l'accompagnement des clients dans leur démarche d'investissements financiers.

Déploiement et utilisation

Les utilisateurs potentiels du chatbot sont déterminés sur la base de critères d'éligibilité définis par l'institution financière. Ces utilisateurs peuvent présenter des niveaux hétérogènes d'aversion au risque et de connaissances liées aux produits financiers (profits et risques).

Le chatbot est basé sur un modèle de langage pré-entraîné (LLM) optimisé avec les documents et les données des clients de l'entreprise pour assurer des réponses précises et contextualisées. Un système de feedback est également mis en place pour recueillir les avis des utilisateurs, permettant des ajustements continus pour optimiser l'utilité et l'efficacité du chatbot.

4.9.3. Matrice des risques

L'analyse des risques ci-après propose une évaluation sur la base de niveaux d'impact propres à ce cas d'usage précis.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	Données d'entraînement **non représentatives** du périmètre d'application ou **non exhaustives** : langue différente de la langue cible, textes non représentatifs de la diversité de la clientèle cible en termes d'âge, … **Risque financier** : La non-représentativité de la base peut donner lieu à des réponses non adaptées à certaines catégories de clients et donc in fine à la perte de clients potentiels. **Risque social :** L'absence de certaines catégories de la base d'entraînement peut créer des problèmes de biais. Exemple : clients/prospects qui ne maîtrisent pas la langue de l'IA, personnes fournissant des prompts moins précis à cause d'une culture financière limitée. **Risque légal** : L'absence de données concernant certains profils de clients et proposition de produits non adaptés.	3	4	3	2	1	3	1	• Vérifier la représentativité des données d'entraînement et leur exhaustivité par rapport au cas d'usage ; • Mettre en place des contrôles appropriés pour détecter d'éventuels biais, en mesurant à titre d'exemple les disparités de performance du chatbot pour différents groupes d'utilisateurs.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Modèle	**Contenu généré inexact (hallucination) ou incomplet** Proposition de **contenu toxique, irrespectueux ou inapproprié ou en dehors du cas d'usage**, non lié à l'objectif de conseil financier du chatbot. Exemple : proposition de conseil médical. Proposition d'un **contenu qui manipule l'utilisateur et l'oriente dans des directions non adaptées à son profil de risque**. Exemple : proposition de conseils financiers à fort risque de perte en capital pour des clients présentant une forte aversion au risque.	3	4	4	3	1	2	1	Tester la performance, les vulnérabilités et la robustesse du modèle via un certain nombre de techniques : • Tests contrefactuels en agissant sur les données pour évaluer la capacité du chatbot à limiter les hallucinations ; • Ajouter un module de détection de prompts toxiques ou non-adaptés à l'objectif du chatbot (guardrails) ; • Adversarial prompting : génération de prompts adverses pour étudier la robustesse du chatbot et sa capacité à contrer ces attaques et pour tester la performance du module de détection de contenus toxiques ; • Tests sur les réponses du chatbot afin de s'assurer que les réponses sont complètes et explicites en termes de risques associés aux conseils financiers.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Humain	**Dans l'usage :** Manque de sensibilisation des utilisateurs aux risques des IA génératives. Exemple : présence de biais de confirmation et/ou d'automatisation. Absence de limitation dans les sujets de conversation possibles. Non connaissance des utilisateurs des responsabilités des différentes parties prenantes. **Durant la phase de développement et monitoring :** Manque de surveillance de l'utilisation de l'IA. Exemple : une mauvaise utilisation du chatbot peut conduire à un impact environnemental négatif, dû à l'utilisation excessive non justifiée.	2	3	3	2	1	2	3	• Ajouter un module de topic identification pour éviter le mauvais usage du chatbot ; • Assurer la transparence : s'assurer que les utilisateurs sont bien informés qu'il s'agit d'un chatbot, que ce système présente des risques ; • Mettre en place des contrôles pour s'assurer que l'utilisateur est une personne éligible et apte à utiliser ce système ; • S'assurer que l'ensemble des utilisateurs sont conscients de leurs responsabilités et leurs droits ; • Mettre en place un score de pertinence pour accompagner les utilisateurs et éviter les biais d'automatisation ; • Mettre en place un cadre de monitoring qui surveille la performance et les risques du chatbot ; • Exploiter le user feedback pour améliorer les sorties du chatbot ; • Création d'une librairie de prompts pour former les utilisateurs à faire de bons prompts (en fournissant des exemples sur des questions récurrentes) et prioriser les réponses pré-générées pour les questions fréquemment posées.

4.10. Juridique

4.10.1. Introduction

En premier lieu, il est intéressant de rappeler que la CNIL s'était déjà (en décembre 2017) interrogée sur l'utilisation des algorithmes et l'encadrement des risques individuels et collectifs possibles au regard des discriminations et des biais potentiels ainsi que de la « démutualisation et de perte de logique collective » qu'ils pouvaient provoquer[65].

En second lieu, rappelons que dans un article[66] paru dans Dalloz IP/IT en octobre 2017, Sandrine Chassagnard-Pinet s'interrogeait sur l'algorithme, « prestataire de services juridiques » pouvant devenir un « producteur de droit ».

Enfin, dans un article[67] paru dans Dalloz IP/IT en octobre 2017, Gaël Chantepie écrivait que « les algorithmes seront utiles aux juristes pour leur permettre de dire le droit qui pourrait être, s'ils ne prétendent pas dire le droit qui devrait être ».

Plus proche de nous, lors d'une enquête[68] effectuée entre le 13 et le 21 juillet 2023, le fournisseur, LexisNexis Legal & Professional®, a interrogé 643 professionnels du droit en France, dont 221 du segment Avocat, 281 du segment Entreprise et 141 du segment Notaire, et a constaté que 77% des répondants ont considéré que les outils d'IA générative (IAG) augmenteront l'efficacité des professionnels du droit.

Alors que beaucoup de cas d'usage se sont développés depuis 2017 de nombreuses questions demeurent encore actuelles.

4.10.2. Cas d'usages

Les cas d'usages idoines au sein d'une direction juridique

Nous avons relevé 7 cas d'usages qui sont, semble-t-il, les plus recherchés, au sein d'une entreprise, à savoir :

---

[65] CNIL. Comment permettre à l'Homme de garder la main ? Décembre 2017. https://www.cnil.fr/sites/cnil/files/atoms/files/cnil_rapport_garder_la_main_web.pdf

[66] Sandrine Chassagnard-Pinet. Les usages des algorithmes en droit : prédire ou dire le droit. Dalloz IP/IT. Octobre 2017. https://lilloa.univ-lille.fr/handle/20.500.12210/2730

[67] Gaël Chantepie. Le droit en algorithmes ou la fin de la norme délibérée. Dalloz IP/IT. Octobre 2017. https://lilloa.univ-lille.fr/handle/20.500.12210/32729?locale-attribute=en

[68] LexisNexis. L'IA Générative & les Professionnels du Droit. Septembre 2023. https://go.lexisnexis.fr/etude_ia_intelligence_artificielle_lexisnexis_france

- Rechercher une information juridique via des bases de données juridiques, de la jurisprudence. L'objectif est d'accompagner les services juridiques afin d'obtenir des informations juridiques pertinentes ;
- Analyser un contrat car la lecture et l'actualisation juridique de clauses essentielles des contrats peuvent permettre la réduction des risques d'erreurs humaines ;
- Traduire afin que la traduction automatique de documents et la génération de contrats selon la langue du pays concerné soit plus rapide et plus efficace. La mise en place de glossaires permet de gagner du temps et donc de réduire les coûts ;
- Prédire une décision judiciaire qui va permettre non seulement la mise en place d'une stratégie juridique plus pertinente en analysant la jurisprudence existante mais aussi une prédiction de conclusions de litiges ;
- Automatiser des tâches ayant pour conséquence la réduction de tâches à faible valeur ajoutée comme la rédaction de mails, de documents standard ;
- Aider à la décision en s'appuyant sur l'analyse et le conseil dont la finalité est un gain de temps humain et financier ;
- Gestion de la relation client par l'intermédiaire de l'analyse des données des clients aboutissant à une amélioration de la satisfaction client.

Toutefois, l'utilisation de l'IA Générative peut présenter certains risques et nous vous en proposons quelques-uns en sachant que cette liste est non exhaustive :

- Dans la qualité des données d'entraînement, si les données sont biaisées ou incomplètes (hallucinations, biais, …), les résultats peuvent s'avérer erronés ou partiels.
- Dans le biais algorithmique, des recommandations juridiques ou des décisions générées peuvent être erronées voire inventées ;
- Dans la traduction, des erreurs de traduction peuvent se produire. De même, un manque de compréhension du contexte juridique peut survenir. Enfin, on peut s'interroger sur la confidentialité et la sécurité des données transmises et traduites ;
- Dans le manque de transparence qui questionne à propos de la pertinence d'une recommandation ou d'une suggestion. La question de l'acceptabilité et de la responsabilité peut aussi être posée ;
- Dans la confidentialité des données où le partage de données à caractère personnel voire sensible peut interroger. Ces données peuvent être compromises dans l'hypothèse de cybersécurité par exemple ;

- Dans la fiabilité des résultats où des résultats imprévisibles voire incohérents peuvent être proposés.
- Dans le cadre de la responsabilité, car en cas d'erreur ou de préjudice causé par une décision générée par une IA Générative, il peut être difficile de déterminer qui est responsable, le concepteur du modèle, l'utilisateur ou les deux ;
- Dans le cadre du droit de la propriété intellectuelle, l'IA Générative interroge sur la création et la protection des œuvres, sur le plagiat, par exemple.

Il est donc souhaitable voire nécessaire de mettre en place des "processus" de validation, de transparence et de gouvernance adaptés aux conséquences éthiques, légales et juridiques qu'entraîne l'utilisation de l'IA Générative dans le secteur juridique d'une entreprise. De même, il est impératif de former et de sensibiliser toute partie prenante lors de l'utilisation de l'IA Générative.

Les cas d'usages potentiels dans les différents domaines du droit

L'IA Générative pose question dans toutes les catégories du droit tant par ses enjeux que par ses risques et interroge sur son impact sur les libertés fondamentales telles que la liberté d'expression, le droit à la vie privée, à un procès équitable ou la liberté de réunion ou d'association.

L'utilisation de l'IA Générative se développant, elle a et aura pour conséquence la nécessité d'une régulation de celle-ci eu égard aux risques inhérents qu'elle entraîne. Tant les risques individuels, comme les risques pour la vie privée ou la fuite de données, que les risques collectifs comme les risques sociaux sont nombreux et non exhaustifs. Il est nécessaire de s'interroger sur la protection des libertés fondamentales et l'articulation du développement de l'IA Générative au regard de la responsabilité civile et pénale.

Dans le cadre du droit pénal, par exemple, les notions de sécurité, d'éthique et de légalité interrogent. Quelles mesures législatives, techniques et éducatives doivent être mises en place pour garantir le respect des libertés fondamentales ?

En articulant cette réflexion sur les différentes matières du droit, nous pouvons lister quelques risques, là encore, non exhaustifs comme :

- En droit pénal, s'interroger sur la preuve et la décision judiciaire ;
- En droit civil, sur la responsabilité civile et l'automatisation de contrats ;
- En droit de la propriété intellectuelle sur la protection des œuvres et le plagiat ;

- En droit de la vie privée et de la protection des données sur la collecte, l'utilisation et la sécurité des données ;
- En droit de la concurrence sur les pratiques des entreprises qui vont s'appuyer sur des IA Génératives pour concurrencer sur un même marché d'autres entreprises ;
- En droit du travail sur les conditions de travail, l'automatisation de tâches juridiques ;
- En droit administratif sur les décisions administratives et leur transparence ;
- En droit constitutionnel sur le respect des libertés fondamentales ;
- …

Autrement dit, nous vous proposons, à partir de la matrice des risques, une déclinaison possible des risques potentiels liés à l'utilisation d'une IA Générative en analysant les causes pouvant générer ces risques, leurs impacts et les pistes de remédiation proposées.

Dans le cadre de la "donnée", du "modèle" et de "l'humain" de cette matrice des risques, nous établissons deux matrices envisageables :

- L'une sur les cas d'usages possibles dans une direction juridique ;
- La seconde dans le droit pénal au travers de la preuve et de la décision judiciaire.

En conclusion, « Le défi de la régulation à venir sera de parvenir à encadrer les usages de l'IA Générative et à s'assurer que son développement demeure compatible avec les principes européens, sans entraver l'innovation et l'émergence de nouveaux acteurs, français ou européen »[69].

Les cas d'usages potentiels en droit pénal dans le domaine de la preuve et de la décision judiciaire

Nous avons relevé quelques cas d'usages qui sont, semble-t-il, les plus attendus, dans le domaine de la preuve comme :

- Collecter de grandes quantités de données issues de sources différentes afin de pouvoir identifier des éléments de preuves ;
- Détecter des attitudes, comportements pouvant être considérés comme équivoques ;

[69] Philippe Pradal et Stéphane Rambaud. Rapport d'information sur les défis de l'intelligence artificielle générative en matière de protection des données personnelles et d'utilisation du contenu généré. Présidence de l'Assemblée nationale. 14 février 2024. https://www.assemblee-nationale.fr/dyn/16/rapports/cion_lois/l16b2207_rapport-information.pdf?v=1709804244

- Reconstituer des scènes de crimes via des modèles 3D générés par l'IA Générative ;
- Détecter des fraudes financières ;
- Contribuer à une meilleure prise de décision humaine au regard des éléments matériel, moral et légal constitutifs d'une infraction ;
- Analyser des vidéos de masse ;
- Rapprocher des entités par l'analyse de données et établir des scores de rapprochement ;
- Classer des données très nombreuses et permettre une meilleure efficacité dans la recherche de l'élément matériel de l'infraction ;
- Utiliser l'IA Générative dans le cadre de l'identification biométrique.

Concernant la décision judiciaire, quelques cas d'usages possibles sont :

- Aider à la décision ;
- Permettre un meilleur respect des délais dans le cadre de la procédure pénale ;
- Utiliser des modèles prédictifs afin d'évaluer une probabilité de récidive ;
- Surveiller les réseaux sociaux ou autres plateformes si risque potentiel d'attentat par exemple ;
- Assister les professionnels dans le processus pénal à travers une acceptabilité de l'IA Générative « raisonnable et « raisonnée ».

Toutefois, l'utilisation de l'IA Générative peut présenter certains risques et devenir une atteinte aux libertés individuelles et les matrices des risques qui suivent en illustrent quelques-uns (en sachant que cette liste est non exhaustive).

4.10.3. Matrices des risques

Direction juridique

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Humain	**Dépendance** : son utilisation ne doit pas remplacer l'analyse juridique par la partie prenante car possibilité de prendre des décisions erronées, fausses et/ou aléatoires. Ex : Outil d'aide à la rédaction de documents juridiques, de courriels, de conclusions.	4	4	3	2	1	1	1	Analyse humaine indispensable : l'humain doit être le dernier lecteur et celui qui prend la décision individuellement ou collectivement.
Modèle	**Evolution du droit** : attention à l'interprétation du droit ! Quelle transparence sur la construction et le fonctionnement du logiciel proposant un système d'IAG ?	3	3	2	2	2	1	1	L'utilisateur de la plateforme doit prendre en considération l'évolution du droit et son interprétation : une veille juridique est toujours nécessaire et doit se faire via un logiciel IAG **et** humain (complémentarité).
Données & Modèle	La **qualité des données** utilisées par l'algorithme peut être imprécise, incomplète, peu pertinente et entraîner un risque de l'utilisation d'un modèle d'IAG peu fiable.	3	3	2	1	1	1	2	• Prévoir une correction des biais, des discriminations ; • Mettre en place des recours si des biais sont constatés ; • Nécessité d'actualiser en processus continu tout modèle d'IAG.
Données	**Confidentialité des données** : le système d'IAG peut nécessiter l'accès à des données sensibles ou confidentielles. Il faut garantir la sécurisation et la protection des données contre tout accès non autorisé.	4	4	2	2	2	1	2	• Prévoir une double authentification ; • Garantir une sécurisation et une protection des données contre tout accès non autorisé.

Droit pénal au travers de la preuve et de la décision judiciaire

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Humain	Risque de ne pas pouvoir vérifier, expliquer et justifier le système.	1	1	1	2	2	3	1	Supervision et dimension humaines afin de ne pas porter atteinte, entre autres, aux libertés individuelles.
Modèle	Recours excessif à des modèles d'IAG dans un but de gain de temps et de réduction de coûts. Cette volonté de productivité peut entraîner une trop grande dépendance aux modèles d'IAG.	3	3	1	2	1	2	1	Le développement des systèmes d'IAG doit être encadré eu égard aux risques d'atteintes à des principes fondamentaux : mettre en place un « dispositif de garantie humaine ».
Données & Modèle	Grande quantité de données (masse de données) : risques de biais, déséquilibre entre les données discriminantes et non discriminantes permettant de répondre aux éléments légaux constitutifs d'une infraction (matériel, légal et moral).	2	4	1	3	1	2	1	• Formation des professionnels dans le cadre du processus pénal et développement de leur technicité ; • Evaluation des données : ○ Examiner et analyser les données utilisées afin de vérifier si celles-ci peuvent inclure des données pertinentes et nécessaires limitant les biais ; ○ Evaluer les biais en vérifiant si ces données peuvent conduire à une discrimination intentionnelle ou pas (ethnie, genre, …) ; ○ Vérifier et évaluer que ces données sont proportionnelles, objectives et impartiales ; ○ Pouvoir justifier de l'utilisation d'une donnée ou son exclusion.
Données	Multiplicité des sources des données pouvant entraîner une non-maîtrise de celles-ci dans leur qualité et dans leur collecte.	3	4	1	3	1	2	1	Formation de professionnels capables de les analyser et de distinguer les données discriminantes et non discriminantes.

4.11. Service Client

4.11.1. Introduction

Très rapidement après l'avènement de ChatGPT, au début de l'année 2023, les premières expérimentations de l'IA générative dans les métiers du service client ont été mises en place. Certaines études[70] ont alors estimé que les professions du service client, aux côtés de celles du développement informatique, seraient parmi les plus impactées.

Cette rapide adoption a probablement été facilitée par la présence déjà bien établie de l'IA "traditionnelle" (machine learning, NLU…) dans ce secteur, comme en témoigne l'utilisation fréquente de chatbots (par exemple, TOUTOUI[71] proposé par la SNCF). Désormais, le potentiel de l'IA générative ouvre des perspectives entièrement nouvelles et à une échelle inédite : il n'est plus nécessaire de relier des centaines, voire des milliers, de situations ou d'intentions différentes à la réponse la plus pertinente.

Cette évolution marque un tournant vers la **personnalisation à grande échelle** des interactions, permettant à l'IA de s'adapter en temps réel aux besoins spécifiques de chaque utilisateur, enrichissant ainsi l'expérience des clients par une **compréhension** et une anticipation plus fine de leurs demandes.

Dans le prolongement de cette personnalisation, il devient primordial de reconsidérer **la formation des agents** du service client. L'intégration de l'IA générative exige une évolution des compétences : les agents doivent désormais développer une capacité à gérer des problématiques complexes où l'intervention humaine est indispensable. Les programmes de formation continue et de développement professionnel doivent donc être repensés pour aligner les compétences humaines avec les capacités de l'IA, assurant une **synergie optimale entre les agents et la technologie**.

Mais cet avancement significatif introduit également une série de nouveaux risques que les entreprises doivent apprendre à gérer pour éviter des conséquences négatives importantes, notamment en termes d'image de marque et de pertes financières.

---

[70] Avishek Adhvaryu, Allison Bailey, Frank Breitling, Tim Fenton, JinK Koike. The Path to Generative AI Value Begins with a Workforce Diagnostic. Boston Consulting Group. October 31, 2024. https://www.bcg.com/publications/2023/assessing-the-impact-of-generative-ai-on-workforce-productivity

[71] https://tout-oui.sncf.com/

Par exemple, les **hallucinations**, caractéristiques de l'IA générative, peuvent entraîner une volatilité des réponses pour une même question posée par le client, ce qui implique la mise en place d'un **contrôle qualité** particulièrement complexe.

4.11.2. Cas d'usage

Le chatbot eCommerce

Le chatbot est conçu pour interagir directement avec les clients d'une entreprise de distribution dans le cadre d'un achat sur le site e-commerce de cette entreprise. Ce chatbot est accessible sur une partie du site internet de l'entreprise (pages « contact » ou « FAQ ») pour traiter les demandes de service client après-vente.

Cet outil numérique vise à améliorer l'expérience des clients en fournissant une assistance instantanée et disponible 24/7.

Le chatbot est basé sur un modèle de langage (LLM) pré-entraîné spécifiquement avec les documents et les données de l'entreprise, notamment les procédures de service client. Pour assurer des réponses précises et contextualisées. Des prompts définissant les règles de rédaction de réponses (« Bonjour Madame » ou « Bonjour Sophie ») sont également fournies au LLM pour respecter la tonalité de réponse de l'entreprise.

4.11.3. Matrice des risques

Causes		Impacts							Remédiations
Famille	**Description**	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	**Données incomplètes** (procédure inexistante) qui entraînent des hallucinations où le modèle va inventer la procédure (par exemple, autoriser un échange en boutique alors que c'est impossible).	2	1	2	3	1	1	1	Ne jamais solliciter le modèle en cas d'absence de procédure.
Données	Envoi non minimisé de la **donnée personnelle** au modèle : envoi du prénom et nom du client au modèle alors que cette donnée n'est pas nécessaire pour générer la réponse.	1	3	4	1	1	1	1	Pseudonymiser les données personnelles avant envoi au modèle et reconstruction de la réponse juste avant restitution au client final.
Modèle	**Modèle non contraint** sur le périmètre du service client : réponses inutiles (hors périmètre) générant des coûts inutiles ou réponses inappropriées (insultes…).	3	2	4	1	1	1	3	Contraindre le modèle sur un périmètre fixe avec un langage approprié.
Humain	**Aucune possibilité d'escalader** à un humain en cas de désaccord.	2	4	4	1	1	1	1	Offrir une "porte de sortie" au client final (par exemple si le client pose plusieurs fois la même question ou dans le cas où le client s'énerve).
Humain	**Manque de transparence** sur l'utilisation de l'IA.	1	2	4	1	1	1	1	Affichage clair et lisible de l'utilisation d'une IA.
Humain	Aucune ou mauvaise **supervision des performances du chatbot** : qualité des réponses aléatoires avec des réponses hors procédures (remboursements inappropriés ou solutions impossibles).	3	2	4	3	1	1	1	• Boucle de contrôle automatisée basée sur les retours des utilisateurs finaux ; • Mise en place d'analyses des performances avec des alertes automatisées (sur les volumes et les retours des utilisateurs, sur des mots clés interdits) ; • Contrôle aléatoire humain des réponses.

4.12. Conseil

4.12.1. Introduction

La fonction conseil est chargée de fournir des analyses stratégiques, des recommandations et un accompagnement opérationnel visant à optimiser la performance et l'efficacité des clients qu'elle adresse. Souvent, le consultant apporte la vision « benchmark du marché », et aide à diffuser les bonnes pratiques.

L'IA générative pourrait certainement augmenter et accélérer un certain nombre de tâches propres à la fonction conseil. Face à ce constat, il convient d'identifier les tâches du consultant pouvant être en partie déléguées à cette technologie de celles où l'apport du consultant restera essentiel.

Aujourd'hui, l'IA générative n'est pas encore suffisamment mature pour fonctionner de manière autonome. Le consultant avisé est celui qui l'utilise pour accélérer son travail, mais continue d'appliquer son esprit critique et son expertise sur les résultats proposés.

Elle représente tout de même un levier de performance majeur pour un cabinet de conseil. Elle peut permettre d'améliorer son organisation interne par une **meilleure gestion de ses ressources**, **par l'extension de sa force commerciale** et par **l'amélioration de la qualité** des **prestations** vendues.

4.12.2. Cas d'usages

Accélération des processus de « staffing »

Un des enjeux majeurs des cabinets de conseils et de savoir identifier, lorsqu'une mission est gagnée, les consultants pouvant être mobilisés sur celle-ci. Ce processus, appelé « staffing », pourrait ainsi être accéléré par une IA générative, en intervenant sur le processus d'identification du CV du consultant correspondant aux besoins de la mission. Elle pourrait en effet faire le tri automatiquement entre les profils de consultants, en comparant la description de ceux étant disponibles et la description de la mission. Une validation humaine serait conservée en bout de chaîne pour un contrôle de la qualité. Le prérequis indispensable pour que ce cas fonctionne est la **qualité des données** utilisées, à savoir, les cahiers des charges des missions, et les fiches profils des consultants.

Traduction des ressources

L'intelligence artificielle générative pourrait offrir des avantages considérables en matière de traduction. En permettant une traduction précise et rapide de documents, de CV de consultants et de retours d'expérience de missions, l'IA générative pourrait faciliter la communication et la collaboration avec des clients et des partenaires dans différentes régions du monde. Cela permettrait non seulement de surmonter les barrières linguistiques, mais aussi de présenter des livrables adaptés à l'environnement (e.g. noms des acteurs économiques clés, prise en compte d'éléments macro-économiques) de chaque pays, renforçant ainsi l'efficacité et la pertinence des services proposés à l'échelle globale.

Levier de performance pour la relation client

L'usage de l'IA générative permet également d'assembler de manière cohérente les données de plusieurs sources. Dans le cas du conseil, elle pourrait par exemple faire l'historique des échanges avec un prospect, ou l'ensemble des réponses techniques éditées pour les clients d'un secteur en lien avec un savoir-faire particulier.

- Notamment, elle pourrait éditer des campagnes de prospection ultra personnalisées, à l'aide des historiques de conversation, de données CRM, et des données open-source. Elle pourrait aussi aider à digérer le contenu des appels d'offres particulièrement volumineux pour en extraire l'essentiel ;
- Aussi, grâce à sa capacité à trouver des informations dans plusieurs sources de données, et en l'assistant avec un dispositif de RAG (*Retrieval Augmented Generation*), elle pourrait permettre d'analyser plus rapidement des rapports, aider à anticiper les évolutions du marché, et proposer des recommandations pertinentes pour permettre à leurs clients de s'adapter.

L'intégration d'outils IA Générative pour accroître la valeur des services proposés dans le cadre des missions

Sans aucun doute, et ce changement est déjà amorcé, l'IA générative pourrait améliorer toute la chaîne des services proposés par les cabinets de conseil.

Pour le domaine du *Project Management Office* (PMO), l'utilisation d'outils comme COPILOT permet déjà d'automatiser les tâches récurrentes, d'optimiser la planification des réunions, de rédiger des comptes-rendus de manière quasi automatisée, ainsi que d'améliorer la rapidité d'édition et la pertinence des livrables destinés au client.

Les résultats des analyses et des *benchmarks* pourraient être plus facilement restitués grâce aux fonctions de visualisation de données.

Surtout, l'IA Générative pourrait aider à résoudre un des problèmes majeurs dans les cabinets de conseil : la gestion des connaissances. En effet, d'abord en facilitant la consolidation et l'anonymisation des supports produits dans les missions. Ensuite, à l'aide d'un RAG propre, de faciliter la production de nouveaux livrables similaires.

L'intégration de l'intelligence artificielle générative dans la fonction de consultant pourrait représenter une avancée significative pour l'efficacité opérationnelle et la pertinence des services offerts par les cabinets de conseil. Cependant, il est crucial de reconnaître que l'IA Générative ne peut remplacer entièrement le rôle du consultant, notamment en ce qui concerne la compréhension des dynamiques organisationnelles, les compétences interpersonnelles et la capacité à fournir des solutions personnalisées.

4.12.3. Matrice des risques

Cette analyse est réalisée sur l'un des quatre cas d'usages évoqués : l'intégration d'outils d'IA générative pour accroître la valeur des services proposés dans le cadre des missions. Ce cas d'usage a été choisi car il s'expose à de nombreux risques qui méritent une illustration.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	Production de livrables contenant des informations confidentielles d'autres clients que celui pour qui sont réalisés les livrables.	1	4	4	1	1	1	1	• Vérifier l'anonymisation des livrables clients utilisés pour l'entrainement / réentrainement du modèle, et le dispositif de RAG ; • Limiter les données fournies au modèle pour son entraînement et son RAG avec des données strictement nécessaires pour de futures missions.
Modèle	Production de livrables comportant des informations erronées ou non vérifiables.	1	1	3	3	1	1	1	• Inclure des liens sourcés dans le livrable pour permettre la vérification des réponses par un humain ; • Mettre en place un dispositif de garantie humaine pour contrôler aléatoirement des recommandations du modèle et les comparer à des recommandations faites par des humains ; • Mettre en place un RAG pointant vers un ensemble de documents dont la qualité a été vérifiée avant leur inclusion.
Humain	Surutilisation du modèle par les consultants (paresse intellectuelle, prompts peu efficaces, ...), générant un nombre de requêtes trop important.	1	1	1	4	1	1	2	• Former les consultants au prompt engineering pour limiter la répétition de requêtes ; • Sensibiliser aux limites de ce que l'IA générative peut produire et à son impact environnemental.

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Humain	Perte de compétences au sein des consultants du fait d'une dépendance trop importante aux outils d'intelligence artificielle.	1	1	1	4	1	1	2	• Maintenir des formations aux consultants sur la production de livrables (étude de marché, analyse et synthèse de données) ; • Assurer la transparence du modèle pour permettre aux consultants de pouvoir étudier la méthode de production du livrable et pouvoir la vérifier.
Humain	Production de livrables de mauvaise qualité du fait d'une mauvaise utilisation du modèle par les consultants.	2	1	3	3	1	1	1	• Former les consultants au prompt engineering pour améliorer la pertinence des prompts et des réponses ; • Sensibiliser les consultants aux limites de ce que peut produire l'IA générative.

4.13. Data Science

4.13.1. Cas d'usage

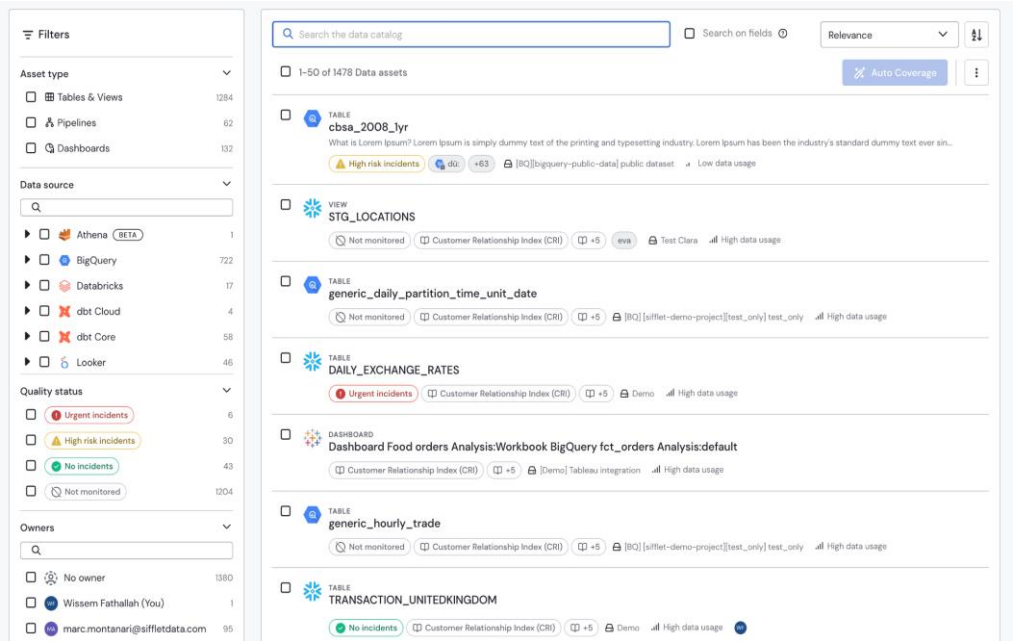
Description automatisée des données dans un logiciel de Data Observability et de Data Quality

Un logiciel de *Data Observability* et de *Data Quality* offre aux entreprises la possibilité de se fier pleinement à leurs données, quelle que soit leur échelle. Conçu pour être utilisé par les équipes techniques et les équipes métier, ce logiciel permet la mise en œuvre d'une surveillance complète de la qualité des données, prenant en compte les aspects techniques et métier de celle-ci. Cette surveillance peut être effectuée via une interface *no-code/low-code* ou de manière programmatique grâce aux APIs.

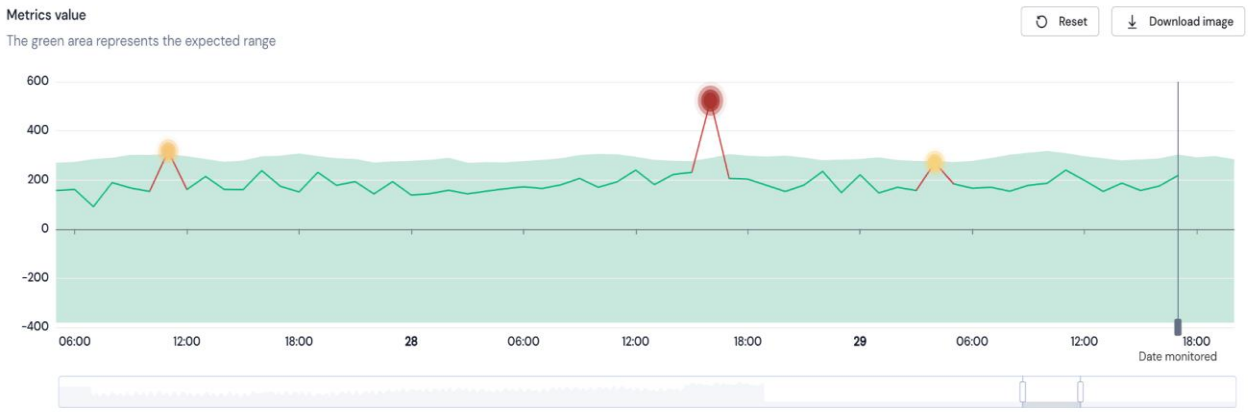
Pour assurer la qualité des données au sein des entreprises traitant un volume important de données, les logiciels de *Data Observability* s'appuient fortement sur l'intelligence artificielle.

Une plateforme de *Data Observability* et de *Data Quality* se compose généralement de trois éléments clés pour offrir une expérience utilisateur complète :

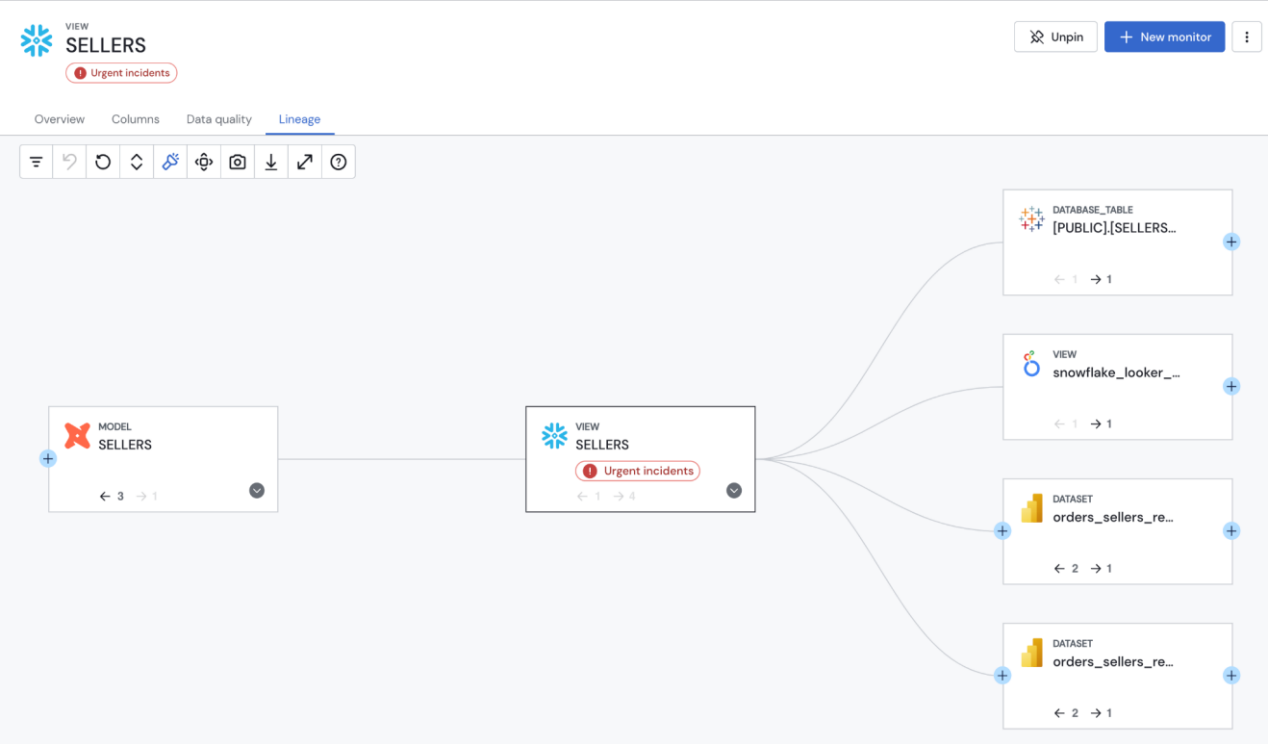
- Un ***Data Catalog*** : reflétant les objets de données présents dans les plateformes utilisateurs, tels que des tables dans des entrepôts de données ou des tableaux de bord dans un outil de visualisation de données. Ce composant permet une meilleure compréhension des données pour mettre en place les contrôles de qualité les plus pertinents.



- Un système automatique de **monitoring de données** : permettant de détecter les anomalies grâce à l'intelligence artificielle.



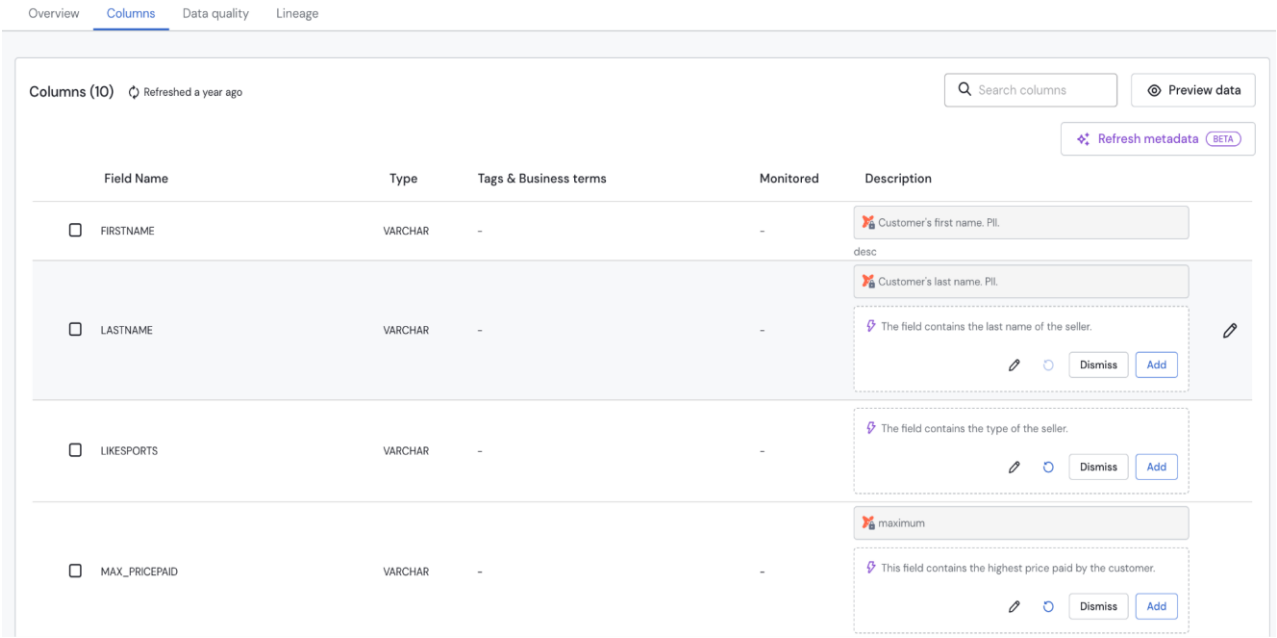
- Un ***Data Lineage*** : établissant les dépendances de données à travers toutes les technologies de données connectées à la plateforme. Cet élément accélère la résolution des anomalies de qualité de données en détectant la source de l'anomalie et son impact sur les utilisateurs de la donnée.



Déploiement et utilisation

Dans le cadre du *Data Catalog*, une fonctionnalité a été mise en place pour permettre la description et la classification automatiques des données. Cette fonctionnalité peut, par exemple, décrire un champ dans une table de données ou détecter des données personnelles telles que des noms, prénoms ou adresses.

Pour ce faire, la fonctionnalité utilise un modèle d'Intelligence Artificielle Générative auquel sont exposées des métadonnées et des échantillons de données provenant des champs que l'utilisateur souhaite décrire ou classifier.



4.13.2. Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	La qualité des propositions est fortement dépendante de la qualité des données utilisées pour fine-tuner le modèle ou faire de l'augmentation de contexte par RAG.	4	3	4	4	1	1	1	Faire du monitoring de la qualité des données exposées aux modèles afin de proactivement détecter les anomalies et les résoudre avant que ça impacte le résultat final.
Modèle	Proposition de contenu toxique, irrespectueux ou inapproprié ou en dehors du cas d'usage, non lié à l'objectif de la fonctionnalité.	4	4	4	3	1	1	1	Ajouter un module de détection de descriptions toxiques ou non-adaptées à l'objectif du chatbot (guardrails).
Humain	Il est essentiel que les descriptions de données mises à disposition des utilisateurs dans le milieu professionnel soient précises afin de garantir une utilisation adéquate. En effet, il existe un risque de fournir des descriptions erronées ou incomplètes, ce qui peut induire les utilisateurs en erreur. Par exemple, il pourrait être indiqué qu'un champ contenant une erreur peut être utilisé pour calculer un chiffre d'affaires d'une entreprise.	4	4	4	3	1	1	1	Montrer les descriptions comme suggestions aux utilisateurs et non pas comme une vérité absolue, et reposer sur les experts pour les confirmer ou les modifier.

4.14. Logistique et transport

4.14.1. Introduction

Dans le monde du transport maritime, le personnel navigant doit pouvoir faire face à tout type de panne sur le navire, pour cela l'expérience est primordiale. Mais avec l'arrivée à bord des navires de technologies de plus en plus modernes, la maintenance et la réparation reposent sur des procédures complexes. Certains cas imprévisibles requièrent pour le personnel navigant de trouver une solution rapide.

Plutôt que de consulter sur papier ou tablette un mode opératoire documenté de plusieurs centaines de pages et complexe à lire, l'intelligence artificielle peut apporter une réponse rapide et précise sur une grande quantité de documents, mais aussi apporter une solution guidée de dépannage, hors guide technique, pour proposer une réponse rapide à un problème bloquant.

L'IA générative pré-entrainée a plusieurs avantages pour le personnel navigant :

- Compréhension de documentation dans diverses langues et restitution dans la langue de l'utilisateur ;
- Capacité à engranger une masse de données sur les modes opératoires et documentations techniques adaptés au métier ;
- Interprétation des photos et images fournies ;
- Maîtrise de la donnée utilisée via le pré-entrainement (données de la société et des navires, réglementation à bord, gestion des incidents, etc.) ;
- Génération de rapports d'incident/résolution d'incident et de checklists ;
- Génération de KPIs ;
- Accès rapide aux historiques de maintenance ;
- Gestion de l'inventaire du navire ;
- Possibilité d'avoir une conversation orale avec l'IA, ce qui permet d'avoir les mains libres pour la réparation.

4.14.2. Cas d'usage

Agent conversationnel pour la gestion des pannes et de la maintenance sur un porte-conteneur

L'objectif ici est de fournir un agent conversationnel pré-entrainé sur les données techniques et sur les processus de la compagnie maritime qui encadre les modes opératoires internes du personnel de navigation. Cette documentation est conséquente et peut être une charge pesante pour le marin. L'idée est donc d'optimiser et de fluidifier ces tâches à bord grâce à l'agent qui questionnera toute cette documentation. Le pré-prompt intégré à l'IA fonctionne de manière à bien identifier la machine, la panne, l'environnement et les compétences du marin, pour adapter au mieux sa réponse et son accompagnement. L'IA est entraînée pour se référer uniquement aux données techniques et en cas d'absence de solution dans la documentation, avertira systématiquement le marin que la solution proposée doit être vérifiée et validée par un humain. L'agent conversationnel rappellera à chaque intervention les risques d'utiliser l'IA et les principes de sécurité à bord selon la réparation à faire (EPI, travail en milieu clos, etc.).

Contexte

Un marin est à bord d'un porte-conteneur en pleine mer. Une panne survient sur l'un des moteurs de l'unité de réfrigération des conteneurs qui cesse de fonctionner, ce qui menace la cargaison périssable. Dans cette situation critique, un agent conversationnel basé sur l'IA générative peut jouer un rôle crucial pour diagnostiquer et résoudre le problème rapidement.

Description du cas d'usage

1. Détection et Signalement de la Panne

Le marin remarque que la température dans un conteneur réfrigéré a commencé à augmenter. Il contacte immédiatement l'agent conversationnel via son appareil mobile.

Marin : "Agent IA, le conteneur réfrigéré numéro 56 montre une température anormale. Que dois-je faire ?"

Agent IA : "Bonjour. Pouvez-vous m'envoyer une photo du panneau de contrôle du conteneur 56 et toute information supplémentaire sur le problème ?"

Le marin prend une photo du panneau de contrôle et la télécharge via l'interface de l'agent.

2. Analyse de la panne

L'agent IA utilise la reconnaissance d'image pour analyser la photo du panneau de contrôle. Il détecte une alerte de surchauffe et une indication de dysfonctionnement du compresseur.

Agent IA : "D'après la photo, il semble que le compresseur du conteneur 56 soit en surchauffe. Voici une liste de vérifications à effectuer pour diagnostiquer plus précisément la panne :

- Vérifiez les niveaux de réfrigérant ;
- Inspectez les connexions électriques pour détecter tout signe de dommage ;
- Assurez-vous que les filtres à air ne sont pas obstrués. »

3. Préparation avant l'intervention

Avant de commencer l'intervention, l'agent IA rappelle au marin les précautions de sécurité à prendre :

Agent IA : "Avant de commencer, assurez-vous de suivre ces consignes de sécurité :

- Portez des équipements de protection individuelle (EPI) : gants, lunettes de protection et vêtements adaptés ;
- Vérifiez que la zone de travail est bien ventilée ;
- Assurez-vous que l'équipement est hors tension avant de procéder à toute inspection ou réparation ;
- Ayez un extincteur à proximité en cas d'incident.

Une fois que vous aurez appliqué ces règles de sécurité, veuillez me confirmer que nous pouvons continuer."

4. Assistance à la maintenance

Le marin suit les étapes fournies par l'agent IA et découvre que le niveau de réfrigérant est bas. Il signale cette découverte à l'agent.

Marin : "Le niveau de réfrigérant est bas. Que dois-je faire ensuite ?"

Agent IA : "Vous devez maintenant recharger le réfrigérant. Voici les instructions pour effectuer cette opération en toute sécurité. Veuillez suivre chaque étape attentivement."

L'agent IA fournit au marin un guide détaillé avec des photos et des descriptions textuelles pour recharger le réfrigérant.

5. Documentation et suivi

Après avoir rechargé le réfrigérant, le système revient à la normale. Le marin signale la résolution du problème à l'agent.

Marin : "Le réfrigérant a été rechargé et le système fonctionne à nouveau normalement."

Agent IA : "Excellent travail. Je vais consigner cette intervention dans le journal de maintenance du navire. Avez-vous besoin d'autre chose ?"

L'agent enregistre automatiquement toutes les informations pertinentes de la panne et de l'intervention dans le système de gestion de maintenance du navire, assurant une traçabilité complète pour les inspections futures.

Bénéfices de l'IA Générative dans ce cas d'usage

- Réactivité et Disponibilité 24/7 : l'agent IA est disponible à tout moment pour assister l'équipage, réduisant le temps de réaction en cas de panne ;
- Précision du diagnostic : en utilisant des photos et des checklists, l'IA peut aider à diagnostiquer les problèmes avec précision ;
- Guidage détaillé : fournit des instructions détaillées et spécifiques basées sur les données de l'équipement du navire ;
- Documentation automatique : enregistre automatiquement les interventions, facilitant le suivi et la conformité ;
- Réduction des erreurs humaines : en suivant des instructions précises, les marins peuvent éviter des erreurs coûteuses.

4.14.3.Matrice des risques

Causes		Impacts							Remédiations
Famille	Description	Financier	Juridique	Réputationnel	Opérationnel	Organisationnel	Social	Environnemental	
Données	Manque de fraîcheur de la donnée : pas de mise à jour en temps réel.	1	1	1	3	1	1	1	• Réentrainement régulier avec data à jour (vis-à-vis du matériel, des opérations de maintenance, de l'inventaire) ; • RAG : mise à jour du contexte avec les nouveaux documents (ex : mises à jour sur la liste des problèmes identifiés sur le matériel et des solutions associées).
Données	• Stockage des données à l'international ; • Perte de la confidentialité des données (cf. réglementation locale) ; • Problèmes d'accès à la donnée.	1	3	4	3	3	1	1	Faire en sorte de travailler avec un hébergeur adapté au travail international, qui respecte la confidentialité, la sécurité, en accord avec la politique de l'entreprise ou stocker en local dans chaque navire pour consulter les données.
Modèle	Hallucinations et conseils contre-productifs.	1	4	1	3	1	1	1	Entraînement sur la documentation technique et formation du personnel en amont, ce qui permet d'éviter des mauvaises manipulations dues à une hallucination.
Modèle	Manque d'exhaustivité des informations produites par l'IA. Ex : liste des risques (électrocution, brûlures, incendies, etc.)	2	3	1	4	1	1	4	Les marins sont formés au travail sur les équipements électriques, l'IA devra rappeler les risques à chaque fois et les EPI nécessaires.
Humain	Le marin n'est pas assez formé ou ne garde pas son esprit critique pour challenger la réponse de l'IA. En appliquant les conseils de l'IA, il peut endommager l'équipement via une mauvaise réparation.	2	3	1	4	2	1	4	• L'IA est basée sur les données techniques de chaque équipement ; • Formation du marin à l'usage du nouvel outil numérique.

HUB FRANCE IA

5. Synthèse des remédiations

Contributeurs :

- ***Thomas Argheria**, Manager, Wavestone*
- ***Gérôme Billois**, Partner, Wavestone*
- ***Martin D'Acremont**, Consultant – Wavestone*

Les risques de l'IA Générative
Juillet 2024

5. Synthèse des remédiations

Les cas d'usage d'IA générative se multiplient sur toutes les verticales métiers. Il ne fait aucun doute que cette technologie prend toute sa place dans nos organisations, et qu'elle s'implémentera durablement. Les risques associés aux IA génératives ne sauront être intégralement atténués. En particulier, pour les *Large Language Models* (ou LLM), leur fonctionnement intrinsèque et leur méthode d'entraînement ne permettent pas d'en faire des agents vraiment « intelligents ». Ce sont de simples « mémoires vives » qui restituent une connaissance apprise lors de la phase d'entraînement. Cette mémoire sait être certes savamment adaptée et redistribuée, mais ça n'en reste pas moins simplement une mémoire, et ces systèmes sont dépourvus d'intelligence humaine. Ils peuvent donc commettre des erreurs, sources de risques pour l'entreprise. Cependant, il existe des mesures « de **remédiation** » qui permettent de réduire considérablement les risques associés à cette nouvelle technologie. Elles sont présentées dans la section qui suit, mais il convient de préciser son contour :

1. Pour mettre en place leur projet d'IA générative, la grande majorité des organisations va **s'appuyer sur un modèle élaboré par un fournisseur** (Google, OpenAI, Meta, Microsoft, Anthropic, Mistral…). Ces modèles dits « de fondation » sont déjà (pré) entrainés. Ainsi, les mesures présentées ici interviennent donc exclusivement après la phase d'entraînement du modèle. Ce sont des mesures actionnables pour les organisations, qui ne sont pas dépendantes des décisions d'un fournisseur.
2. Aussi, les mesures de remédiation présentées ne cherchent à couvrir que les risques spécifiques à l'IA générative. Comme toute application, un système d'IA générative doit également **mettre en place des mesures de protection contre les risques** usuels, dont les risques **de cybersécurité** (sécurité des API, des plugins, chiffrement des données et des flux, journalisations d'évènements…). Ce qui suit considère que l'ensemble du socle cybersécurité « classique » est déjà mis en œuvre pour les composants du système d'IA (documentation, évaluation des risques, plan d'action de mitigation …).

3. Pour **aider la compréhension et la lecture**, les mesures de remédiation sont présentées de la manière suivante :

- Par grandes familles de risques : modèles, utilisateurs, données ;
- Par ordre croissant de complexité : complexité technique et d'implémentation ;
- Rattachées à un risque principal, même si certaines peuvent en couvrir plus largement.

5.1. Réduire les risques liés au modèle

5.1.1. Limiter la génération de contenu non désirable

L'un des risques principaux lors de l'utilisation d'un système d'IA générative est la génération d'un **contenu non désirable** (e.g. contraire à l'éthique ou aux valeurs de l'organisation, biaisé, illégal, inexact…). Au premier rang de ce problème le phénomène **d'hallucination**[72], qui correspond à la production d'un contenu cohérent intellectuellement, mais factuellement incorrect. Il existe aussi des problématiques de génération de contenu illégal, notamment en lien avec la production de contenus protégés par licence.

Quelques exemples d'hallucinations

- Lors de sa sortie en août 2023, Google Bard (désormais Gemini), affirmait que le télescope James Webb (opérationnel depuis 2021), avait pris des photographies de la première exoplanète (alors que celles-ci ont été prises en 2004)[73] ;
- Gemini à nouveau, souhaitant augmenter la représentativité des personnes de couleurs dans les réponses produites, a finalement conduit à des aberrations historiques : des images de Vikings et de Nazis noirs. [74]

Comme mentionné en introduction, ces risques sont liés intrinsèquement à la manière dont les modèles de génération de contenu sont entrainés. Les LLM notamment fonctionnent selon une méthode de génération qui construit une suite de mots constituant la réponse la *plus probable* selon (1) l'entraînement réalisé sur des quantités énormes de données, souvent open-source, et parfois protégées par le droit d'auteur, et (2) le contenu du prompt de l'utilisateur. Le fonctionnement est cependant non

[72] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, vol. 55, n° 12, pp. 1-38. November 2022. https://arxiv.org/pdf/2202.03629

[73] Le Monde. Google perd 7 % à la Bourse de New York après une erreur de Bard, son nouveau robot conversationnel. 9 février 2023. https://www.lemonde.fr/pixels/article/2023/02/09/google-perd-7-a-la-bourse-de-new-york-a-la-suite-d-une-erreur-de-son-tout-nouveau-robot-conversationnel_6161118_4408996.html

[74] Radio France. Intelligence artificielle : Google suspend la création d'images de personnes sur son IA Gemini après des critiques. 22 février 2024. https://www.francetvinfo.fr/internet/intelligence-artificielle/intelligence-artificielle-google-suspend-la-creation-d-images-de-personnes-sur-son-ia-gemini-apres-des-critiques_6382006.html

déterministe car une part d'aléatoire est ajoutée (par l'intermédiaire d'un paramètre de « **température** »[75]).

Parce que les modèles apprennent sur ces quantités énormes de données, pour lesquelles la fiabilité et l'absence de biais ne sont pas toujours contrôlables (d'autant plus que les modèles de fondation ne donnent pas d'information précise sur les données d'entraînement qu'ils ont utilisées), les hallucinations, mais aussi l'apprentissage sur des données qui sont normalement protégés par droit d'auteur sont inévitables. Ce risque, même avec l'application des mesures de réduction présentées ci-dessous, ne sera jamais complétement supprimé.

Mesure n°1 : Spécialiser le modèle pour ses cas d'usage

Selon le cas d'usage à implémenter, choisir le modèle le plus adapté en fonction de ses capacités, et si possible, le réentraîner sur des données fiables, actualisées et spécialisées pour le domaine du cas d'usage, pour lui permettre de gagner en précision, et donc de réduire le risque d'hallucination.

De la même manière que lorsqu'on construit une application classique il est nécessaire de s'assurer de l'utilisation des librairies appropriées, il convient pour les applications utilisant de l'IA générative de s'assurer de la pertinence du modèle utilisé. En effet, certains modèles sont plus adaptés que d'autres pour certaines tâches comme le montre par exemple le comparatif du *Center for Research on Foundation Model* de l'université de Standford[76] C'est un premier filtre à prendre en compte selon le cas d'usage que l'on étudie.

Ensuite, les modèles génériques ne sont pas adaptés pour tous les cas d'usage et parfois, ils doivent être spécialisés. En l'occurrence, certains modèles de LLM rendus disponibles en open-source (Llama, Mistral) peuvent être réentraînés partiellement pour mieux correspondre à la mise en place d'un cas d'usage en particulier. Ainsi, le projet Bhashini en Inde a pour objectif de créer des jeux de données dans les 22 langues officielles, afin qu'ils puissent être utilisés pour réentraîner des modèles de fondation. Ce réentraînement est indispensable, car les données sur lesquelles sont entraînés les

[75] Cf Glossaire.

[76] Percy Liang et al. Holistic Evaluation of Language Models. *Center for Research on Foundation Models, Standford University.* November 2021. https://arxiv.org/pdf/2211.09110

modèles de fondation ne contiennent pas assez de contenus dans ces langues pour permettre aux modèles d'être nativement assez performants pour les traductions. [77]

Mesure n°2 : Mettre en place la « Retrieval Augmented Generation [78] », ou RAG.

*Constituer une source de données de confiance sur lesquelles le modèle peut s'appuyer, exclusivement ou partiellement, pour générer ses réponses afin d'accroître leur pertinence***.**

Les hallucinations produites par les modèles sont en partie liées à un problème de qualité des données d'entraînement, et notamment de fraîcheur et d'actualisation. Pour pallier ce problème, il est possible de paramétrer le modèle pour qu'il utilise uniquement ou prioritairement ses capacités sur des sources de données de confiance, et ainsi réduise la priorité de la mémoire issue de son entraînement. S'appuyer sur ces sources de données de confiance, c'est faire du RAG, ou « *Retrieval Augmented Generation* ». Cette technique permet non seulement de réduire les hallucinations mais aussi de faire levier sur les données internes à l'entreprise.

Cependant, pour que le RAG soit efficace, il faut évidemment maintenir la qualité des documents qui le composent.

Mesure n°3 : Durcir les paramètres de génération de contenu du modèle

Durcir les paramètres du modèle (température, définition du master prompt, verbosité des réponses…) pour orienter le comportement de l'agent génératif, et maîtriser les réponses émises.

Lors du déploiement des systèmes d'IA générative, il est possible de durcir les paramètres de base du modèle pour orienter son « comportement » face aux requêtes des utilisateurs. Parmi ces paramètres, il convient de mentionner la **température**, la définition du ***master prompt***[79], et la **verbosité** des réponses. Les réglages sur ces mesures sont souvent facilités par les interfaces (plus ergonomiques) proposées sur les plateformes d'IA générative des fournisseurs de Cloud. Pour les modèles récupérés en *open-source*, ce durcissement doit faire l'objet de développements supplémentaires.

---

[77] Milin Stanly. India turns to AI to capture its 121 languages for digital services. *Indi/ai.* December, 20 2023 . https://indiaai.gov.in/article/india-turns-to-ai-to-capture-its-121-languages-for-digital-services

[78] En français, génération augmentée de récupération.

[79] Cf. Glossaire.

La température permet de régler le niveau d'improvisation ou de créativité du modèle dans ses réponses. Au minimum, cela contraint le modèle à se reposer exclusivement sur les données qui sont fournies pour produire ses réponses. C'est ce qu'il est possible d'utiliser pour un *chatbot* qui agit comme conseiller juridique. En revanche pour un *chatbot* marketing, peut-être que plus de créativité est souhaitable. À nouveau, tout dépend du cas d'usage, mais plus la température est basse, plus le risque d'hallucination et la créativité sont faibles.

Le *master prompt* désigne, quant à lui, les instructions données au modèle pour l'aider à répondre aux questions des utilisateurs. Il s'agit d'un *prompt* générique, fournit à l'IA avant même qu'il traite une question, pour l'initialiser et la cadrer. Ces instructions peuvent contenir le rôle ou le format de réponse à respecter, ou la demande de ne pas répondre si le modèle n'a pas l'information demandée (et ainsi de ne rien « inventer »). Un *master prompt* correctement rédigé permet de préciser le comportement souhaité, et d'insérer des exigences de sécurité (e.g. « ne livre aucune information sur l'entreprise », « ne fournis aucune information à caractère personnel », …).

Enfin, pour éviter aux modèles d'être trop « verbeux », c'est-à-dire de fournir des réponses trop amples avec des informations superficielles, il est possible aussi de régler la verbosité des réponses. Les systèmes d'IA générative décomposant les entrées des utilisateurs en jetons, ou « *tokens* », pour les traiter puis renvoyer une réponse constituée de nouveaux jetons, il est possible, en limitant les réponses à un certain nombre de jetons, de limiter le risque de diffusion de résultats non désirables.

Mesure n°4 : Filtrer les réponses non désirables

Mettre en place un filtrage des réponses produites par le modèle, pour écarter les non désirables. Cette capacité peut être outillée avec un pare-feu IA.

Même avec la mise en place des trois premières mesures, et parce que le risque d'hallucination est impossible à réduire complétement, il peut être nécessaire de mettre en place un filtrage sur le contenu généré. Cette couche permet de bloquer l'émission d'un contenu non désirable si celui-ci est malgré tout produit. Des solutions émergent sur le marché, et notamment des « *AI Firewall* », ou pares-feux pour IA. Ces outils sont eux-mêmes des LLM, localisés entre l'utilisateur et le modèle, et entraînés pour être des filtres de sécurité. Ils vont filtrer des prompts malicieux ou des réponses indésirables produites par le modèle. Ainsi, ils agissent à la fois sur les entrées et les sorties.

Ce filtrage nécessite cependant d'avoir un cadre de modération clair et bien rédigé (quelles valeurs doivent être respectées, quel contenu doit être considéré comme indésirable) sur lequel le pare-feu peut s'appuyer pour filtrer. Dans le cadre d'un *chatbot* pour des conseils financiers par exemple, il sera pertinent de filtrer des conseils médicaux, ou alors des réponses qui ne respectent pas les obligations légales en matière de transparence.

Les limites de cette mesure résident d'abord dans le fait qu'on ne peut jamais totalement dresser la liste exhaustive des cas de contenus non désirables. Ensuite, le cadre de modération n'est jamais complétement étanche et peut être contourné par des effets de forme, du fait de la nature de ces modèles : c'est **le *prompt injection***[80].

Mesure n°5 : S'assurer de la pertinence des données de réentraînement

S'assurer que les données utilisées pour le réentraînement du modèle soient fiables et de bonne qualité, en vérifiant l'absence de biais, de tentative d'empoisonnement, et la pertinence des données.

De la même manière que pour la phase d'entraînement sur un jeu de données initial, tout réentraînement du modèle doit être fait à partir de jeux de données fiables. Ce réentraînement peut avoir lieu au moment de la spécialisation du modèle (mentionnée plus haut) ou d'un réentraînement sur tout ou partie des données de production (e.g. les conversations avec les utilisateurs). Pour ces cas particuliers où les modèles sont réentraînés, il est nécessaire que la donnée utilisée fasse l'objet de vérification, pour s'assurer de leur pertinence, de l'absence de biais, et de tentatives d'empoisonnement.[81] On peut imaginer par exemple qu'un *chatbot* de conseils financiers soit réentraîné sur la base d'échanges uniquement avec une population présentant une aversion au risque. Ainsi, in fine, il n'émettra que des conseils adaptés à cette population mais non adaptés à ceux ayant une forte appétence au risque.

[80] OWASP. LLM01: Prompt Injection. LLM Top 10 risks. https://genai.owasp.org/llmrisk/llm01-prompt-injection/

[81] C'est l'exemple fameux de Tay, ce *chatbot* de Microsoft qui en quelques heures s'est transformé en un activiste d'extrême droite. Voir
Morgane Tual. A peine lancée, une intelligence artificielle de Microsoft dérape sur Twitter. Le Monde. 24 mars 2016. https://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html

Plusieurs techniques sont possibles pour appliquer ces vérifications : un contrôle par échantillonnage, par recherche de mots clés (liste d'insultes…) ou même un retraitement des données par un LLM de vérification.

Mesure n°6 : Mettre en place du Reinforcement Learning from Human Feedback, ou RLHF

Améliorer les performances d'un système d'IA en appliquant les principes de l'apprentissage supervisé : un agent humain vient vérifier les prévisions faites par le modèle, corrige les écarts trop importants, et catalyse sa progression.

Le RLHF ou « *Reinforcement Learning from Human Feedback* » (Apprentissage par renforcement à partir de rétroaction humaine) est l'une des techniques les plus efficaces, mais aussi les plus chères, pour améliorer les modèles. Elle consiste à la mobilisation d'analystes humains, qui vont noter plusieurs réponses émises par le modèle, afin d'améliorer sa précision. Ce sont les experts du domaine métier sur lequel porte le cas d'usage qui valident la pertinence des résultats proposés (c'est une des techniques utilisées par OpenAI pour ChatGPT).

Une autre forme de RLHF consiste à utiliser directement les commentaires et les évaluations des utilisateurs vis-à-vis des réponses proposées pour guider l'apprentissage. Ces retours utilisateurs permettent d'identifier les meilleures réponses et les plus mauvaises. Ils sont utilisés pour entraîner un autre système d'IA, qui jouera le rôle de modèle de « récompense » et sera en mesure d'évaluer la qualité des réponses du premier modèle.

Mesure n°7 : Mettre en place un dispositif de garantie humaine

Mettre en place un collège de supervision pour évaluer les performances d'un modèle en mobilisant des experts du domaine métier concerné. Ces derniers produisent des réponses à des requêtes utilisateurs, les comparent à celles de l'IA, et identifient les écarts et des pistes d'amélioration.

Dans certains cas, l'impact d'hallucination de la part du système peut entraîner des conséquences importantes, par exemple pour un système d'IA générative de conseils médicaux. Une réflexion a été lancée par l'AFNOR pour imaginer un dispositif permettant de contrôler étroitement la conception et le fonctionnement d'un système d'IA en santé. Cette réflexion a pour origine l'article 17 de la loi de bioéthique de 2021 qui encadrait le

recours à l'IA en matière de santé publique, et l'AI Act européen qui prévoyait des exigences spécifiques de contrôle humain pour les systèmes d'IA à haut risque. [82] [11]

Ce type de dispositif est appelé **dispositif de garantie humaine**. L'objectif est de permettre au concepteur[83] de détecter et corriger les défaillances du système d'IA lorsqu'il répond aux utilisateurs, c'est à dire les écarts entre les recommandations faites par le système d'IA et celles faites par des experts humains.

Cette mesure permet d'accompagner le concepteur tout au long de la mise en place d'un cas d'usage : de la conception à la supervision en production. C'est en cela qu'elle va plus loin que le RHLF, qui intervient uniquement après la mise en production du modèle. Un **collège de supervision** est mis en place pour suivre les risques et mesures identifiés en amont par l'analyse de risque projet :

- **Des experts réviseurs** : experts du domaine métier, ils vont comparer les résultats du modèle avec des recommandations qu'ils auraient eux-mêmes produites. Dans le cas d'un *chatbot* de conseils en matière de santé, ce seront des médecins ;
- **Des représentants utilisateurs** : ici ce sont les professionnels du domaine métier en question, qui vont mettre à disposition le produit à des usagers (par exemple pour notre cas d'usage santé : les pharmaciens, d'autres médecins, des cliniques). Ils apportent une vision sur les cas d'usage pour lesquels ils mettent le système à disposition des bénéficiaires ;
- **Un tiers expert** du domaine métier à l'image des experts réviseurs mais avec une position d'indépendance vis-à-vis des autres protagonistes ;
- Le **concepteur** du système a la charge de l'implémentation des éventuelles mesures correctives identifiées ;
- Des **représentants des bénéficiaires ou utilisateurs finaux** : associations d'usagers, représentants associatifs, etc… Ils permettent de remonter des retours sur leur expérience du parcours usager et des points de difficulté dans l'utilisation du système d'IA.

Ce collège sera chargé d'analyser un échantillon aléatoire de dossiers traités par le système d'IA. Les mesures correctives peuvent prendre des formes variées et aller au-delà d'une correction purement technique au niveau de l'algorithme : formation,

---

[82] AFNOR. Garantie humaine des systèmes fondés sur l'intelligence artificielle en santé. AFNOR Spec 2213. Mai 2024. https://www.boutique.afnor.org/fr-fr/norme/afnor-spec-2213/garantie-humaine-des-systemes-fondes-sur-lintelligence-artificielle-en-sant/fa205274/419909

[83] Ici, le concepteur peut être toute personne physique ou morale (ex. autorité publique, entreprise, chef de projet, chef de produit …) qui développe ou fait développer un système d'IA.

modification de la notice d'utilisation … Si l'analyse a permis d'identifier un nouveau risque pour le système d'IA, celui-ci est inclus dans l'évaluation des risques, de même que la ou les mesures correctives associées.

Mesure n°8 : Mettre en place un système d'IA constitutionnelle

Améliorer les performances d'un système d'IA en entraînant un second modèle à estimer si les réponses du premier respectent un ensemble donné de règles, et utiliser ces évaluations pour améliorer le modèle en continu.

Les systèmes d'IA dite « constitutionnelle »[84] sont des modèles construits à partir d'un modèle de fondation à qui l'on a fourni une « **constitution** ». Cette constitution regroupe un ensemble de règles que doit suivre le modèle pour générer sa réponse. C'est un peu comme si le modèle était réentraîné, sur la base de son propre cadre de modération. Le principe est similaire au RLHF, mais de manière automatisée et le référentiel n'est plus l'évaluation humaine, mais la constitution.

Lors de la phase d'entraînement, le modèle est entraîné à (1) produire des réponses, y compris des réponses toxiques, puis à (2) estimer si ces réponses respectent sa constitution et enfin à (3) les corriger jusqu'à avoir des réponses satisfaisantes. À partir de ce jeu de données, il est possible d'entraîner un deuxième système d'IA générative, qui notera la conformité des réponses proposées vis-à-vis de la constitution, et permettra au premier modèle d'ajuster ses réponses.

De tels modèles sont utiles pour des cas d'usage où il est nécessaire que les réponses générées par l'IA générative adhèrent strictement à des règles (légal, audit…).

5.1.2. Se protéger des tentatives malicieuses (incl. prompt injection, jailbreak)

Une des particularités des systèmes d'intelligence artificielle est que, contrairement à la plupart des systèmes informatiques classiques, il n'est pas nécessaire de gagner des droits sur un système (e.g. obtenir un mot de passe par exemple, rentrer dans le réseau, etc.) pour conduire une action malicieuse. Cela peut être réalisé directement à travers l'interface utilisateur, comme la fenêtre de dialogue du *chatbot*.

Avec les techniques de *prompt engineering* (l'art de requêter un modèle pour obtenir la meilleure réponse possible) sont nés aussi les techniques de *prompt injection* (l'art de

[84] Yuntao Bai et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint : 2212.08073*. December 2022. https://arxiv.org/pdf/2212.08073

requêter un modèle pour le faire sortir de son cadre de modération). Certaines de ces techniques vont jusqu'à permettre, par un jeu d'instructions dans le prompt, de reprogrammer complétement le modèle (*jailbreak* ou DAN, pour *Do Anything Now*).

Ces attaques ne sont possibles que parce que ces modèles ne sont pas véritablement « intelligents » et ne savent pas faire la différence entre des requêtes légitimes et malicieuses, et parce que les cadres de modération peuvent toujours être contournés par des effets de formulation et de forme.

À noter que le *prompt injection* peut être utilisé à plusieurs fins : extraire des données d'entraînement, obtenir des informations sur le paramétrage du modèle, faire agir le modèle en dehors de son cadre de modération, faire tenir au modèle des propos illégaux ou nocifs, etc.

Exemples connus de *prompt injection* :

- L' « exploit » de la grand-mère sur ChatGPT[85] : En demandant au modèle de prendre le rôle d'une grand-mère défunte, il était possible de lui faire produire des réponses qu'il s'interdisait par ailleurs (« comment fabriquer une bombe », « comment détruire l'humanité » …). Cet exploit a été rendu impossible depuis.
- Une équipe de chercheurs de l'université de Cornell, Technion, et de l'Israël Institute of Technology, a élaboré un prompt capable d'extraire les données personnelles des utilisateurs des assistants de messagerie, mais surtout de s'auto-répliquer pour se diffuser d'un utilisateur à un autre. [86]

Evidemment, toutes les mesures évoquées précédemment permettent aussi de réduire ce risque, mais les mesures décrites dans la suite vont plus loin dans la sécurisation.

Mesure n°9 : Maintenir ses modèles à jour

Mettre en place un processus de gestion de l'obsolescence du modèle utilisé, afin de le maintenir à jour pour éviter de conserver dans son application des vulnérabilités déjà corrigées par le fournisseur mais également connues des acteurs malveillants.

Comme toute application, un système d'IA générative doit être intégré dans les processus de sécurité, et notamment la gestion des vulnérabilités. Les tests de *prompt*

[85] Bastien L. ChatGPT jailbreak : toutes les techniques pour désactiver la censure. lebigdata.fr. 4 juin 2024. https://www.lebigdata.fr/chatgpt-dan

[86] Stav Cohen, Ron Bitton, Ben Nassi. Here Comes The AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications. arXiv preprint arXiv:2403.02817. March 2024. https://arxiv.org/pdf/2403.02817

injection effectués sur les modèles 3.5 et 4.0 de ChatGPT montrent que le dernier modèle est bien plus robuste face à des attaques. Et c'est logique, puisque les fournisseurs de modèles ont des équipes spécialisées, parfois de plus de 100 personnes, pour écumer internet, trouver les techniques de *prompt injection* échangées sur les forums, et corriger les failles.

Réaliser soi-même la veille des vulnérabilités sur les modèles de *Machine Learning* peut être fastidieux. Cependant, il est indispensable de se tenir informé auprès des fournisseurs sur les mises à jour des modèles, et en source ouverte sur les vulnérabilités existantes sur ces modèles, afin de choisir les versions les moins vulnérables.

Mesure n°10 : Offusquer les paramètres du modèle

Dissimuler au maximum les paramètres du modèle pour limiter la capacité d'un attaquant à comprendre le fonctionnement d'un modèle, et l'exploiter à des fins malveillantes.

Pour tirer avantage d'un modèle et construire des attaques efficaces, un attaquant va essayer d'obtenir un maximum d'informations sur la manière dont ce dernier est paramétré. Par des techniques de *prompt injection*, il peut extraire les règles de fonctionnement internes de ce dernier (ce qu'on appelle alors « *prompt leaking* »). C'est ce qu'il s'est passé pour le *chatbot* de Bing, Sidney, en 2023[87]. Avec ces informations, un attaquant identifie plus facilement les vulnérabilités de paramétrage, et peut ainsi construire des attaques plus performantes.

Ainsi, il convient de mettre en place une stratégie d'offuscation de ces données, et du *master prompt* notamment. Si cette mesure est relativement simple à mettre en œuvre, encore faut-il y penser !

Mesure n°11 : Contrôler les entrées utilisateurs

Réduire la marge de manœuvre d'un attaquant en limitant le format des requêtes utilisateurs, et/ou en filtrant les requêtes malveillantes.

Pour se prémunir en partie contre des attaques adverses ou de *prompt injection*, il est possible de contrôler le format des requêtes que les utilisateurs peuvent soumettre. Les

[87] Benj Edward. AI-powered Bing Chat spills its secret via prompt injection attack. *Ars Technica*. October 2, 2023. https://arstechnica.com/information-technology/2023/02/ai-powered-bing-chat-spills-its-secrets-via-prompt-injection-attack/

attaques « Dauphin » par exemple, où des assistants vocaux ont été déclenchés via l'émission de signaux inaudibles pour l'oreille humaine[88], auraient pu être évitées si le format des données permettant de déclencher les assistants avait été limité à une fréquence audible humainement.

Pour les LLM, limiter le type, la taille, le format, la langue utilisée dans les prompts permet de restreindre la marge de manœuvre des utilisateurs lors de sa conception. Ainsi, vous réduisez la surface d'attaque et le potentiel malveillant des requêtes. On peut citer d'autres méthodes de contrôle des entrées comme interdire l'utilisation de code python ou limiter le nombre de caractères possible pour le prompt.

Mesure n°12 : Transformer les prompts envoyés par les utilisateurs

Modifier les requêtes envoyés par les utilisateurs (changement de certains caractères, application d'opérations mathématiques, …) afin de dissimuler à un acteur malveillant des informations sur le traitement des requêtes par le modèle.

Il est recommandé lorsque c'est possible de modifier les prompts soumis sous d'autres formes. Le *prompt injection* est très sensible à des altérations sur des mots (exemple enlever un caractère, le modifier, etc.). Ainsi, si l'entrée malicieuse est modifiée, le taux de succès du *prompt injection* est considérablement réduit. Alors que si la requête est légitime, cette modification a peu d'impact sur la qualité.

Le retraitement des requêtes par le *master prompt* est déjà une forme de transformation, lorsque la requête est encore sous forme de phrase. Il est recommandé d'aller plus loin, en modifiant à nouveau le prompt une fois que celui-ci est transformé en données mathématiques traitées par l'algorithme. Cette transformation permet de cibler et filtrer les informations les plus importantes qui constituent le cœur de la question posée pour qu'elles soient les seules à être traitées par l'algorithme. Ainsi, le modèle sera moins sensible aux effets de forme induits par le *prompt injection*.

Pour aller plus loin, il est possible de systématiser la modification des caractères d'une demande de 10% par exemple, et de soumettre les prompts modifiés à plusieurs LLM. Les

[88] Damien Leloup. Une faille de sécurité permet de contrôler les assistants vocaux de Google, d'Apple ou d'Amazon. *Le Monde*. 7 septembre 2017. https://www.lemonde.fr/pixels/article/2017/09/07/une-faille-de-securite-permet-de-controler-les-assistants-vocaux-de-google-d-apple-ou-d-amazon_5182348_4408996.html

réponses sont ensuite moyennées, et ce résultat est redistribué in fine à l'utilisateur. La figure suivante extraite de [89] illustre cette méthode.

Cependant, parce qu'elle agit sur la composition des demandes, cette mesure peut avoir un impact sur la qualité des réponses proposées, en augmentant les coûts et les temps de latence du fait de l'utilisation de plusieurs LLM. Il est nécessaire de tester quel niveau de transformation constitue un bon équilibre entre protection du modèle et impact sur la qualité.

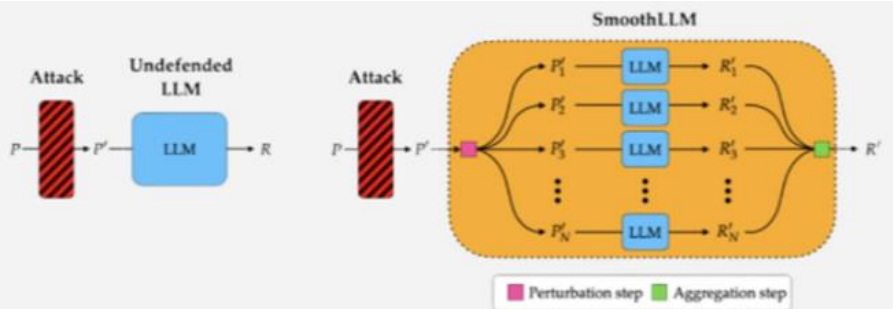


Figure : Architecture de SmoothLLM (d'après [89]). On perturbe le prompt P' qu'on fournit au LLM. Pour ne pas dégrader les performances, on effectue n perturbations différentes, puis on agrège les n réponses. Ainsi les propriétés sémantiques du prompt initial sont conservées.

Mesure n°13 : Réaliser un *redteam* IA

Comme souvent en cybersécurité, s'il est indispensable de réfléchir à la sécurité par défaut, et aux mesures nécessaires pour assurer la confiance d'un système dans la durée, il est indispensable de tester si cette réflexion a permis en pratique de sécuriser l'application. Le premier niveau de vérification est de vérifier que le système d'IA générative résiste à des prompts simples (e.g. « donne-moi le salaire de telle personne », « explique-moi comment faire une bombe », « donne-moi de la donnée sensible », ou ces éléments mais dans d'autres langues comme le japonais). Ces tests fonctionnels permettent de se prémunir des abus classiques et attendus des utilisateurs lambda. Il est nécessaire de les conduire via plusieurs profils utilisateurs avec des droits différents sur le système.

Ensuite, il faudra tester la résistance du système à des techniques de prompt injection plus avancées. Par exemple, le « *payload splitting* » consiste à diviser la demande en plusieurs parties en apparence non malicieuses, puis à demander à l'IA de combiner ses

[89] Alexander Robey, Eric Wong, Hamed Hassani, George J. Pappas. SMOOTHLLM : Defending Large Language Models Against Jailbreaking Attacks. arXiv preprint arXiv:2310.03684. October 5, 2023. https://arxiv.org/pdf/2310.03684v4

parties ce qui forme alors une demande malicieuse. Le « *context switching* » consiste lui à faire croire à l'IA à l'aide d'un prompt que l'on est dans un cadre légal, rassurant et éthique, pour lui faire produire des réponses peu éthiques ou illégales. En mixant ces concepts, et en adoptant une posture d'essai /erreur, il est tout à fait possible de contourner les cadres de modération en place.

Aujourd'hui, il existe sur le marché des équipes spécialisées dans les techniques de prompt injection semi-automatisé et outillé. Un article récent[90], publié par des chercheurs de l'Université Carnegie Mellon et du Centre pour la sécurité de l'IA, expose une méthode de création de prompts utilisant des techniques très poussées. Elles permettent de maximiser la probabilité que le modèle produise une réponse affirmative à des requêtes qui auraient dû être filtrées. Les prompts malicieux ainsi créés ne sont même plus compréhensibles par un cerveau humain, et pourtant ils permettent d'obtenir des résultats redoutables ![91]

Mesure n°14 : Détecter les tentatives malicieuses

Mettre en place des capacités de journalisation sur les projets d'IA générative, de détection des requêtes malicieuses, mais aussi d'investigation et de réaction.

Il est impératif de mettre en place un système de journalisation des événements sur les systèmes d'IA générative. Comme souvent en cyber, ce sont ces systèmes qui permettent de réaliser des investigations, de retracer le chemin emprunté par un attaquant, et de corriger le tir pour éviter qu'une attaque ne se reproduise.

Cela étant dit, c'est seulement pour les cas où une application d'IA générative est mise en place au niveau d'un processus critique pour l'entreprise, ou présente un risque d'atteinte à la réputation particulièrement élevé, qu'il peut être nécessaire de mettre en place des mécanismes de détection de tentative malicieuse.

En effet, aujourd'hui la mise en place de telles capacités est complexe et coûteuse, et la technologie IA générative n'est pas encore suffisamment mature pour intervenir sur des processus critiques pour les entreprises. Les *chatbots* IA agissent aujourd'hui plutôt

[90] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*. December 2023. https://arxiv.org/pdf/2307.15043

[91] Thomas Argheria, Pierre Aubret, Youssef Khouchaf. Quand les mots deviennent des armes : prompt Injection et Intelligence artificielle. *Risk Insight. Wavestone.* Octobre 2023. https://www.riskinsight-wavestone.com/2023/10/quand-les-mots-deviennent-des-armes-prompt-injection-et-intelligence-artificielle/

comme copilotes, et sont rarement autonomes pour réaliser des actions concrètes sur les systèmes, sans supervision humaine. Ainsi, il est difficile de justifier le coût de l'investissement sur la détection au regard du coût généré par la matérialisation du risque (ex : interruption de la disponibilité sur des processus non critiques).

Dans le cas où c'est effectivement nécessaire d'implémenter cette capacité, il est préférable de se tourner vers des fournisseurs spécialisés sur le marché[92]. Par ailleurs, le marché commence à voir l'apparition de CISRT spécialisés en AI[93]. En effet, la détection des tentatives malicieuses pour les systèmes d'IA implique une connaissance fine des menaces en la matière, du fonctionnement des algorithmes de *Machine Learning*, et de la remontée et l'interprétation des alertes.

5.2. Réduire les risques liés aux données utilisées par le modèle

5.2.1. Limiter la génération de contenu sensible, confidentiel, ou personnel

L'IA générative peut, dans certains cas, révéler des données censées rester confidentielles.

Bien sûr, cela n'est possible que si le système a accès à un moment donné, à ces données. Et cela peut intervenir dans plusieurs cas :

1. Si le modèle est réentraîné sur des données sensibles pour l'entreprise ou sur des données personnelles ;
2. Si les données du RAG sont des données sensibles ou personnelles ;
3. Si les modèles mettent en place une fonctionnalité de « mémoire ».

Dans ce dernier cas, les conversations avec le modèle sont conservées pour affiner la connaissance du modèle vis-à-vis de l'utilisateur. Parfois, ces données sont réinjectées dans le RAG.[94] La question se pose alors d'être en mesure de filtrer les données sensibles ou personnelles volontairement fournies par l'utilisateur.

---

[92] Gérôme Billois, Sleh-Eddine Choura, Henri du Périer. Radar 2024 de la sécurité de l'IA : panorama des solutions pour une IA de confiance. https://wwa.wavestone.com/fr/insight/radar-solutions-cyber-ia-de-confiance/

[93] Software Engineering Institute Establishes AI Security Incident Response Team. *Carnegie Mellon University*, November 28. 2023. https://www.cmu.edu/news/stories/archives/2023/November/sei-aisirt

[94] Bernard Marr. ChatGPT Gets a Memory – Here's All you Need To Know About This Groundbreaking Innovation. *Forbes*. April 15, 2024. https://www.forbes.com/sites/bernardmarr/2024/02/14/chatgpt-gets-a-memory--heres-all-you-need-to-know-about-this-groundbreaking-innovation/

Mesure n°15 : Protéger les données par défaut

Assurer une protection des données par défaut, en limitant l'utilisation de données sensibles ou personnelles dans l'apprentissage des modèles, soit en les écartant, soit en les transformant (minimisation, pseudonymisation, anonymisation).

Pour le premier cas, Il convient bien sûr de s'assurer que l'utilisation de données sensibles et personnelles pour le réentrainement d'un modèle est évitée, si ceci n'est pas absolument nécessaire.

Si l'utilisation des données personnelles ne peut être évitée, il conviendra d'essayer d'anonymiser ou de pseudonymiser les données. Cependant, les procédures de mise en conformité sont coûteuses en ressources humaines et en temps. Pour une alternative, voir les mesures 20 et 21.

Mesure n°16 : Respecter les politiques de gestion des identités et des accès

S'assurer que l'ensemble des populations utilisant le modèle ou intervenant sur son fonctionnement (e.g. administrateur, chef de projet, data scientist) aient des droits sur les données ou les composants du modèle strictement conformes aux politiques de gestion des identités et des accès appliqués au reste du système d'information de l'organisation.

Nous ferons ici un écart à notre posture introductive affirmant ne pas mentionner des mesures de cybersécurité classique. Pour le cas spécifique du RAG, les retours du terrain montrent que les politiques de gestion des accès et des identités (notamment le *Role Based Access Control*[95], ou RBAC) sont souvent insuffisamment respectées.

Premièrement, il convient d'assurer que les données sont protégées par défaut via une bonne implémentation des contrôles d'accès. Cela implique très classiquement d'identifier, de cartographier les accès, et de chiffrer et tracer les flux qui doivent l'être.

Il convient ensuite de s'assurer que les droits d'accès sur la donnée positionnée dans le RAG sont les mêmes que ceux nécessaires pour l'utilisation du *chatbot*. Par exemple, si je souhaite mettre en place un *chatbot* RH à destination de l'ensemble de mes employés pour répondre aux questions courantes, je ne mettrai pas dans mon RAG le fichier confidentiel avec l'ensemble des salaires des employés. En revanche, ce peut être le cas si le Chabot est réservé à une population qui a normalement accès au fichier.

[95] Cf. Glossaire

Par ailleurs, il faudra également s'assurer que les composants IA de l'application font eux aussi l'objet d'un contrôle RBAC. Par exemple, le chef de projet ne devrait pas avoir accès au modèle (cet accès est réservé au propriétaire de la plateforme GenAI ou le *Model owner* du modèle en interne), et la gestion du RAG ne doit être accordée qu'à des personnes triées sur le volet. Cette mesure protège aussi d'un risque d'empoisonnement de la donnée du RAG.

Mesure n°17 : S'assurer du respect du RGPD

En cas de manipulation de données personnelles, s'assurer que leur exploitation par le système d'IA générative soit conforme aux exigences du RGPD, quitte à réaliser une analyse d'impact sur les données personnelles.

Un modèle peut être amené à exploiter des données personnelles. Dans ce cas, les principes du RGPD devront être respectés. Il y a plusieurs cas de figure à prendre en compte :

1. Le cas où l'utilisateur lui-même fournit ses données personnelles dans ses prompts.

   Les utilisateurs peuvent eux-mêmes livrer un certain nombre de données personnelles dans le contenu de leur prompt. Si les discussions avec les utilisateurs sont conservées et utilisées pour améliorer le modèle :

   i. D'abord, s'assurer, si possible, de filtrer toute la donnée non nécessaire à la requête (par exemple un numéro de carte de sécurité sociale, des informations d'identification ou d'authentification pour des accès sur le système d'information, une date de naissance, etc.), en paramétrant les outils de filtrage intelligents déjà en place (e.g. pare-feu IA, évoqué plus haut, sur les entrées et sur les sorties) pour reconnaître les données personnelles et les filtrer.
   ii. Si des données personnelles sont collectées, il faudra justifier d'une base légale pour la collecte (consentement, contractualisation, ou intérêt légitime). L'intérêt légitime est la piste à privilégier. Dans ce cas, il faut réaliser un LIA (*Legitimate Interest Assessment*).[96] Si ce n'est pas possible, il faut se rapprocher le plus possible de la capacité à révoquer les données.

[96] IA : Mobiliser la base légale de l'intérêt légitime pour développer un système d'IA. *CNIL*, 10 juin 2024. https://www.cnil.fr/fr/base-legale-interet-legitime-developpement-systeme

Une fois entraîné, un modèle ne peut pas « désapprendre » ce qu'il a appris, et c'est là que les choses se compliquent. Il faut alors se tourner vers les mesures 19, 20, 21 et 22.

2. Le RAG mis en place contient des données personnelles

   Dans certains cas, le RAG utilisé représente une large quantité de données du système d'information (pour Microsoft Copilot 0365, c'est l'intégralité des données qui est insérée dans le RAG). Dans ces cas, et dès lors qu'un système d'IA est susceptible de porter atteinte à la vie privée des personnes dont on traite les données, il est nécessaire de réaliser une analyse d'impact sur les données personnelles des personnes concernées. Ceci permet de définir, sur tout le cycle de vie de la donnée (de l'entraînement de l'IA générative à l'exploitation de la données sortante), comment elle est traitée, et sécurisée, et comment les différents principes du RGPD sont respectés.

3. Le modèle est réentraîné sur des données personnelles

   Afin d'optimiser les performances d'un modèle pour une tâche donnée, il est possible de le réentraîner sur des jeux de données spécifiques. Evidemment, si la présence de données personnelles n'est pas nécessaire pour le réentrainement, il convient de s'assurer que le jeu de données n'en comporte pas, notamment si des données issues d'interactions avec les utilisateurs y sont intégrées. Si le réentraînement doit se faire avec des types de données susceptibles de contenir des données personnelles (par exemple, des fiches de patients dans le cadre d'un *chatbot* de conseils médicaux), l'utilisation d'un jeu de données synthétiques est possible (voir mesure 20 dans la suite du document). Autrement, si l'utilisation de données personnelles est indispensable, il convient de réaliser comme dans le cas du RAG une analyse d'impact sur les données personnelles concernées, pour déterminer si leur exploitation est licite et mettre en place les mesures de conformité correspondantes le cas échéant.

Mesure n°18 : S'assurer d'avoir des fonctions mémoires hermétiques

Dans les cas où le système d'IA générative conserve en mémoire ses interactions avec les utilisateurs, s'assurer d'un cloisonnement strict pour éviter qu'un utilisateur puisse avoir accès au contenu des interactions d'un autre utilisateur avec le modèle.

En mars 2023, un bug dans ChatGPT permet à des utilisateurs de voir les titres de l'historique de conversations d'autres utilisateurs[97]. Evidemment, plus les *chatbots* auront une connaissance fine de l'utilisateur avec qui ils échangent, plus leurs réponses seront personnalisées et précises. Cependant, il est primordial de vérifier que cette fonction « mémoire » est hermétique.

Une des solutions les plus simples et efficaces aujourd'hui consiste à simplement stocker ces conversations dans une base de données SQL et de s'assurer que le contrôle d'accès sur cette base est bien implémenté.

À l'avenir, et pour aller plus loin sur le sujet, le *Federated learning* pourrait constituer une solution efficace. Il permet d'utiliser un modèle en local sur l'appareil de l'utilisateur. Dans ce cas, à la fois l'historique personnel et le modèle pourront être conservés localement, assurant une protection par défaut des données de l'utilisateur. Aujourd'hui, si certains modèles peuvent être implémentés selon les principes du *Federated Learning*, ces cas sont encore rares pour les IA génératives.

Mesure n°19 : Appliquer les principes de *differential privacy*, ou « confidentialité différentielle »

Implémenter une solution de confidentialité différentielle, permettant l'exploitation des propriétés statistiques d'une base de données, tout en protégeant les données sensibles qu'elle contient.

La *differential privacy* peut permettre de résoudre le problème de protection des données personnelles lors de l'apprentissage des systèmes d'IA générative. C'est un ensemble de techniques qui permettent l'analyse et l'entraînement d'un modèle sur une base tout en protégeant les données prises individuellement (incl. personnelles). Dit autrement, la *differential privacy* permet d'exploiter les propriétés statistiques d'une base mais empêche les données sensibles de transparaître dans le système entraîné. Par exemple, on pourra exploiter le fait qu'une base de données soit composée de 40% d'hommes et de 60% de femmes, en s'assurant de ne jamais révéler le nom d'une personne de cette base.[98]

---

[97] Kyle Barr. ChatGPT Bug Let People See Other Users' Chat History Titles. *Gizmodo*, March 21, 2023. https://gizmodo.com/openai-chatgpt-gpt4-chatbot-microsoft-1850247184

[98] Gbola Afonja, Robert Sim, Zinan Lin, Huseyin Atahan Inan, Sergey Yekhanin. The Crossroads of Innovation and Privacy: Private Synthetic Data for Generative AI. *Microsoft Research Blog*, May 29, 2024.

Sans rentrer dans les détails techniques, le concept repose essentiellement sur l'utilisation d'un bruit aléatoire finement ajusté et paramétré sur la base de données. Ce bruit permet de masquer les contributions spécifiques des individus tout en préservant les tendances ou motifs généraux présents dans les données [99].

C'est ici une solution possible au problème de rétention des données personnelles lors de l'entraînement des modèles. Si la base d'entraînement est supprimée, et que le modèle a été entraîné selon les techniques de confidentialité différentielle, alors il peut être considéré qu'il n'y a pas de rétention des données.

De plus, le principe peut s'appliquer au-delà de la protection des données personnelles. Par exemple, il est possible d'entraîner un LLM sur une base de données de transactions, et la confidentialité différentielle permettra de s'assurer que les données d'une transaction individuelle n'apparaîtront pas dans les résultats du LLM.

Il existe cependant une limite : si le bruit injecté pour permettre la confidentialité différentielle est trop fort, cela perturbe l'apprentissage. Cependant, plus il y a de données dans la base, plus il est facile de trouver un bon compromis entre performance et confidentialité.

Les applications concrètes commencent à émerger, grâce à des fournisseurs spécialisés sur le sujet à l'image de la startup Sarus[100]. La technique s'est par exemple montrée efficace dans le cas d'une application mobile permettant de discuter avec un médecin en tant que patient. L'application facilite la tâche du médecin en préremplissant les réponses, sur la base de conversations passées, mais en permettant de protéger le détail de l'identité des patients de ces conversations précédentes.

Mesure n°20 : Utiliser des données synthétiques

Générer des données synthétiques pour remplacer les données sensibles d'une base de données d'entraînement et ainsi les protéger d'une absorption par le modèle.

---

https://www.microsoft.com/en-us/research/blog/the-crossroads-of-innovation-and-privacy-private-synthetic-data-for-generative-ai/

[99] Tianqing Zhu, Dayong Ye, Wei Wang, Whanlei Zhou, Philip S. Yu. More Than Privacy: Applying Differential Privacy in Key Areas of Artificial Intelligence. *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, n°6, pp. 2824-2843. June 2022. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9158374

[100] https://www.sarus.tech/

Les données synthétiques sont des données générées artificiellement plutôt qu'issues du monde réel. En *Machine Learning*, elles sont utilisées pour entraîner des modèles lorsque les données réelles sont limitées, inaccessibles ou sensibles.

Pour les générer, il est possible d'avoir recours à des modèles statistiques (utiliser des distributions statistiques pour créer des données qui suivent les mêmes distributions que les données réelles) ou de l'apprentissage profond (avec l'utilisation de *Generative Adversarial Networks*[101] pour créer des données à partir des données existantes).

Par exemple, dans le cas où des comptes-rendus de réunion doivent être partagés avec un tiers, on précédera de la sorte : on entraîne une IA sur le contenu des comptes-rendus originaux, puis on utilise cette IA pour générer des données synthétiques sur la base de cet apprentissage. Cela permet de transmettre cette nouvelle donnée, similaire sur le fond, mais sans l'information sensible.

Mesure n°21 : Implémenter du *confidential computing*

Exploiter la technologie de « confidential computing » pour assurer le chiffrement et le déchiffrement des données exploitées par le modèle directement au niveau du matériel de la machine effectuant le traitement.

Aujourd'hui, si le chiffrement homomorphe (FHE : Fully Homomorphic Encryption) n'est pas encore complétement industrialisable pour des raisons que nous développerons, le *confidential computing* est déjà mis en œuvre et permet de très bons résultats.

Le *confidential computing* permet de chiffrer la mémoire de la machine avec une clé secrète, qui est stockée directement dans son hardware. Le processeur (Intel, ND, Puce M) déchiffre à la volée les données avant de les exécuter. L'avantage de cette technique, c'est qu'elle impacte relativement peu les performances (environ 5%). Il n'est pas nécessaire d'écrire des programmes pour exécuter la technique, et elle fonctionne avec n'importe quel programme déjà existant. Pour le traitement au niveau du processeur, ça ne fait aucune différence : la donnée est déchiffrée juste avant le traitement. En revanche, un administrateur ne verra aucune donnée en clair sur la machine : elles seront toutes chiffrées dans la mémoire.

Grâce au *confidential computing*, il est possible de déployer des modèles chiffrés sur des machines dans un cloud public par exemple. Mais il est également possible de chiffrer le RAG associé. Ainsi, des connexions sont établies depuis le poste utilisateurs

[101] Cf. Glossaire

jusque dans la mémoire chiffrée et tout est chiffré, sauf sur la machine de l'utilisateur final : tout est opaque pour l'administrateur.

Des entreprises spécialisées, comme Cosmian[102], proposent ces services. Ils permettent de mettre en place des opérations de manière sécurisée sur des cloud publics ou chez des tiers, et fournissent une couche de services complémentaires qui fonctionne malgré la couche de chiffrement (e.g. synthèse de document, traduction, vérifiabilité, traçabilité des modifications sur le hardware…).

Le risque résiduel de cette technique réside dans le fait que la clé est gravée dans le CPU, et donc reste présente sur le serveur. Il y a toujours un risque de venir casser la couche hardware et de récupérer la clé.

Mesure n°22 : Implémenter le chiffrement homomorphe

Effectuer les opérations de traitement sur des données demeurant chiffrées, en s'appuyant sur les technologies de chiffrement homomorphe.

Contrairement au *confidential computing*, le chiffrement homomorphe est une technique de cryptographie qui permet de réaliser des opérations sur des données chiffrées sans avoir à les déchiffrer au préalable. S'il y a quelques années, les équations mathématiques pour permettre ces techniques n'étaient pas résolues, aujourd'hui nous sommes plutôt aux portes de l'industrialisation. Un certain nombre de problèmes doivent encore être résolus pour permettre cette dernière :

- La performance : si le *confidential computing* présente 5% de performance en moins, le chiffrement homomorphique n'a pas encore des performances suffisantes pour les calculs informatiques classiques[103] (deux entreprises dans le monde en sont capables : Optalysis et Cornami[104]) ;
- Même avec la résolution du problème de rapidité de traitement, un autre défi réside dans l'expansion, voir l'explosion du volume de données lorsqu'on passe du calcul en clair au calcul en chiffré. La taille des données étant beaucoup plus importante, si le chiffrement homomorphe est mis en œuvre sur un téléphone portable, la consommation en termes de bande passante sera difficilement maîtrisable ;

[102] https://cosmian.com/

[103] Sidorov, Vasily, Ethan Ethan Wei Yi Fan, Wee Keong Ng. Comprehensive performance analysis of homomorphic cryptosystems for practical data processing. *arXiv preprint arXiv:2202.02960*. February 2022. https://arxiv.org/pdf/2202.02960

[104] https://optalysys.com/ https://cornami.com/

- Le chiffrement homomorphe est mono client. C'est-à-dire que l'entité qui chiffre doit être la même qui déchiffre. En d'autres termes, la clé de chiffrement doit être la même des deux côtés. Il n'est pas possible de faire du chiffrement homomorphique sur des sources de données chiffrées sous des clés différentes (par exemple, une fiche patient sous une clé, et le résultat d'un diagnostic sous une autre clé, dans la main du médecin habilité) ;
- Enfin, il n'y a pas aujourd'hui d'outillage approprié pour les développeurs. Par exemple, il n'est pas possible d'adapter un code python et de le faire fonctionner en chiffrement homomorphe. Cela nécessite des compilateurs, et c'est très compliqué à mettre en œuvre. Ainsi, c'est très problématique pour faire des choses à l'échelle.

Ainsi, le chiffrement homomorphe aujourd'hui attend la prochaine révolution de l'outillage et du hardware pour pouvoir être industrialisé. Cependant, c'est une technique d'avenir car elle permet de faire de la **preuve formelle**. Au contraire du *confidential computing*, il n'y a pas de secret qui reste sur le serveur. Le dernier risque résiduel est supprimé. L'entreprise Zama[105] travaille beaucoup sur ces problématiques, avec dans ses équipes, Pascal Pallier, un cryptologue français reconnu.

Conclusion

La mise en œuvre de toutes ou d'une partie des mesures présentées ci-dessus, permet de couvrir de manière drastique les risques liés aux données utilisées par les IA génératives. Cependant, certaines d'entre elles peuvent impacter la performance et la fiabilité des réponses. Il est indispensable de suivre un processus itératif lors de l'implémentation de ces mesures pour aboutir au bon équilibre entre sécurité et performance.

5.2.2. Se protéger de la génération de contenu protégé légalement

Si l'on a déjà évoqué la production de contenu toxique, erroné ou contenant des données sensibles, il existe aussi la possibilité de voir un système d'IA générative produire du contenu normalement protégé par la propriété intellectuelle.

Les systèmes d'IA générative ayant été entraînés sur de grandes quantités de contenus, parfois protégés par des droits de propriété intellectuelle, il est possible que les éléments en sortie du modèle soient proches de ces contenus protégés.

[105] https://www.zama.ai/

Par exemple, plusieurs cas où des systèmes d'IA générative, comme MidJourney ou GPT 3.5 ont été accusés de plagiats d'œuvres protégées par la propriété intellectuelle sont déjà survenus, occasionnant des contentieux judiciaires [106]. Des auteurs ont par ailleurs attaqué en justice OpenAI, pour des raisons similaires. [107]

Mesure n° 23 : Se protéger contractuellement des risques légaux

Prendre des mesures de protection légale pour se prémunir en cas de poursuites liées à l'utilisation de données protégées légalement pour l'entraînement d'un modèle, ou liées à la production d'un contenu protégé par des droits d'auteur.

Le problème de la génération de contenu non libre de droits par les systèmes d'IA générative n'est à l'heure actuelle pas réglé. Il n'existe pas aujourd'hui de solution permettant par exemple de filtrer les sorties d'un modèle pour bloquer tout contenu couvert par la propriété intellectuelle.

Dans le cas du recours à un modèle d'un fournisseur, avec qui il existe un contrat, il convient de s'assurer que les clauses contractuelles protègent contre ce risque. La plupart des fournisseurs (OpenAI, Microsoft) ont d'ailleurs publiquement déclaré qu'ils prendraient à leur compte tous les procès de ce type. Microsoft, a pris la décision de s'engager à défendre ses clients utilisant son service Azure OpenAI en cas d'attaque pour des questions de violation de la propriété intellectuelle. [108]

En revanche, pour les modèles récupérés en open source (Llama, Mistral), la tâche peut s'avérer plus compliquée.

---

[106] Aayush Mittal. The Plagiarism Problem: How Generative AI Models Reproduce Copyrighted Content. *Unite.AI*. January 9, 2024. https://www.unite.ai/the-plagiarism-problem-how-generative-ai-models-reproduce-copyrighted-content/

[107] Antoine Oury. Des auteurs attaquent en justice ChatGPT, accusé de violations du copyright. *Actualitté*. 30 juin 2023. https://actualitte.com/article/112438/droit-justice/des-auteurs-attaquent-en-justice-chatgpt-accuse-de-violations-du-copyright

[108] Microsoft Legal Resources. Customer Copyright Commitment Required Mitigation. *Microsoft*, May 21, 2024. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/customer-copyright-commitment

5.3. Réduire les risques liés à une mauvaise utilisation de l'IA générative

5.3.1. Garantir une connaissance suffisante pour l'utilisation des outils d'IA générative

La promesse de l'IA générative, à savoir un gain considérable en termes de productivité, ne pourra être remplie qu'à condition que les utilisateurs comprennent et maîtrisent complètement ces outils.

Et la bonne utilisation de ces systèmes suppose une connaissance élémentaire de leur fonctionnement intrinsèque, de ce qu'ils peuvent faire et ne pas faire, de la bonne manière de les utiliser et de leurs limites. Un manque de maîtrise de la part d'un utilisateur peut sans doute conduire à la production de résultats erronés et/ou peu fiables, et in fine à une frustration et un rejet des outils.

En 2023, le cas de l'avocat Steven Schwartz avait attiré l'attention du public, lorsqu'il avait produit une plaidoirie s'appuyant sur des jurisprudences n'existant pas, et qui lui avait été fournies par ChatGPT. L'avocat avait expliqué ne pas avoir été conscient que le modèle pouvait générer des jurisprudences fictives. [109]

Mesure n°24 : Sensibiliser et former les utilisateurs aux risques et limites des outils d'IA générative

Assurer la bonne maîtrise de l'IA générative par les utilisateurs en sensibilisant à ses limites et ses potentielles défaillances, et en formant les utilisateurs à optimiser leur utilisation de l'outil (e.g. ingénierie de requête ou prompt engineering) pour améliorer les performances du modèle.

Il est indispensable de former les utilisateurs aux outils d'IA générative. Cette formation doit porter sur la manière dont cette technologie fonctionne et sur ses limites. Il est notamment indispensable de former les utilisateurs à l'ingénierie de requête, qui consiste à améliorer les instructions données à un système d'IA générative pour accroître la fiabilité et la qualité des résultats fournis.

[109] Benjamin Weiser, Nate Schweber. The ChatGPT Lawyer Explains Himself. *New York Times.* June 8, 2023. https://www.nytimes.com/2023/06/08/nyregion/lawyer-chatgpt-sanctions.html

Par exemple, pour l'agent conversationnel fournissant des conseils financiers, la possibilité de fournir une bibliothèque de requêtes aux utilisateurs a ainsi été évoquée. Cela constitue une base de départ pour un néophyte. Cependant, la base de prompts génériques n'est pas suffisante, et il s'agit de faire monter en compétence les utilisateurs sur l'art du *prompting* (e.g. donner un rôle à l'IA générative, demander à l'IA de s'auto-corriger, voire ajuster son niveau de politesse[110], …).

Par ailleurs, certains messages doivent être diffusés régulièrement auprès des populations non averties, notamment sur le risque d'hallucination et le fait qu'il est indispensable de porter un regard critique sur tous les résultats proposés par un outil d'IA générative.

Enfin, en complément des mesures de sensibilisation et de formation aux outils, la collecte et l'analyse des retours des utilisateurs d'un système d'IA générative peut permettre d'identifier les potentielles difficultés dans son utilisation, et d'ajuster les messages de sensibilisation.

Mesure n°25 : Afficher un message d'avertissement contre les risques d'hallucination

Rappeler à chaque utilisation d'un système d'IA générative le risque d'hallucination et d'erreur afin d'encourager les utilisateurs à porter un regard critique sur le contenu généré et à revérifier les informations en cas de doute.

En février 2024, pour la première fois, une entreprise est tenue responsable d'informations erronées fournies par un *chatbot* à ses utilisateurs. C'est Air Canada qui a été visée par la procédure, alors que son *chatbot* avait halluciné en proposant à tort une procédure de remboursement de billet d'avion[111]. La justice a considéré notamment que le site n'avait pas suffisamment mis en avant l'avertissement qui mettait en garde les utilisateurs quant aux risques d'hallucination et les encourageait à aller consulter en complément la politique sur le site.

Ce type de mesure sera indispensable pour tout projet d'IA générative, en particulier ceux qui sont rendus disponibles au grand public. Il permettra de réinsister sur les

[110] Ziqi Yin, Hao Wang, , Kaito Horio, Daisuke Kawahara1, Satoshi Sekine. Should We Respect LLMs? A Cross-Lingual Study on the Influence of Prompt Politeness on LLM Performance. *arXiv preprint arXiv:2402.1453.* February 2024. https://arxiv.org/pdf/2402.14531

[111] Maria Yagoda. Airline held liable for its chatbot giving passenger bad advice – what this means for travellers. *BBC*. February 23, 2024. https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know

risques, éventuellement de réduire le risque juridique lié à des erreurs pour les organisations, et surtout d'insister pour que les utilisateurs aillent vérifier la pertinence de la réponse dans le corpus documentaire. Il est désormais facile de rajouter un lien vers les documents dans lesquels le *chatbot* est allé piocher sa réponse.

Mesure n°26 : Faire signer à ses employés une charte d'utilisation de l'IA

Publier et faire signer une charte d'utilisation des systèmes d'IA à tous ses employés, qui rappelle les grands principes en la matière et les règles d'or d'utilisation.

Cette charte pourra par exemple rappeler les usages autorisés et non autorisés de l'IA, la responsabilité des employés quant à l'utilisation des systèmes, l'impératif de porter un regard critique sur les réponses proposées, et plus globalement l'ensemble des bonnes pratiques et les contacts clés en cas de problème ou de question.

5.3.2. Assurer la continuité en cas d'indisponibilité des outils d'IA générative

Si l'usage de systèmes l'IA générative prend une place trop importante dans les processus de l'organisation, l'activité opérationnelle de l'organisation pourrait être affectée si la disponibilité de ces systèmes fait défaut. Le 4 juin 2024, ChatGPT devient indisponible pour quelques heures, semant la panique dans l'écosystème IA[112]. Si le système a été rapidement rétabli, il pose la question de la dépendance à ces systèmes.

Dans le cas de l'IA générative, certains cas d'usage vont permettre le remplacement d'agents humains et ce à grande échelle (par exemple pour faire du support client). L'impact opérationnel en cas d'indisponibilité du système peut être conséquent.

Mesure n°27 : Assurer la continuité d'activité

Maintenir une capacité à fonctionner sans les outils d'IA générative, quitte à ce que cela soit dans un format dégradé, pour limiter les impacts en cas de défaillance de ces derniers, et conserver des savoir-faire au sein des équipes.

Le système d'IA générative doit être également considéré comme un composant à part entière du système d'information, et être intégré aux processus de signalement des incidents. Comme pour tout autre système, il faudra définir des processus de réaction à ces incidents pour pouvoir réagir au plus vite en cas d'attaque avérée sur le système

[112] https://status.openai.com/incidents/qvp3rhvc3vwk

d'IA, et identifier clairement les modes de fonctionnement dégradés en cas d'indisponibilité partielle ou totale du système.

Aussi, il faut assurer la diversification des systèmes et des modèles. Cela permet non seulement de diversifier les risques d'erreur et de biais, mais aussi les risques en termes de disponibilité.

Dans notre exemple, cela peut passer par la création d'une *task force* d'agents de clientèle qui peut être mobilisée rapidement en cas de problème, ou le fait de réorienter temporairement les demandes des utilisateurs par voie de courriel classique.

5.4. Récapitulatif

Ce tableau récapitule l'ensemble des mesures évoquées dans notre synthèse. La complexité de mise en œuvre est évaluée en fonction de l'ampleur de l'effort nécessaire pour implémenter la mesure au sein de l'organisation, et du degré de maturité de la mesure évoquée (adoption au sein des organisations, taille du marché des solutions associées, etc…). Il indique également si la mesure doit plutôt mobiliser les équipes cybersécurité, les équipes data science, ou les deux à la fois.

Objectif de la mesure	Mesures	Complexité de mise en œuvre	Equipes à mobiliser
Réduire les risques liés au modèle			
Réduire la génération de contenu non désirable	1. Spécialiser les modèles pour leurs cas d'usage	🟢	Data science
	2. Mettre en place de la « Retrieval-Augmented Generation », ou RAG	🟢	Data science, owner de la documentation
	3. Durcir les paramètres de génération de contenu du modèle	🟢	Data science, équipe cybersécurité
	4. Filtrer les réponses non désirables	🟠🟠	Data science, cybersécurité
	5. S'assurer de la pertinence des données de réentrainement	🟠🟠	Data science

Objectif de la mesure	Mesures	Complexité de mise en œuvre	Equipes à mobiliser
	6. Mettre en place du Reinforcement Learning From Human Feedback	🟠🟠	Data science
	7. Mettre en place un dispositif de validation humaine	🟠🟠	Data science
	8. Mettre un place un système d'IA constitutionnelle	🔴🔴🔴	Data science
Se protéger des tentatives d'utilisation malicieuse	9. Maintenir ses modèles à jour	🟢	Data science, cybersécurité
	10. Offusquer les paramètres du modèles	🟢	Data science
	11. Contrôler les entrées utilisateurs	🟢	Data science, cybersécurité
	12. Transformer les prompts envoyés par les utilisateurs	🟠🟠	Data science, cybersécurité
	13. Réaliser un red-team IA	🟠🟠	Data science, cybersécurité
	14. Détecter les tentatives malicieuses	🔴🔴🔴	Data science, Cybersécurité
Réduire les risques liés aux données utilisées par le modèle			
Prévenir la diffusion de contenu sensible, confidentiel ou personnel	15. Protéger les données par défaut	🟢	Cybersécurité
	16. Respecter les politiques de gestion des identités et des accès	🟢	Cybersécurité
	17. S'assurer du respect du RGPD	🟢	Cybersécurité
	18. S'assurer d'avoir des fonctions mémoires hermétiques	🟢	Data science, cybersécurité
	19. Mettre en place la confidentialité différentielle	🟠🟠	Cybersécurité
	20. Utiliser des données synthétiques	🟠🟠	Data science
	21. Implémenter du *confidential computing*	🟠🟠	Cybersécurité

Objectif de la mesure	Mesures	Complexité de mise en œuvre	Equipes à mobiliser
	22. Implémenter du chiffrement homomorphe	🔴🔴🔴	Cybersécurité
Se prémunir contre les conséquences légales de la génération de contenu avec de la donnée protégée légalement	23. Se protéger contractuellement des risques légaux	🟠🟠	Data science
Réduire les risques liés à une mauvaise utilisation de l'IA générative par les utilisateurs			
Pallier le manque de connaissance des utilisateurs sur l'utilisation de l'IA générative	24. Sensibiliser et former les utilisateurs	🟢	Data science, cybersécurité
	25. Afficher un message d'avertissement sur les risques d'hallucination	🟢	Data science, cybersécurité
	26. Faire signer à ses employés une charte d'utilisation de l'IA	🟢	Data science, cybersécurité
Eviter les cas de dépendance trop élevée aux outils d'IA générative	27. Assurer sa continuité d'activité et la capacité à fonctionner en mode dégradé	🟠🟠	Data science, cybersécurité

HUB FRANCE IA

6. Conclusion générale

Contributeurs :

- ***Eric Savignac**, Expert, Airbus DS*

6. Conclusion générale

L'analyse des différents usages des LLM à travers les risques causés par cette technologie dans de nombreux domaines, telles que les ressources humaines, la finance ou la santé montre une situation en pleine évolution. Alors que ChatGPT a fait son entrée massive dans tous les secteurs de l'économie, gagnant des millions d'usagers en quelques mois, sans avoir d'abord fait la preuve de ses qualités ou de ses travers, les utilisateurs ont entamé un parcours d'expérimentations pour identifier les bénéfices qu'ils pouvaient en tirer ainsi que les risques encourus.

Nous apportons un début de réponse à ces interrogations légitimes, à travers les différents cas d'usage étudiés, dans le chapitre 4 de ce document, et au-delà de l'intérêt de l'utilisation des LLM, en mettant en exergue, pour chacun des domaines abordés, les risques liés au modèle, ceux qui sont liés aux données utilisées et enfin ceux provenant d'une mauvaise utilisation du LLM. Chaque risque identifié a fait l'objet d'une analyse d'impact (1-4) pour chacune des 7 catégories d'impact étudiées (financier, réputationnel, juridique ou environnemental ...). Une synthèse de mesures transverses de remédiation, permettant de limiter la probabilité d'occurrence du risque ou son impact, pour chacune des 3 grandes classes de risques (données, modèles et facteurs humains) a été présentée au chapitre 5 de ce document et constitue la boîte à outil indispensable pour la mise en place d'un LLM de manière sécurisée.

L'année 2023 a été l'année des LLM. Mais c'est l'année 2024 qui va permettre d'établir les bonnes pratiques pour le **déploiement des LLM** : on verra certainement des **LLM plus petits**, moins coûteux donc, **affinés** (*fine-tuned*) sur les données des entreprises et donc faisant moins d'erreurs, ou bien des architectures LLM faisant appel à du **RAG** pour obtenir des informations de qualité et pertinentes, au regard du prompt de l'utilisateur, directement issues du corpus informationnel de l'entreprise. Il est fortement probable que nous serons aussi bientôt amenés à aborder le concept du professionnel augmenté, et ce peu importe son domaine d'emploi, car il semblerait que cela soit une trajectoire qui se dessine pour l'IA générative.

7. Glossaire

ANSSI	L'agence nationale de la sécurité des systèmes d'information est l'autorité nationale en matière de cybersécurité. Elle est placée sous l'autorité du Premier ministre et rattachée au secrétaire général de la défense et de la sécurité nationale. https://cyber.gouv.fr/decouvrir-lanssi
ChatGPT	*Chatbot* développé par OpenAI, fondé sur un grand modèle de langage.
CNIL	Commission Nationale de l'Informatique et des Libertés.
CPU	*Central Processing Unit* : microprocesseur principal d'un ordinateur.
Deep Learning	Sous-ensemble du *Machine Learning* fondé sur l'utilisation de réseaux de neurones dits profonds, c'est-à-dire utilisant de nombreuses couches de neurones.
Fine-tuning	Le *fine-tuning* d'une IA générative pré-entraînée consiste à lui faire exécuter un entraînement supplémentaire sur des données labellisées spécifiques d'une tâche ou d'un domaine particulier afin d'améliorer sa performance.
GAN	*Generative Adversarial Networks* : architecture de Deep Learning dans laquelle deux réseaux neuronaux sont entraînés et mis en compétition.
GED	Gestion Electronique de Documents : solution logicielle visant à organiser et gérer des informations sous forme de documents électroniques.
GPT	*Generative Pretrained Transformer* : c'est une famille de *Large Language Models* développée par OpenAI.
Guardrails	Ce sont des protections qui permettent de contrôler les entrées et sorties d'une IA générative afin de réduire les risques liés à son utilisation.

Hallucination	Information fausse, inexacte ou incohérente créée par une IA générative.
IA générative	Sous-ensemble du *Deep Learning*, visant à produire du contenu, que ce soit du texte, une image, de l'audio ou une vidéo, à partir de données en entrée (on parle alors de *prompt*), elles-mêmes du texte, une image, de l'audio ou une vidéo.
Large Language Model (LLM)	Un type d'IA générative capable de générer et d'analyser du texte (par exemple : langage naturel, langage de programmation …)
LOD	*Line Of Defense* : niveau de contrôle composant le contrôle interne d'un établissement.
Machine Learning	Apprentissage automatique à partir d'un ensemble de données.
Master prompt	C'est un prompt de haute qualité et bien informé, conçu avec précision, contexte et clarté qui guide le modèle d'IA et influence de manière significative la qualité de la sortie de l'IA permettant de générer une réponse spécifique très pertinente.
Méthode EBIOS Risk Manager	Méthode d'analyse de risque française de référence, permettant aux organisations de réaliser une appréciation et un traitement des risques. https://cyber.gouv.fr/la-methode-ebios-risk-manager
Model Owner	Acteur clé qui a la responsabilité de s'assurer que le développement du modèle d'IA, son implémentation, son usage, et son suivi dans le temps soient conformes avec les politiques et procédures de l'organisation.
NIST	*National Institute of Standards and Technology*: agence du département du Commerce des États-Unis dont la mission est de promouvoir l'économie en développant des technologies, la métrologie et les normes pour l'industrie. https://www.nist.gov/

Prompt	Le *prompt* est l'instruction ou la requête en langage naturel fournie à l'IA générative dans le but d'obtenir une réponse (un contenu).
RAG	Retrieval *Augmented Generation* : génération augmentée via la récupération d'informations d'une base de connaissances qui n'a pas été utilisée lors de l'entraînement de l'IA générative.
RIA	Règlement sur l'Intelligence Artificielle – *AI Act* en anglais. Applicable sur le marché de l'Union Européenne.
RGPD	Règlement Général sur la Protection des Données.
Role Based Access Control	Modèle de contrôle d'accès à un système d'information dans lequel l'accès à une ressource est basé sur le rôle de l'utilisateur concerné.
SLA	*Service Level Agreement* : contrat de service entre un prestataire informatique et un client.
SSI	Sécurité des Systèmes d'Information, voir la norme internationale ISO/CEI 27001 ainsi que l'autorité nationale de sécurité des systèmes d'information (ANSSI).
Température	La température dans le cadre d'une IA générative est un paramètre du modèle permettant de gérer le caractère aléatoire d'un texte généré (par exemple). La température varie généralement entre 0 et 1 ; une valeur proche de zéro générera un texte quasi identique à chaque génération, alors qu'une valeur proche de 1 générera un texte avec plus de créativité ou variabilité.
Token	Sous-ensemble d'un mot constituant une unité de traitement par un *Large Language Model*.

8. Remerciements

Le Hub France IA remercie l'ensemble des participants au groupe de travail IAG, et tout particulièrement les contributeurs de ce livrable.

La pilote :

- **Imen Fourati**, Expert lead, Risque de modèle, Société Générale

Les contributeurs :

- **Thomas Argheria,** Manager – Wavestone
- **Gérôme Billois,** Partner, Wavestone
- **Benjamin Bosch**, Manager - Model risk Management – Data Science, Société Générale
- **Anis Bousbih,** Cofondateur - Aicademia
- **Kati Bremme,** Head of Innovation – France Télévisions
- **Thibault Cattelani,** Cofondateur – Emocio.hr
- **Martin D'Acremont,** Consultant – Wavestone
- **Wissem Fathallah,** Cofondateur & Chief Product Officer - Sifflet
- **Thomas Gouritin,** Consultant - Tomg Conseil
- **Jeanine Harb,** CTO – Beink Dream
- **Vanessa Hespel**
- **Belkacem Laïmouche,** Chargé de mission innovation – Direction Générale de l'Aviation Civile
- **Pascal Lainé,** CTO - Talkr
- **Jacques Mojsilovic,** CMO - Numalis
- **Cyril Nicolotto,** Chef de projets – Hub France IA
- **Kevin Paci,** Responsable des services informatiques – Mediaco Vrac
- **Nicolas Pellissier,** Cofondateur – Klark
- **Alexandre Pouymayon,** Consultant, Wavestone
- **Constance Relmy,** Etudiante – Université Paris 1 Panthéon Sorbonne
- **Laurence Relmy**
- **Eric Savignac,** Expert – Airbus DS
- **Kevin Soler,** CEO - Virteem
- **Yael Suissa,** CEO & Cofondateur - MAP-Monitoring And Protection

Les relecteurs :

- **Fatiha Gas,** Directrice Innovation Data/IA & Programme IA Générative Groupe – La Poste Groupe
- **Didier Girard,** co-CEO – SFEIR
- **Etienne Guibout,** Group Data Office – Société Générale
- **James Rebours,** Cofondateur - Klark
- **Françoise Soulié-Fogelman,** Conseiller Scientifique – Hub France IA
- **Bastien Zimmermann,** Ingénieur R&D – Craft AI

La touche finale :

- **Mélanie Arnould,** Responsable des opérations – Hub France IA
- **Louise Paurise,** Stagiaire – Hub France IA

Mention spéciale :

- **Jean de Bodinat,** Fondateur – Rakam AI ; pour son implication dans le pilotage des premières semaines (les plus difficiles !) du groupe de travail

LES RISQUES DE L'IA GENERATIVE

Juillet 2024

HUB FRANCE IA