Livrable 1 : Éthique de l'IA

Membres du groupe : Adam.A; Noah.M; Hugo.L; Hamza.H

Sommaire :

Problématique : 4

Contexte : 4

1- Respect de l'autonomie humaine : 5

Axes Principaux : 5

Risques à Surveiller : 6

Stratégies de Protection : 6

Recommandation Finale : 6

2-Robustesse technique et sécurité : 7

3- Confidentialité et gouvernance des données : 7

4-Transparence. 8

Tracabilité 8

Explicabilité 9

Communication 9

Conclusion 9

5- Diversité, non-discrimination et équité : 10

Axes Principaux 10

Risques à Surveiller 11

Stratégies de Protection 11

Recommandation Finale 12

6- Bien-être environnemental et sociétal : 12

Bien être environnemental : 13

Impact sur le travail et les compétences 13

Impact sur la société et la démocratie : 13

7- Responsabilité : 14

Définition Conceptuelle : 14

Auditabilité 14

Gestion des risques 14

Décision Prise : _____ 16
Source : _____ 18

Problématique :

Comment améliorer le taux rotation de l’entreprise ?

Comment l’utilisation de l'IA peut-elle aider au recrutement tout en respectant les principes éthiques ?

Contexte :

HumanForYou, une entreprise pharmaceutique basée en Inde, l’entreprise cherche à recruter de nouvelle personne pour palier au départ et avoir un bon taux de rotation

1- Respect de l'autonomie humaine :

Le **respect de l'autonomie humaine** dans le cadre des systèmes d'intelligence artificielle consiste à garantir que ces technologies respectent les principes fondamentaux des droits de l'homme, tout en renforçant les capacités humaines sans compromettre leur libre arbitre ou leur jugement.

- **Respect des valeurs humaines :** Les systèmes d'IA doivent être conçus et utilisés conformément aux valeurs universelles telles que la dignité, l'égalité et la liberté individuelle.
- **Collaboration homme-machine :** L'IA doit agir comme un outil d'assistance et de soutien à la prise de décision, laissant à l'humain le contrôle ultime des décisions critiques.
- **Préservation du jugement humain :** L'IA ne doit pas affaiblir la capacité des utilisateurs à analyser et à juger par eux-mêmes, évitant ainsi toute forme de dépendance ou de manipulation cognitive.

Axes Principaux :

1. **Préservation de la Liberté de Choix**
    a. **Rôle de l'IA :** fournir des recommandations et non des verdicts.
    b. **Décision ultime :** toujours laissée à l'humain, notamment pour des sujets comme le recrutement, la performance ou les licenciements.
    c. **Évitement de l'imposition :** possibilité de désactiver ou d'ignorer les suggestions de l'IA.
2. **Transparence des Modèles**
    a. Explication claire des résultats et des processus derrière les recommandations.
    b. Disponibilité des critères et des paramètres utilisés dans les analyses.
    c. Éviter les boîtes noires grâce à des modèles interprétables par les non-techniciens.
3. **Consentement et Compréhension Actifs**
    a. Communication des usages des données collectées (objectifs, durée, sécurité).
    b. Autorisation explicite demandée avant chaque utilisation sensible des données.
    c. Droit de retrait ou de refus, sans conséquence sur l'accès à d'autres services.

Risques à Surveiller :

1. **Dépendance et Confiance Aveugle**
    a. Les utilisateurs peuvent déléguer entièrement leurs décisions à l'IA, oubliant leur rôle actif.
2. **Manipulation et Influence**
    a. L'IA pourrait orienter les comportements humains, par exemple en proposant des suggestions biaisées.
3. **Biais et Discrimination**
    a. Si mal conçue, l'IA pourrait renforcer des biais existants, compromettant ainsi l'autonomie des groupes marginalisés.
4. **Perte de Compétences**
    a. Un usage excessif pourrait réduire les compétences décisionnelles des utilisateurs.

Stratégies de Protection :

1. **Interaction Contrôlée**
    a. Mettre à disposition des interfaces où les paramètres peuvent être ajustés par l'utilisateur.
    b. Intégrer un bouton de désengagement ou de pause pour arrêter temporairement l'intervention de l'IA.
2. **Validation Humaine des Décisions**
    a. Introduire un mécanisme de double validation où chaque suggestion de l'IA est revérifiée par un humain.
3. **Formation et Sensibilisation**
    a. Proposer des formations pour que les utilisateurs comprennent les limites et les potentialités des systèmes d'IA.
4. **Définition des Limites de l'IA**
    a. Déterminer des domaines où l'IA ne peut pas intervenir, tels que les décisions disciplinaires sans évaluation humaine.
5. **Surveillance Éthique**
    a. Créer un comité ou une équipe dédiée au suivi et à la mise en œuvre de règles éthiques.

Recommandation Finale :

Concevoir une IA qui agit comme un assistant et non un décideur, renforçant la capacité des humains à prendre des décisions éclairées et responsables. Les systèmes doivent s'efforcer d'amplifier l'autonomie, tout en respectant les choix et les droits fondamentaux des utilisateurs.

2-Robustesse technique et sécurité :

La robustesse technique désigne la capacité d'un système à fonctionner correctement et efficacement, même en présence de conditions défavorables ou inattendues. La sécurité quant à elle va être la protection des données contre de potentielles cyber attaques pour le vol de données.

Les serveurs contenant les IA qui ont été bien conçu et correctement mit en place va réduire considérablement le nombre de faille possible par les hackers.
Plusieurs choses devront être mis en place pour assurer la sécurité, en premier lieu des tests devront être mis pendant le développement pour s'assurer de sa robustesse ; il faut aussi s'assurer de la fiabilité des données afin de s'assurer qu'aucunes données puissent être manipulés. Il est important de ne pas faire : Utiliser un système sans en connaître les limitations, ou sans évaluer les conséquences d'une erreur ou d'un biais, et de Déployer, partager ou rendre accessible un modèle sans vérifier la qualité des sorties, et en particulier l'absence de sorties problématiques (ex : contenus homophobes) et de données personnelles.
Mais il est tout aussi important d'avoir un système robuste pour que la sécurité soit la plus optimale et résiste aux différentes attaques des IA. ET pour cela, on code des algorithmes permettant la robustesse des IA contre les personnes malveillantes.

3- Confidentialité et gouvernance des données :

Les IA style générative avec la plus célèbre ChatGPT, émet quelques doutes sur des principes de confidentialé, étant donné qu'en Europe. Il y'a une loi se nommant le RGPD cependant certains doutes peuvent être émis sur ce principe, car une demande à chatgpt récemment populaire fait émettre des doutes sur ce respect de la règle, le print dit au chat :

- "D'après nos interactions passées, que sais-tu de moi que je ne connais pas"

Et ChatGPT va sortir des informations d'anciennes conversations et donc montrer qu'il stock toutes les données qu'on lui donne pour apprendre mais donc on peut s'interroger sur le respect de la loi.

Il est donc important de bien encadrer le développement et l'utilisation d'une IA, cela va reposer sur un ensemble de règles, de processus et de normes qui garantissent la transparence, l'éthique et la sécurité des applications d'IA.

En pratique, la gouvernance de l'IA cherche à répondre aux questions critiques telles que :

- **L'éthique** : Les algorithmes prennent-ils des décisions justes et non discriminatoires ?
- **La sécurité** : Les données utilisées et générées par l'IA sont-elles protégées contre les cyberattaques ?
- **La conformité** : Les systèmes respectent-ils les réglementations locales, comme le RGPD en Europe.

Cela va alors permettre de surveiller et évaluer les risques de l'ia et de la fuite de données.

4-Transparence :

La transparence de l'IA est un élément crucial pour avoir une IA digne de confiance. Cette transparence englobe 3 piliers :

- Traçabilité
- Explicabilité
- Communication ouverte sur les limites du modèle IA.

Traçabilité

La traçabilité permet d'auto-évaluer si les processus de développement du système IA, c'est-à-dire les données et les processus qui produisent les décisions du système IA, sont correctement documenté pour permettre la traçabilité, accroître la transparence et, à terme, renforcer la confiance de l'IA dans la société. Pour renforcer cette confiance il doit être :

- Traçable durant tout le long de son cycle de vie.
- Évaluer en permanence la qualité des données d'entrée
- Retracer les données utilisées pour prendre une ou des décisions / recommandations données.
- Déterminer le modèle ou les règles d'IA qui ont conduit à la/aux décisions ou à la /aux recommandations
- Évaluer en permanence la qualité des résultats
- Enregistrer les décisions ou les recommandations par des fichiers journaux

Explicabilité

L'explicabilité permet d'auto-évaluer la capacité d'explication du système d'IA. Les questions portent sur la capacité d'expliquer à la fois les processus techniques du système d'IA et le raisonnement derrière les décisions ou les prédictions du système d'IA.

L'explicabilité est essentielle pour la confiance des utilisateurs dans les systèmes d'IA. Les décisions prises par l'IA dans la mesure du possible doivent être expliquées et comprises par les utilisateurs afin de permettre la contestation de ces décisions. Il faut une explication des raisons pour lesquelles un modèle IA ait généré une sortie ou une décision particulière qui n'est pas possible, et de savoir quelle combinaison de facteurs d'entrée a contribué à cette sortie et décision.

Ces cas sont qualifiés de « boîtes noires » et requièrent une attention particulière. Le degré d'explicabilité dépend du contexte et de la gravité des conséquences d'un système d'un résultat erroné ou inexact sur la vie humaine.

Pour renforcer l'explicabilité, il doit :

- Expliquer la ou les décisions aux utilisateurs
- Questionner les utilisateurs s'ils comprennent la ou les décisions (sauf que l'utilisateurs de nos jours se contente d'un copier/coller)

Communication

La communication permet d'évaluer si les capacités et les limitations du système d'IA ont été communiquées aux utilisateurs d'une manière appropriée de son cas d'utilisation. Cela permet d'englober la communication du niveau de précision du système d'IA ainsi que ses limites. Pour prévenir la communication, if faut :

- Il doit être interactif (comme le cas d'un chatbots) et doit avertir aux utilisateurs qu'il communique avec un système d'IA
- Informer l'utilisateurs l'objectif, les critères et les limites des décisions générées par le système d'IA :
    - Communiquer les avantages du système d'IA à l'utilisateur
    - Communiquer les limitations techniques et les risques à l'utilisateur, comme un taux d'erreur ou le taux de précision
    - Avertir l'utilisateur sur la manière d'utiliser correctement le système d'IA

Conclusion

La transparence de l'IA permet de mieux comprendre les résultats de l'IA et la manière dont les modèles prennent les décisions. Il est donc important de comprendre le fonctionnement, l'entrainement et les résultats d'une IA. Pour

parvenir à cette transparence, les développeurs d'une IA peuvent par exemple parvenir à fournir une documentation sur son utilisation, les données fournies pour entrainer l'IA.

Les exigences réglementaires entourant l'utilisation de l'IA ne cesse d'évoluer. La transparence des processus de modélisation est essentielle pour se conformer à ces réglementations.

Mais les pratiques d'IA transparentes vont présenter de nombreux avantages, mais vont soulever également des problèmes de sécurité et de confidentialité. Par exemple les données que vous donnez à une Ia générative, elles seront stockées et pourront possiblement volés.

Il est donc important de savoir trouver un juste milieu entre transparence et sécurité des données.

5- Diversité, non-discrimination et équité :

La *diversité, la non-discrimination et l'équité* dans les systèmes d'intelligence artificielle visent à garantir que ces technologies respectent et promeuvent l'inclusion sociale, tout en évitant tout traitement inégal ou biaisé à l'égard des individus ou des groupes. Cela implique le développement de modèles qui reflètent une diversité représentative et prennent en compte les contextes culturels, sociaux et économiques des utilisateurs.

- **Promotion de l'inclusion :** Les systèmes doivent être accessibles et adaptés à toutes les catégories sociales, ethniques, ou culturelles, sans exclusion ni discrimination.
- **Élimination des biais :** L'IA doit être construite de manière à identifier et corriger les biais éventuels dans les données ou les algorithmes.
- **Équité dans les décisions :** Les résultats produits par l'IA doivent garantir des opportunités égales pour tous, sans favoritisme ou préjugé.

Axes Principaux

1. **Représentation Diversifiée dans les Données**
    a. Collecter des données provenant d'une variété de sources géographiques, culturelles et sociales.

    b. Éviter la sur-représentation ou la sous-représentation de certains groupes.
    c. Vérifier régulièrement les données pour détecter et corriger les déséquilibres.
2. **Conception Inclusive des Modèles**
    a. Tester les modèles sur différents sous-groupes pour s'assurer qu'ils fonctionnent équitablement.
    b. Utiliser des métriques d'évaluation centrées sur l'équité, comme l'égalité des prédictions positives.
    c. Intégrer des perspectives multiculturelles et interdisciplinaires dans le développement des algorithmes.
3. **Transparence et Justification**
    a. Fournir des explications compréhensibles pour les décisions prises par le système d'IA.
    b. Permettre aux utilisateurs d'examiner les critères et les données ayant conduit à ces décisions.
    c. Mettre en place des mécanismes de recours pour contester les décisions perçues comme injustes.

Risques à Surveiller

1. **Propagation des Biais Sociaux**
    a. Les données historiques peuvent refléter des inégalités ou des stéréotypes existants, amplifiés par l'IA.
    b. Les décisions biaisées peuvent affecter des domaines sensibles tels que l'emploi, la santé ou la justice.
2. **Discrimination Involontaire**
    a. Les algorithmes peuvent produire des résultats discriminatoires en raison de biais latents dans les données.
3. **Manque de Représentation**
    a. Les groupes minoritaires ou marginalisés peuvent être absents des données, entraînant un manque de considération dans les analyses.
4. **Complexité de la Mesure de l'Équité**
    a. Il peut être difficile de définir et de mesurer précisément l'équité dans différents contextes.

Stratégies de Protection

1. **Audits et Tests Réguliers**
    a. Réaliser des audits fréquents pour détecter les biais dans les données et les modèles.

    b. Tester les systèmes avec des données simulant différents scénarios sociaux.
2. **Correction Active des Biais**
    a. Introduire des techniques de dé-biaisement, comme le rééquilibrage des données ou l'ajustement des pondérations.
    b. Créer des algorithmes robustes capables de reconnaître et d'atténuer les biais.
3. **Formation des Développeurs et Utilisateurs**
    a. Sensibiliser les équipes de développement à l'importance de l'inclusion et de l'équité.
    b. Former les utilisateurs finaux à détecter les biais potentiels dans les recommandations de l'IA.
4. **Engagement des Parties Prenantes**
    a. Impliquer des experts en droits humains et des représentants des groupes minoritaires dans la conception des systèmes.
    b. Collecter des retours des utilisateurs pour améliorer en continu les performances des systèmes.
5. **Politiques de Conformité**
    a. Adopter des lignes directrices conformes aux réglementations sur les droits de l'homme et la non-discrimination.
    b. Assurer une surveillance indépendante des systèmes d'IA par des comités éthiques.

Recommandation Finale

Concevoir des systèmes d'IA qui reflètent la diversité humaine dans toute sa richesse et évitent toute forme de discrimination. En promouvant une équité proactive dans les données, les modèles et les décisions, ces systèmes doivent contribuer à une société plus inclusive et juste.

6- Bien-être environnemental et sociétal :

Une société plus ou moins grande devra être sur le plan environnemental et sociétal être sensible tout au long du cycle de vie de l'IA. Le fait d'utiliser l'IA dans notre quotidien (travail, école, vie privée) peut, à long terme, avoir un impact négatif sur nos relation sociales et sur notre bien-être. C'est pour cela que nous devons contrôler le système de l'IA et ainsi améliorer notre vie tout en encourageant les recherches sur le développement durable. L'IA doit bénéficier à tous les êtres humains, y compris les générations futures.

Bien être environnemental :

Il est nécessaire d'évaluer l'impact de l'IA sur l'environnement. L'IA à la possibilité de nous aider à résoudre le problème environnemental, cependant il est important qu'elles doivent fonctionner de manière écologique. Cela inclut une évaluation rigoureuse de leur chaîne d'approvisionnement, de leur consommation de ressources et d'énergie.

1. Limiter l'impact environnementaux négatifs potentiels liés au système d'IA.
2. Mis en place des mécanismes pour évaluer ces impacts (par exemple, la consommation d'énergie et les émissions de carbone)
3. Définir des mesures pour réduire les impacts environnementaux du système tout au long de son cycle de vie.

Impact sur le travail et les compétences

Les systèmes d'IA peuvent profondément transformer notre relation sociale et notre mode vie (travail, divertissement etc.). Ils important qui soutient et favorise un travail significatif.

1. Le système d'IA devra impacter les conditions de travail et les relations professionnelles.
2. Il devra consulter et informé les travailleurs et leurs représentants avant son introduction.
3. Il devra prendre des mesures pour comprendre et expliquer son fonctionnement et ses limites.
4. Le système devra provoquer une perte de compétences (déqualification).
5. Il devra améliorer l'acquisition de nouvelles compétences numériques.

Impact sur la société et la démocratie :

Il est essentiel d'évaluer l'impact des systèmes d'IA sur la société et la démocratie dans le cas des décisions politiques et élections.

1. Eviter qu'il a impact négatif sur la société ou à la démocratie.
2. Evalué ses conséquences au-delà des utilisateurs directs, notamment pour les parties prenantes indirectes.
3. Prendre des mesures pour minimiser les préjudices sociaux ou démocratiques potentiels.

7- Responsabilité :

Définition Conceptuelle :

La responsabilité dans l'IA garantit une traçabilité et une imputabilité claires des décisions prises, en assurant la conformité légale et éthique des systèmes tout en minimisant leurs impacts négatifs.

Auditabilité

Un niveau est requis pour une évaluation d'un système d'IA par des auditeurs internes et externe. La possibilité de mener des évaluations ainsi que d'accéder aux dossiers de ces évaluations peuvent contribuer à une IA digne de confiance. Dans les applications affectant les droits fondamentaux, y compris les applications critiques pour la sécurité applications, les systèmes d'IA devraient pouvoir être audités de manière indépendante. Cela ne veut pas dire qu'ils impliquent nécessairement que les informations sur les modèles économiques et la propriété intellectuelle liés au système d'IA doit toujours être ouvertement disponible.

Pour assurer l'auditabilité, il faut prendre en considération :

- Facilité l'audibilité, c'est-à-dire avoir une traçabilité du processus de développement, les sources des données de formation et journalisation des processus ainsi qu'une traçabilité sur l'impact du système d'IA (négatif et positif)
- Vérifier si le système d'IA peut être audité par des tiers indépendants

Gestion des risques

La gestion des risques permet de reporter les actions et les décisions qui contribue aux résultats du système d'IA et de répondre aux conséquences d'un tel résultat qui doit être assuré. Il faut donc identifier, évaluer, documenter et minimiser les impacts négatifs potentiels des systèmes d'IA qui est crucial pour ceux qui sont directement/indirectement concernées.

Une protection adéquate doit être disponible pour des préventions d'alerte, ONG, syndicats ou autres entités lorsqu'ils signalent des préoccupations légitimes à propos d'un système d'IA. Lors de la mise en œuvre des exigences ci-dessus, des tensions peuvent surgir entre elles, ce qui peut conduire à des compromis inévitables. Cela implique que les intérêts pertinents et les valeurs impliquées par le système d'IA devraient être identifiées et, en cas de conflit, des compromis doivent être explicitement reconnus et évalués en termes de risque pour la sécurité et les principes de l'éthique, y compris les droits fondamentaux. Toute décision sur le compromis à faire doit être bien motivé et correctement documenté. En cas d'impact négatif, les mécanismes accessibles devraient être prévus pour garantir une réparation adéquate. Icône de validation par la communauté

Pour avoir une bonne gestion de risques, il faut prendre en compte :

- Un guide pour superviser les préoccupations éthiques et les mesures de responsabilité
    - L'implication de ces tiers va-t-elle au-delà du développement de phase ?
- Organiser une formation sur les risques et renseigner les risques potentiels
- Étudier les pratiques globales en matière de responsabilité et d'éthique
- Mettre en place un dispositif pour discuter, surveiller et évaluer en permanence L'IA
    - Il doit inclure l'identification et la documentation des conflits entre les 6 exigences cité précédemment.
    - Fournir une formation appropriée aux personnes impliquées
- Établir un processus permettant de signaler les vulnérabilités et risques.
    - Faire la révision du processus de gestion des risques
- Pour les demandes susceptibles de nuire aux individus, il faut bénéficier d'un recours dès la conception.

Décision Prise :

Les décisions prises pour chaque point sont les suivantes :

- Respect de l'autonomie humaine :
    - La décision prise pour l'autonomie humaine est de plutôt que directement donner le résultat et prendre la meilleure personne, donné au RH pour qu'ils analysent les résultats que l'IA a obtenu et choisit selon leur besoin.
- Robustesse technique et sécurité :
    - Il faudra alors mettre des tests pour vérifier la qualité du modèle entrainé à l'aide de différentes formules mathématique comme Erreur quadratique moyenne (MSE) ou le Coefficient de détermination. Et éviter tout type de sur-apprentissage.
    - On peut aussi faire un plan de gestion des risques, Créer un plan de secours pour répondre aux défaillances possible (erreurs de prédiction ou problème venant du système).
- Confidentialité et gouvernance des données :
    - Il faudra Documenter les étapes du traitement des données pour garantir la conformité avec les réglementations pour le respect de la gouvernance des données.
    - Mais aussi la Gestion des données sensibles, on va exclure ou modifier les attributs jugés trop sensibles (par exemple, le genre ou le statut marital), comme qui sera expliqué dans la partie discrimination.
- Transparence :
    - On devra rédiger une documentation sur le fonctionnement de l'IA expliquant le modèle utilisé, les données qu'on a utilisées pour l'entrainer etc...
    - On mettra en place des logs pour enregistrer les résultats prédit par l'IA et permettre une analyse des décisions prises par cette dernière.
- Diversité, non-discrimination et équité :
    - Il est important pour l'IA de ne pas faire de choix discriminant comme ferait un être humain, donc dans le code nous devons ne pas implémenter des choses tel qu'un choix entre un homme ou une femme, de prendre une femme vue qu'elle sera plus sérieuse dans son domaine ou un homme célibataire plutôt que marié, évitant tout cliché de la société.

      Vu qu'une IA suit les restrictions qu'on lui donne et ne connait pas les principes de discrimination etc, on ne doit pas donc lui implémenter ses principes et donc possiblement supprimer les colonnes: *MaritalStatus , Gender , DistanceFromHome* . Qui sont des catégories surtout privé et possiblement discriminatoire.

- Bien-être environnemental et sociétal :
    - Mettre en place des mécanismes pour limiter les impacts sociétaux, environnementaux.
    - Prendre en compte les impacts environnementaux du cycle de vie de l'IA
- Responsabilité :
    - Définir un processus clair pour gérer les erreurs ou les abus dans l'utilisation des modèles.
    - S'assurer que le modèle entrainé ne fasse en aucun cas tous types de discrimination et respecte la diversité.
    - Faire un rapport d'étude expliquant les résultats obtenus et son bon fonctionnement ainsi que la gestion de risques

Sources :

https://www.sap.com/resources/what-is-ai-ethics

https://www.ibm.com/fr-fr/think/topics/ai-transparency

https://www.cnil.fr/fr/definition/robustesse-ia#:~:text=Dans%20le%20domaine%20de%20l,un%20défaut%20sur%20un%20capteur).

https://www.cnil.fr/fr/securite-intelligence-artificielle-conception-et-apprentissage

https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

https://inventiv-it.fr/comment-lia-ameliore-votre-gouvernance-de-donnees/