



Les risques de l'intelligence artificielle

Sorbonne Université - Faculté de Santé

1.4.1 : connaître les grands enjeux liés à l'intelligence artificielle, aux algorithmes, aux biais et aux systèmes d'aide à la décision ainsi que les principes éthiques associés aux traitements des données de santé

Licence CC BY-NC-ND 4.0

<https://creativecommons.org/licenses/by-nc-nd/4.0/>



03/04/2026

Biais, éthique, recommandations pour une éthique by design, guide d'implémentation d'un SIA en santé éthique

Les risques de l'intelligence artificielle

Pr Brigitte SEROUSSI

Sorbonne Université, Faculté de Santé

2025 - 2026



PLAN



1 Les risques de l'IA



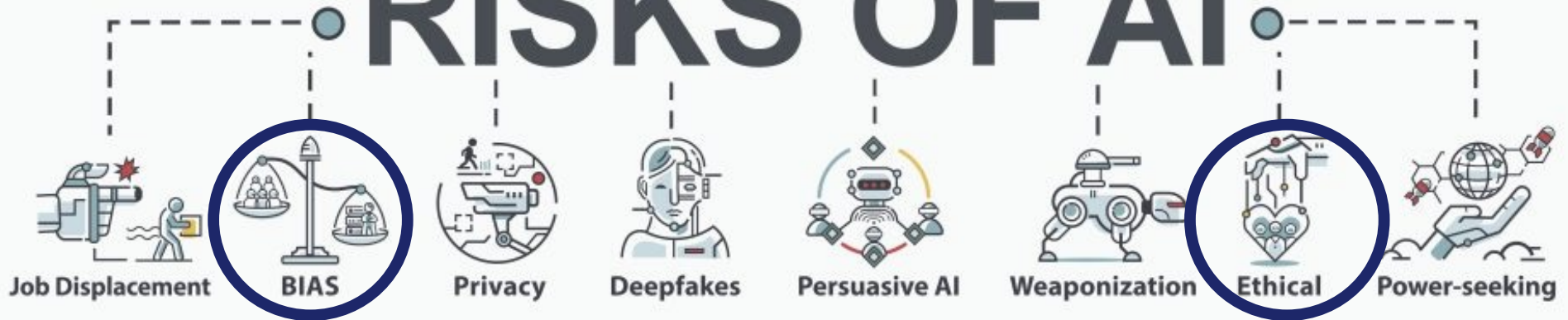
2 Les recommandations pour une éthique by design



3 Le guide d'implémentation d'un SIA en santé éthique



RISKS OF AI



Recommandations Ethique by design pour les SIA en santé (avril 2022)

La DNS publie ses recommandations pour une éthique de l'IA « by design »

ACTUA, 13/07/2022

Partagé par : Beesens TEAM



haas BIENVENUE SUR LE BLOG DU CABINET HAAS AVOCATS

dans le secteur de la e-santé. Enfin, tracer et documenter cette procédure permettra non seulement d'apporter la preuve des efforts déployés pour assurer un traitement éthique et sûr des données de santé, mais également d'améliorer la solution.



Les enjeux d'une éthique « by design » en matière d'IA

L'élaboration d'une solution d'intelligence artificielle éthique « by design » implique qu'au stade de la conception de l'algorithme la question de la transparence de la solution soit abordée.

Implication du conseil scientifique, technique et éthique et des utilisateurs dans le développement et le design de la solution d'IA

Gestion des risques

Cadrage

Définition de la finalité de la solution d'IA

Validation éthique de la finalité

Caractérisation des principes de gouvernance et de responsabilité

Collecte des données

Mise en œuvre des mesures pour assurer :

- Le consentement éclairé des patients à la réutilisation de leurs données au-delà de la finalité première du recueil (conformité RGPD, article 5 et 7)
- La proportionnalité des données collectées par rapport à la finalité du traitement servant l'élaboration de la solution d'IA (conformité RGPD, article 5, e)
- La non réidentification directe des données (agrégées, pseudonymisées, anonymisées) (conformité RGPD, article 32)
- La qualité des données (lutte contre les biais cognitifs)
- La représentativité de la population d'analyse / population cible / prévention des discriminations (lutte contre les biais de sélection)

Mise en œuvre des mesures de sécurité visant à assurer :

- Le transfert sécurisé des données (source unique, sources multiples, challenge, intégrité)
- La qualité de l'hébergement des données, serveurs localisés en France / Europe (Cloud privé/public, centralisé/distribué) (conformité HDS)
- La cyber-sécurité à l'état de l'art

Mise en œuvre de mesures pour garantir la non-réutilisation non éthique des données (par ex. en cas de fusion de la société / modifications législatives (pouvant aller jusqu'à la destruction automatique des données))

Pré-traitement des données

Mise en œuvre des mesures de :

- Traitement des données manquantes (réduction des biais)
- Rééquilibrage des populations minoritaires (réduction des biais)
- Séparation des données (deux jeux totalement distincts, un échantillon pour l'apprentissage et un échantillon pour l'évaluation) et représentativité des deux jeux de données par rapport à la population cible et la finalité du traitement)

Construction de l'algorithme

Choix de l'algorithme d'apprentissage en adéquation avec la finalité

Définition :

- De la politique qualité de l'algorithme
- Des mesures de transparence mises en œuvre
- De la politique de traçabilité de la démarche de construction de l'algorithme
- De la politique d'explicitabilité des résultats explicables, du processus d'auditabilité des résultats non explicables

Implémentation des fonctionnalités et mécanismes visant à assurer :

- L'identification et élimination des biais
- La correction des erreurs
- La traçabilité des traitements (rendre les codes sources publics avec protection par dépôt (AFP))
- L'adaptabilité
- L'intégration des évolutions réglementaires
- L'intégration des avancées médicales
- La maintenance et le versioning

Définition des indicateurs de dérive du système

Évaluation de l'algorithme et préparation de la mise en production

Mise en œuvre des principes d'évaluation externe :

- Technique (bugs), clinique (gold standard, score de précision)
- De l'utilisabilité (PS, patients, usagers)
- De la non-discrimination/équité
- De la robustesse/reproductibilité

Mise en œuvre des procédures en cas de cyber-attaques (analyse d'impact sur la sécurité du système d'IA)

Mise en œuvre des mesures pour assurer l'information (juste et équilibré) des utilisateurs (PS, patients) relative à :

- La finalité, gouvernance, responsabilité
- L'architecture
- L'origine des données et qualité (légalité de la collecte et des traitements)
- L'explication des processus, explication du périmètre de la partie non explicable
- La méthode d'apprentissage, d'inférence, etc.
- Les limites de l'utilisation de l'algorithme (FP, FN si classification)
- Les modalités de recours en cas d'erreurs
- L'implication des utilisateurs

Mise en œuvre des mécanismes de garantie humaine (PS, équipe de soins) pour assurer :

- Le contrôle par l'humain de l'IA
- L'autonomie décisionnelle
- Le maintien des compétences des utilisateurs
- L'intervalle de confiance de l'IA / garde-fou des erreurs de l'IA
- Les audits (désaccords IA / PS)

Définition de l'instance de régulation (audit, Label Ethique-IA)

Analyse d'impact organisationnel sur le parcours de soins

Analyse d'impact sociétal et environnemental

Guide d'implémentation : Libellés courts

1. Cadrage

Mettre en place un CSE

2. Collecte et préparation des données

Transparence sur l'origine des données d'apprentissage du SIA
Minimisation des données vs singularité
Minimisation des données vs impact environnemental

3. Conception de l'algorithme du SIA

Réduire et éliminer les biais du SIA
Explicabilité du SIA
Empêcher l'IA générative de restituer des données à caractère personnel utilisées pour son entraînement (G)
Garantir l'absence de plagiat de l'IA générative (G)
Autonomie pour la désactivation d'un SIA modulaire
Conception et utilisation écoresponsable du SIA
Bilan carbone de la fabrication du SIA

4. Conception des interfaces du SIA

Transparence sur l'interaction avec une IA
Vérifier la bonne compréhension de l'interaction avec une IA*
Visualiser les résultats de l'IA
Traçabilité de la désactivation du SIA modulaire
Garantir l'autonomie décisionnelle de l'utilisateur
Transparence sur le fait de suivre l'IA
Transparence sur le fait de ne pas suivre l'IA et recueil du motif de non suivi
Transparence sur le fait de ne pas utiliser le SIA
Conformité RGAA des SIA avec IHM
Intelligibilité des interfaces des SIA avec IHM

5. Evaluation du SIA

Cohérence des réponses de l'IA générative dans la répétition (G)
Qualité des réponses de l'IA générative (G)
Garantir que l'IA générative incite le patient utilisateur à consulter un professionnel de santé en cas d'alerte (G)
Performance de l'IA identique pour tous les publics couverts

6. Déploiement, formation et utilisation du SIA

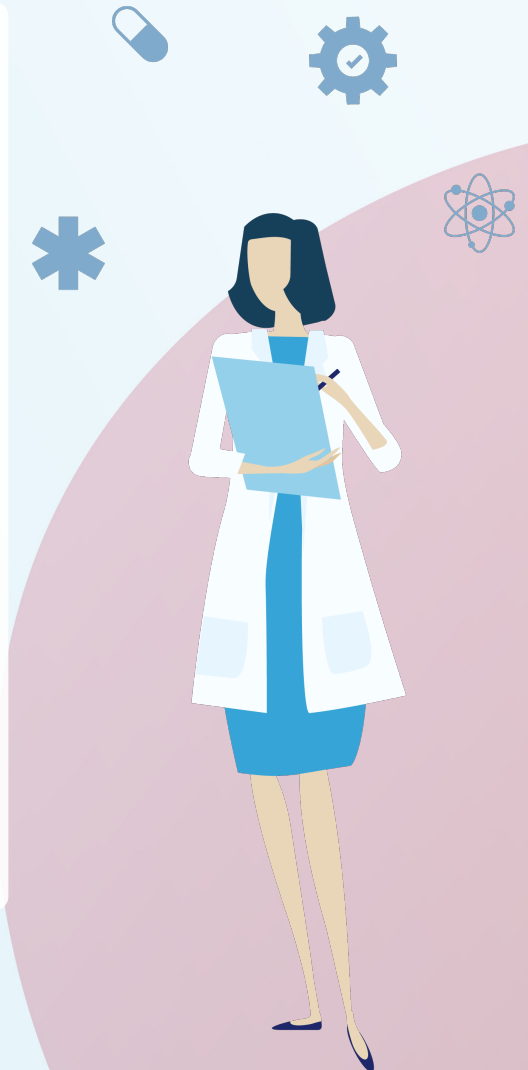
Accompagner les utilisateurs/la structure dans l'évaluation des impacts du déploiement d'un SIA
Co-construire avec les utilisateurs les ressources de formation aux enjeux de l'IA
Former les utilisateurs aux enjeux de l'IA (biais, limites de performance)
Vérifier la bonne compréhension par les utilisateurs des enjeux de l'IA en particulier des limites de performance
Former à l'utilisation du SIA en amont de son introduction
Permettre un accès facile et intuitif à la documentation du SIA
Matrice des responsabilités
Transparence sur le type d'IA employé
Transparence sur la performance de l'IA
Détecter la dépendance à l'IA
Alerter en cas de détection de dépendance
Garder un esprit critique
Rappeler la responsabilité de l'utilisateur sur sa décision
Sensibilisation à un usage écoresponsable
Souveraineté des données de l'organisation utilisatrice (G)

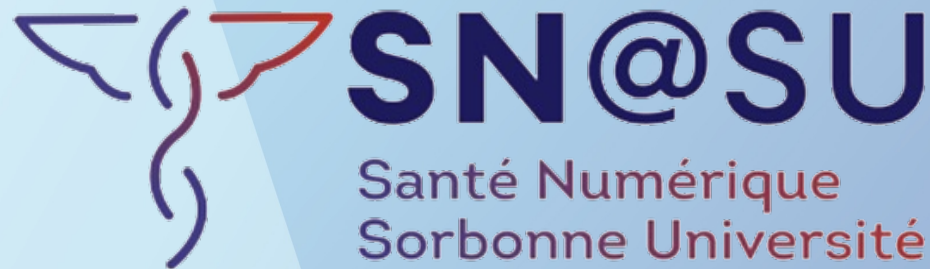
7. Suivi, mise à jour et amélioration continue

Evaluer le SIA en continu
Mettre en place une démarche d'amélioration continue de l'IA
Transparence sur la mise à jour d'un SIA et ses conséquences
Vérifier la bonne compréhension des conséquences d'une mise à jour d'un SIA*

TAKE HOME MESSAGES

- L'IA présente des risques (limites de performance, biais)
- Des recommandations pour une éthique by design des solutions d'IA en santé ont été proposées
- Un guide d'implémentation d'un SIA en santé éthique est en cours de concertation publique





Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre de France 2030 portant la référence ANR-23-CMAS-0001

Cette ressource pédagogique est placée sous la licence CC-BY-NC-ND 4.0