HUB FRANCE IA

L'IA ETHIQUE EN PRATIQUE

OPERATIONNALISER VOTRE SYSTEME D'IA AVEC UNE DEMARCHE ETHIQUE

2ème version - Décembre 2024

L'IA ÉTHIQUE EN PRATIQUE

OPERATIONNALISER VOTRE SYSTEME D'IA AVEC UNE DEMARCHE ETHIQUE

GROUPE DE TRAVAIL – IA & ÉTHIQUE - HUB FRANCE IA

TABLES DES MATIERES

ABREVIATIONS

Abréviations	Description
AI ACT	*Artificial Intelligence Act* (voir RIA)
CNIL	Commission Nationale de l'Informatique et des Libertés
GPU	*Graphical Processing Unit*
HIC	*Human In Command* (voir §3.1)
HITL	*Human In The Loop* (voir §3.2)
HOTL	Human On The Loop (voir §3.3)
HLEG	*High Level Expert Group on Artificial Intelligence*
IAG	Intelligence Artificielle Générative
RGPD	Règlement Général sur la Protection des Données
RIA	Règlement d'Intelligence Artificielle (voir *AI Act*)
SIA	Système d'Intelligence Artificielle
SIAG	Système d'Intelligence Artificielle Générative
XAI	*eXplainable AI*

INTRODUCTION

Préambule

Le groupe de travail éthique du Hub France IA est un groupe actif depuis la création de l'association en 2017. Stratégique et toujours en veille au sein des entreprises et organisations, le groupe éthique est un lieu de partage entre les membres. Les réflexions et les bonnes pratiques en ce qui concerne l'éthique de l'IA ont particulièrement évolué ces dernières années et les entreprises et organisations ont pris conscience de la nécessité, de plus en plus, d'opérationnaliser ce concept.

Courant 2021, à un moment où le Règlement européen sur l'intelligence artificielle (RIA) en était encore au stade de projet, le groupe de travail Éthique a souhaité entamer une nouvelle phase de ses travaux et réflexions : comment anticiper la mise en place du projet de réglementation ? Comment, par le partage d'expérience et de bonnes pratiques, aider les entreprises, institutions et collectivités à appréhender et déterminer un cadre tant de conception, de développement que d'utilisation éthique des systèmes d'intelligence artificielle ?

Les experts du groupe de travail ont développé une première version de livre blanc qui s'est voulue être un guide pour aider lors du développement, de l'implémentation et du déploiement de SIA au sein d'entreprises, d'administrations ou toutes entités ; visant à prendre en compte les enjeux éthiques dès le démarrage d'un projet afin de garantir le respect de principes et de droits fondamentaux et faire de notre écosystème un écosystème de confiance. Cette première version a vu le jour courant 2023, à un moment, donc, où aucune législation n'encadrait encore formellement le développement des SIA.

Pourquoi une seconde version de ce guide pratique ?

Depuis la première version, des événements majeurs sont intervenus : d'une part le RIA a été adopté par le Parlement Européen, et même dans l'attente de sa retranscription en droit français, ce règlement a, de fait, imposé un cadre juridique à l'industrie ; d'autre part nous avons assisté au déferlement de solutions issues de l'IA générative (SIAG), changeant par là-même profondément l'écosystème, les usages et aussi la perception que pouvaient avoir de l'IA tant les professionnels que le public. L'ambition de cette seconde version est donc toujours de faciliter l'opérationnalisation de l'éthique dans le cycle de vie de SIA, en prenant en compte ce nouveau contexte.

D'une part prendre en charge les impacts du développement croissant de SIAG et de leurs for

D'autre part, prendre en compte la mise en place d'un cadre réglementaire et la nécessité persistante de maintenir la réflexion éthique au sein de l'écosystème de l'IA. De fait, le RIA ne sonne pas la fin de la réflexion éthique qui était à son origine. Si le droit nous aide à penser l'IA à travers ce qu'elle peut être, l'éthique nous aide à penser ce progrès technique en tant que ce qu'elle devrait être : un vecteur de progrès humain.

Il ne s'agit donc pas de mettre éthique et droit en concurrence, et le texte du RIA adopté cette année renforce leur complémentarité nécessaire à une opérationnalisation performante

de l'IA. On retrouve d'ailleurs dans ce texte la référence à 7 principes pour une IA de confiance et leur traduction en obligations[1] :

Principes fondateurs du RIA et sa traduction en obligations

- **Facteur humain et contrôle humain** : Des mécanismes de surveillance appropriés doivent être mis en place, ce qui peut être réalisé grâce à des approches où l'humain peut intervenir à différentes étapes de la modélisation et du déploiement du SIA . **Cons. 66 & 73** - **Articles 13 & 14 RIA.**
- **Robustesse technique et sécurité** : Les systèmes d'IA doivent être résilients et sécurisés. Ils doivent être sûrs, garantir un plan de repli en cas de problème, et être précis, fiables et reproductibles. **Cons. 66 & 74 à 78 – Article 15 RIA.**
- **Respect de la vie privée et gouvernance des données :** Des mécanismes adéquats de gouvernance des données doivent être mis en place, en tenant compte de la qualité et de l'intégrité des données, et en garantissant un accès légitime aux données. **Cons. 66 à 70 - Article 10 RIA.**
- **Transparence** : Les systèmes d'IA et leurs décisions devraient être expliqués d'une manière adaptée aux parties prenantes concernées. Les humains doivent être conscients qu'ils interagissent avec un système d'IA et doivent être informés des capacités et des limites du système. **Cons. 66 & 72 -Article 13 RIA.**
- **Diversité, non-discrimination et équité :** En favorisant la diversité des utilisateurs et des données d'entraînement, les systèmes d'IA devraient associer les parties prenantes concernées tout au long de leur cercle de vie. **Cons.27 – Article 10 RIA.**
- **Bien-être social et environnemental** : Les systèmes d'IA devraient tenir compte de l'environnement, y compris d'autres êtres vivants, et leur impact social devrait être soigneusement pris en considération. **Cons. 1 -Article 95 RIA.**
- **Responsabilisation :** L'auditabilité, qui permet d'évaluer les algorithmes, les données et les processus de conception, y joue un rôle clé, en particulier dans les applications critiques. Il convient de garantir un recours adéquat et accessible. **Article 19 RIA**.

En outre, face à une technologie en constante évolution, on ne peut se contenter de seulement préserver la forme originelle que cet outil juridique a prise au moment de son adoption. Le RIA repose aussi sur la notion d'usage, ce qui le destine à évoluer. Il est dès lors particulièrement important que les expérimentations qui seront réalisées en IA dans les années à venir puissent être étudiées de façon critique. Au-delà de constituer une nécessité technologique, ces expérimentations seront déterminantes pour faire évoluer le droit de l'IA en gardant à l'esprit l'objectif de progrès humain et sociétal. Et c'est bien le rôle de l'éthique que de guider cette dynamique en aidant à définir une trajectoire souhaitable et soutenable pour l'IA, une trajectoire centrée sur l'humain, compatible avec le progrès humain.

C'est le but de ce guide pratique : (1) tout d'abord donner un accès pratique à des principes fondamentaux qui sont aussi mis en avant dans le RIA, les opérationnaliser, leur donner du sens ; et aussi, parce que le temps de l'intelligence collective n'est pas le temps du juridique, parce que ce temps nous permet d'être réactif face aux avancées technologiques, parce que l'éthique renvoie justement à un processus réflexif et critique, (2) envisager une trajectoire souhaitable et des pratiques qui la garantissent.

[1] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 14 – Contrôle humain

Comme son prédécesseur, ce livre blanc a été pensé comme un **guide pratique offrant aux concepteurs de SIA un parcours pour aligner l'IA aux valeurs** qu'il se doit d'incarner :

- **Une IA éthique est d'abord une IA sans heurt** : comment bâtir des garde-fous face aux possibles déviances du modèle ? Les résultats et opérations de l'IA se doivent ainsi d'être *sûrs* (sûreté), mais aussi de ne pas entraver la vie privée des citoyens (respect de la vie privée) ni d'induire ou d'aggraver d'injustes inégalités (équité).

- **Une IA éthique est ensuite une IA dont on peut rendre compte**. Des concepteurs aux utilisateurs finaux, le premier défi est de trouver un langage commun et exact pour expliquer la "boîte noire" de l'IA, dont le fonctionnement échappe en partie aux règles humaines (explicabilité). Mais ce n'est pas suffisant : dus justement à cette imprévisibilité, les possibles heurts doivent être *assumés* par l'entreprise d'IA (responsabilité).

- **Une IA éthique est, enfin, une IA qui réponde à un futur désirable**. Dans la situation créée avec l'IA, il est du rôle de l'entreprise de délimiter où l'humain peut intervenir dans les opérations de la machine et y apposer son consentement (autonomie). Au-delà de la place relative de l'humain dans le processus de développement et d'usage des SIA, ce livre blanc donne enfin des pistes aux organisations pour s'assurer que leur IA promeut un avenir compatible avec nos sociétés et leurs économies, notre santé, notre environnement (impact durable).

Le livre blanc que nous vous proposons ici présente plusieurs avantages :

- Donner des **définitions simples et claires** de ces sept principes éthiques de l'IA et de chacune de leurs dimensions ou sous-principes – l'équilibre environnemental étant par exemple sous-principe de l'impact durable de l'IA ;

- Apporter des **conseils pratiques** pour mettre en œuvre ces principes éthiques. Sous-principe par sous-principe, une grille (***check-list***) permet au praticien d'IA de **valider rapidement** qu'il réfléchit, intègre et assume sa vision, par exemple d'une IA qui limite le risque de biais (sous-principe de l'équité).

À travers le modèle de grille de maturité synthétique :

- Donner les moyens d'**auto-évaluer** simplement la maturité de son projet, dans l'ensemble des étapes de son cycle de vie, **établir une feuille de route éthique**

- Donner les moyens de dépasser les dimensions procédurales pour **s'engager dans une démarche proactive d'amélioration continue** en matière d'éthique.

Nous espérons que, partant de ces définitions et recommandations simples, vous serez mieux en mesure de mettre en œuvre une méthodologie de développement et/ou d'usage de l'IA certes conforme à la lettre du RIA, et aussi à l'esprit d'une IA éthique.

Si vous voyez des exigences manquantes, des outils opérationnels plus adaptés...

Vos commentaires nous seront précieux pour **compléter ensemble** le guide en donnant des clefs concrètes aux praticiens d'IA, tout en restant simple à l'adresse :

contact@hub-franceia.fr

COMMENT LIRE CE LIVRE BLANC

Fort de son écosystème riche en *startups*, en entreprises et en administrations centrales, le Hub France IA a souhaité mutualiser l'expression des différents besoins inhérents à l'éthique des algorithmes et de les décliner de manière opérationnelle.

Ce guide pratique est l'aboutissement de ce travail méticuleux. Il permet ainsi de dresser les sept principes éthiques relatifs au déploiement de solutions d'intelligence artificielle au sein d'une structure.

L'intégration de ces principes éthiques dans une entité peut être valorisée par différentes approches :

- Une sensibilisation de vos collaborateurs aux nouvelles obligations qui seront prévues au sein du règlement européen sur l'intelligence artificielle (RIA) :
- Un rôle prééminent de tiers de confiance, en garantissant dès à présent des systèmes d'intelligence artificielle prenant en considération les utilisateurs ;
- Une valorisation par une démarche RSE ou des indicateurs extra-financiers.

Prendre conscience de ces principes éthiques et les déployer au sein de votre structure est déjà un premier pas vers une IA de confiance. Le Hub France IA et l'ensemble des experts du groupe de travail éthique ont donc redéfini ces principes, en repartant des travaux européens du HLEG de la Commission Européenne et de Numeum. Les principes suivants ont ainsi été retenus : (i) sûreté, (ii) impact durable, (iii) autonomie, (iv) responsabilité humaine, (v) explicabilité, (vi) équité et (vii) respect de la vie privée.

Ce livre blanc a été découpé en sept sections dédiées à chacun de ces principes. Dans chaque section, vous trouverez la définition du principe en lui-même, ainsi que les sous-principes afférents. Chaque sous-principe comprend également une définition et les bonnes pratiques pouvant être mises en œuvre.

Nous attirons votre attention sur ces bonnes pratiques. Elles sont envisagées dans une hypothèse complète d'intégration au sein de vos méthodologies projet. Cependant, plusieurs étapes intermédiaires sont nécessaires pour parvenir à leur implémentation complète. De plus, vous pouvez décider d'avoir une intégration partielle de ces principes, garantissant *a minima* les principales mesures attendues.

Déployer une démarche IA éthique suppose plusieurs étapes d'implémentation distinctes. Ainsi, pour vous aider dans la mise en œuvre de ces principes, les experts du groupe de travail ont conçu une grille de maturité. Chaque niveau vous permet de vous situer dans votre implémentation des pratiques IA éthique ainsi que les prochaines étapes que vous pourrez envisagez dans votre démarche.

Bonne lecture !

GRILLE DE MATURIE

A qui s'adresse la grille ?

Le « guide pratique de l'IA éthique » offre des repères pour les organisations qui souhaitent concevoir, développer et/ou mettre en œuvre des systèmes d'IA respectueux des droits et libertés fondamentales, dans le souci de l'éthique. Pour cela, le guide s'attache à définir précisément les sept principes mis en avant par le Groupe d'Experts de Haut Niveau de l'Union Européenne (2018) et à associer à chacun d'eux une série de recommandations pratiques.

Ces recommandations sont reprises dans une grille de synthèse qui les positionne sur le cycle de vie des systèmes d'intelligence artificielle, et pointe quelques spécificités de l'IA générative. Pour des impératifs de lisibilité, nous avons ici présenté comme linéaire ce qui est un processus itératif, notamment dans le cas de l'apprentissage continu. La grille fait apparaître les actions souhaitables à différents niveaux, aussi bien techniques qu'organisationnels (gouvernance des projets, méthodologie de gestion des projets, dispositifs de veille et d'information).

Comment utiliser la grille de recommandations ?

L'entrée par le cycle de vie plutôt que par les principes se veut plus opératoire pour les organisations ayant un projet de conception / de déploiement d'un SIA. Elles pourront en faire usage en fonction de leur positionnement dans les chaînes de valeur, en visualisant ce qu'elles pourraient faire elles-mêmes ou ce qu'elles pourraient exiger de leurs partenaires / fournisseurs. La grille de recommandations est destinée à outiller l'examen critique des SIA développés / déployés et des actions et process mis en œuvre en regard de chaque principe. Utilisée dans une dynamique d'amélioration continue, elle permettra surtout aux organisations d'identifier les axes de progrès possibles / souhaitables à court terme ou à plus long terme.

Ainsi, la grille constitue un outil de synthèse permettant aux organisations ayant un projet de conception / de déploiement d'un SIA de :

- Identifier / hiérarchiser les principes et recommandations pertinentes dans le cadre de leur projet ;
- Evaluer leur niveau de maturité en positionnant ce qui est fait déjà par rapport aux actions souhaitables (identification des écarts) ;
- Etablir leur feuille de route pour s'engager dans une démarche proactive d'amélioration continue en matière d'éthique.

Pourquoi une démarche éthique ne peut-elle se limiter à cocher des cases sur une check-list ?

Le guide pratique de l'IA éthique intègre les avancées réglementaires récentes permises par l'adoption de l'AI Act. Le règlement européen repose sur une démarche de gestion des risques et envisage la certification et l'auditabilité des SIA comme des moyens d'établir la confiance nécessaire à l'adoption et à la diffusion des innovations. Le guide considère quant à lui les principes éthiques non pas comme des moyens mais comme des repères vers un horizon socio-technique désirable et soutenable. Il invite chaque acteur engagé dans une démarche de conception / déploiement d'un SIA à se poser la question non seulement du *comment* mais aussi du *pourquoi.*

Dans cette optique, la grille de recommandations n'est pas à considérer comme une check-list mais bien comme un support de questionnement et de réflexivité pour les organisations qui

conçoivent / déploient des SIA. Que fait-on ? Pourquoi le fait-on ? Que pourrait-on faire de plus ou de différent ? Quels sont les points de blocage actuels ? Où résident les dilemmes et les nécessaires arbitrages ? Comment les justifier ?

Des principes qui invitent à la délibération. S'entendre sur les principes structurant d'une IA éthique et les connaître constitue la première étape d'une démarche éthique. Les mettre en œuvre suppose toutefois une réflexion partagée, une délibération des parties prenantes internes et externes sur l'interprétation qu'il convient de donner à chacun de ces principes dans un cas d'usage particulier. Par exemple, comment déterminer le bien-être collectif ? à l'aune de quel critère apprécier l'équité entre les individus ou les groupes sociaux ? quel est le niveau d'automatisation souhaitable pour un SIA ? Autant de questions qui n'admettent aucune réponse immédiate et qui montrent la nécessité de penser l'IA éthique dans son alignement avec les valeurs de l'organisation et de ses parties prenantes internes et externes.

Des arbitrages nécessaires entre principes. Satisfaire tous les principes en même temps n'est pas toujours possible et certains d'entre eux sont en tension. Par exemple, assurer un haut niveau de fiabilité par des intervalles de confiance très resserrés peut venir à l'encontre du principe d'explicabilité et de transparence des algorithmes, de minimisation de l'empreinte environnementale ou encore être difficilement compatible avec le principe de minimisation des données à caractère personnel. Le principe de transparence peut lui aussi entrer en tension avec le principe de minimisation de l'empreinte environnementale s'il conduit à développer et entraîner des modèles alors que des modèles fondateurs sont déjà existants. Là encore, l'éthique en tant que savoir pratique mis en œuvre dans une situation ambigüe ou incertaine prend tout son sens.

Tendre vers ce qui est désirable plutôt que se limiter à ce qui est déjà possible. Les recommandations émises dans le livre blanc s'appuient sur l'état de l'art en matière d'IA éthique et responsable. Elles se veulent réalistes, tout en se sachant exigeantes et parfois difficiles à atteindre, en raison notamment de la complexité des chaînes de production des SIA. Ainsi, les fournisseurs / concepteurs de solutions algorithmiques qui s'appuient sur des modèles fondateurs se heurtent à l'opacité des conditions dans lesquelles ceux-ci ont été entraînés (corpus et data set d'apprentissage qui ne respectent pas le droit de la propriété intellectuelle et le RGPD par exemple). Les auteurs du présent document n'ignorent pas les rapports déséquilibrés entre les acteurs économiques et l'opacité qui masquent souvent les conditions de production dans les chaînes de valeur. Ils savent que ces éléments sont de nature à favoriser une déresponsabilisation individuelle et collective et c'est précisément pour aller à l'encontre de cet écueil que le livre blanc et la grille de maturité proposent des pistes d'action qui s'apparentent à une démarche d'enquête / de questionnement pour tendre vers une plus grande transparence, gage d'une plus grande maîtrise des systèmes par l'humain.

Ce qui a été fait reste à faire, du fait de l'indétermination des effets des SIA et du rythme des transformations socio-techniques. Comme toute technologie, les impacts socio-économiques, culturels et éthiques des SIA dépendent des usages qui en sont faits, ce à quoi il convient d'ajouter l'indétermination générée par les processus d'apprentissage continu sur lesquels repose le fonctionnement de certains SIA. De ce fait, une démarche éthique suppose une attention continue tout au long du cycle de vie des SIA depuis l'idéation du projet jusqu'à son démantèlement, en passant par la conception et le déploiement. Ce qui n'est pas encore

connu aujourd'hui pourrait devenir certain demain, ce qui nécessite de déployer une veille constante sur les plans techniques mais aussi sociétaux et environnementaux.

En résumé, il s'agit de concevoir une démarche d'amélioration continue, tendue vers l'éthique en tant qu'horizon désirable.

Votre organisation est pleinement engagée dans cette démarche si, pour chacun des cas d'usage envisagé :

- Les finalités et le bien-fondé du projet envisagé sont discutés dans une démarche pluraliste qui associe les parties prenantes internes et externes du projet ;
- Les principes sont connus et compris ;
- Les risques potentiels sont identifiés, principe par principe, ainsi que les critères d'évaluation (les métriques) adaptés au cas d'usage, pour chaque étape du cycle de vie ;
- Les risques sont gérés. Un processus de management des risques est mis en place, avec une documentation systématique des opérations techniques et des choix qui les sous-tendent ;
- Les arbitrages nécessaires entre principes sont identifiés et les choix sont argumentés.
- La démarche est itérative et continue, avec la mise en place d'un processus de monitoring du SIA (événements indésirables, réentraînement, alignement des modèles, feedback des utilisateurs) ;
- La démarche est à l'état de l'art grâce à la mise en place d'une veille ;
- La démarche est systématique et intégrée dans une structure de gouvernance qui tient compte des parties prenantes directes et indirectes tout au long du projet.

La grille de maturité est consultable en détail ci-après.

PRINCIPES	CYCLE DE VIE -->	CADRAGE / DESIGN DU PROJET		COLLECTE ET QUALIFICATION DES DONNEES		MODELISATION & TEST		DEPLOIEMENT		MONITORING & VEILLE		Documents de référence Pour aller plus loin
		Définir les finalités du projet Examiner la pertinence du recours à l'IA dans un cas d'usage identifié	Mettre en place les modalités et structures de gouvernance pour le projet	Collecter, annoter des données, qualifier des corpus d'apprentissage	*cas particulier de l'IA Générative*	Développer, entraîner, tester un SIA dans un environnement contrôlé	*cas particulier de l'IA Générative*	Mettre un SIA en service	*cas particulier de l'IA Générative*	Monitorer le SIA, en particulier dans le cas de l'apprentissage continu	Mettre en place un processus d'amélioration continue et de veille	
SÛRETÉ	**Robustesse** identifier les vulnérabilités internes et externes pour assurer un niveau de performance stable					Indiquer des métriques d'évaluation de la robustesse technique d'une IA en amont de la conception du système. Identifier une checklist de robustesse technique, dans l'environnement de test / recette avant mise en production.				Permettre au SIA de détecter par lui-même ses limites de fonctionnement et informer lorsque ces dernières sont atteintes. Définir un degré de tolérance à l'erreur d'un système informatique. Maintenir la stabilité du système dans son fonctionnement nominal. Permettre au système de rendre la main lorsque les données sont compromises. Gérer les exceptions par des procédures adaptées communiquées en amont. Mettre en place des tests périodiques (fixer une régularité) et à grande échelle (fixer un % des données de test) pour écarter les contenus non pertinents, faux, et discriminants (e.g. avec IA générative).		
	Fiabilité définir des intervalles de confiance dans lesquels les résultats sont fiables et reproductibles			S'assurer que la référence est obtenue par des protocoles fiables et contrôlés : garantir la fiabilité des données avec une annotation humaine soumise à une évaluation inter-juge pour limiter les biais.		Définir les résultats attendus, la référence ou les indicateurs utilisés pour mesurer le niveau de conformité des résultats. Définir des intervalles de confiance pour chaque usage / application.		Faire apparaître l'intervalle de confiance aux utilisateurs.				
	Protection physique contre les imprévus prévenir la mise en danger des personnes et des biens par les SIA	Identifier et évaluer les risques critiques associés à un défaut de fiabilité Qualifier et évaluer le SIA au regard des droits fondamentaux. Identifier les impacts directs lorsque le SIA est intégré dans le hardware ou dans un système de prise de décisions Identifier les impacts indirects physiques (situations de mise en danger potentielle de l'utilisateur).							Permettre le passage en mode "alternate", dans lequel le système donne la main à l'humain.	Evaluer attentivement les réponses générées par l'IA à travers une diversité de prompts, en analysant chaque contenu sous différents angles. Tout contenu inapproprié identifié doit être corrigé ou retiré immédiatement du système.			
IMPACT DURABLE	**Equilibre environnemental** minimiser l'empreinte environnementale par une analyse du cycle de vie	Vérifier que le besoin auquel le SIA répond soit une action positive ou neutre pour l'environnement. Vérifier que le besoin auquel le SIA est censé répondre ne peut pas être adressé par une méthode à moindre impact environnemental				Effectuer les traitements de données en veillant à minimiser leur impact environnemental Estimer le coût environnemental de la production des GPU qui ont été nécessaires à l'entraînement du modèle (que celui-ci soit pré- ou directement entraîné) Réduire le coût environnemental lié à l'entraînement d'un modèle en trouvant, au cas par cas, la bonne option, par exemple en utilisant des modules pré-entraînés, des réseaux de neurones « one-for-all », des algorithmes comprimés, en appliquant des méthodes telles que l'apprentissage fédéré	Limiter le nombre de paramètres dans les modèles LLM : Établir une stratégie de choix des modèles visant à optimiser la consommation énergétique globale, par ex. dans certains cas un modèle à 7 milliards de paramètres peut être optimal pour répondre au besoin caractérisé	Faire de la sobriété énergétique un critère de choix des fournisseurs, prestataires et sous-traitants impliqués dans la production, le développement et l'utilisation du SIA (par exemple en privilégiant des centres de données situés dans des zones nordiques). Adapter en continu la quantité d'information visualisée au besoin, afin de limiter la consommation d'énergie		Mesurer régulièrement l'impact environnemental du SIA, par exemple à l'aide de 4 indicateurs fréquemment utilisés (les ressources abiotiques, le GWP (réchauffement global et émissions de GES), l'eau et l'énergie. Pour ce qui est des algorithmes, il existe quelques outils (Green Algorithms, Carbon Tracker, Code Carbon, MLCO2 Impact ...) Considérer la possibilité d'optimiser le SIA au fil de son évolution, par exemple en supprimant les fonctionnalités non utilisées. Mettre en place une gouvernance des données : optimiser le volume de données récupérées au cours du cycle de vie du SIA et mettre en place une politique d'archivage, d'expiration et de suppression des données. Adapter en continu la quantité d'information visualisée au besoin, afin de limiter la consommation d'énergie. Planifier la fin de vie de tout ou partie du SIA en prêtant une attention particulière à l'avenir des données et des composants logiciels spécifiques au service numérique et à l'avenir des matériels et des ressources libérés.	Faire de la veille active sur les pratiques d'écoconception de services numériques. Diffuser ses connaissances en matière d'impact environnemental du numérique auprès de ses équipes et des utilisateurs.	Référentiel général pour l'IA frugale
	Apport socio-économique garantir la contribution des SIA au bien être collectif	Définir, le plus en amont possible de la conception du SIA, les valeurs socio-économiques à respecter dans le cadre de son développement et de son utilisation, par exemple en prenant en considération son domaine d'application, le nombre de personnes potentiellement impactées, les utilisateurs cibles. Réaliser une évaluation bénéfices-risques concernant l'impact socio-économique du SIA afin de s'assurer que cet impact soit positif par rapport aux valeurs prédéfinies.	Inclure les (futurs) utilisateurs du SIA dans sa conception et dans son développement et/ou son suivi.						Encadrer les usages du SIA afin qu'il ne puisse pas être exploité à des fins préjudiciables pour les individus, en leur qualité de salarié ou de citoyen. Mettre à disposition des utilisateurs du SIA du matériel de formation accessible et compréhensible leur permettant d'en faire une utilisation optimale, en toute conscience des capacités et limites du SIA, not. s'il a vocation à être intégré dans un métier existant Mettre en place une stratégie de compatibilité du SIA avec les terminaux et versions logicielles obsolètes		Évaluer l'accessibilité de l'interface utilisateur du SIA et l'améliorer en continu.		

	Impact sur la santé garantir une contribution directe et indirecte positive des SIA à la santé	Identifier le plus en amont possible les risques potentiels du SIA sur les fonctions cognitives et la santé mentale des utilisateurs et les éliminer ou les réduire au minimum. Mesurer l'impact que le SIA pourrait avoir sur la santé et la psyché Prendre en compte les principes de la bioéthique la bienfaisance, la non-malfaisance, la justice et l'autonomie dès la conception du SIA Prioriser les cas d'usage du SIA visant à favoriser un état de bien-être physique, mental et social, à réduire les inégalités d'accès au système de santé, à aider à identifier, évaluer, surveiller, prévenir ou résoudre les problèmes de santé présents ou à venir (virus, maladies chroniques, perte d'autonomie, vieillissement…).	Inclure les utilisateurs / les bénéficiaires finaux dans la conception du SIA, dans son développement et dans son amélioration continue.	S'assurer que les données personnelles traitées par le SIA ont été / sont collectées dans le respect des lois et réglementations applicables (notamment si elles sont collectées auprès de tiers) S'assurer que les données servant à l'entraînement du SIA sont représentatives de la population cible (mettre en balance avec le principe de minimisation des données à caractère personnel) Mettre en place des mécanismes visant à corriger les biais existant dans les données d'entraînement et les erreurs du SIA (sur-apprentissage, sous-apprentissage). Héberger les données de santé traitées par le SIA dans des centres de données certifiés "Hébergeurs de données de santé" .				Pouvoir expliquer les résultats du SIA à ses utilisateurs ainsi qu'aux patients éventuellement impactés.		monitorer la portée de la solution et réagir en conséquence lorsqu'un problème est signalé Assurer une supervision humaine du fonctionnement du SIA et de ses résultats.	mettre en place une veille sur les risques pour la santé	*CCNS Cadre de l'éthique du numérique en santé*
AUTONOMIE	**Human in command** l'humain conçoit et décide	Déterminer le niveau de délégation adéquat et adapté au regard de la finalité du SIA Définir la marge de manœuvre "idéale" de l'humain dans le SIA en posant un cadre éthique : le publier auprès des utilisateurs. S'assurer que les interdictions quant à la vie privée et aux droits fondamentaux sont respectées	Mettre en place une modalité de gestion de projet adaptée aux spécificités de l'IA							Mettre en place un processus de détection, d'alerte et de correction des défaillances (non-respect des droits fondamentaux, situation à risque). Assister l'utilisateur pour gérer la correction, en cas de détection de défaillance(s).	Définir un processus de reprise en main du SIA en cas de défaillance. Assister l'utilisateur pour gérer la correction, en cas de détection de défaillance(s).	
	Human in the loop l'humain intervient à des étapes choisies	Étudier les impacts de la mise en œuvre et l'exploitation du SIA. Identifier les points de contrôle humain.			Mettre en place un cahier de test & de recettes permettant de vérifier le niveau d'autonomie attendu.				Valider les actions humaines aux points de contrôle par rapport au risque de biais (garder un suivi / historique). Garantir la journalisation du suivi du SIA. Mettre en place une boucle de feedback du comportement du SIA pour remonter à l'utilisateur final ce qu'il se passe en vue de corriger le modèle.		
	Human on the loop l'humain supervise et n'intervient pas	Qualifier le contexte du SIA. Qualifier et évaluer le SIA au regard des droits fondamentaux.	Garantir le suivi du SIA par des protocoles de conception, d'essai et de suivi approuvés par un processus normatif rigoureux		Expérimenter le SIA afin d'évaluer son fonctionnement et ses impacts.				S'assurer que l'humain n'est pas "OUT of the Loop" : synthèses et portes d'entrée d'urgence dans le SIA pour corriger, surveiller, émettre des rapports notamment vers les utilisateurs		
	Consentement humain s'assurer que les utilisateurs / citoyens consentent à l'automatisation et à sa finalité				Mettre en place un message d'information à destination de l'utilisateur final qui soit visible, intelligible et paramétrable.		Par défaut, laisser aux utilisateurs la possibilité de ne pas recourir / être traité par le SIA. Mettre à disposition une méthode de contestation de l'action du SIA par l'utilisateur.			Réaliser un guide d'évaluation du SIA pour l'utilisateur final (consommateur du service).	
RESPONSABILITE	**Assumer les conséquences** mettre en œuvre les moyens d'être redevable des conséquences indésirables potentielles		Mettre en place une structure référente de gouvernance des SIA au sein de l'entreprise . Définir une procédure d'identification des responsabilités dans la chaîne de conception-mise en œuvre d'un SIA. Nommer un model owner qui sera le représentant technique unique des différents acteurs humains impliqués dans le cycle de développement du modèle , en relation avec la structure de gouvernance. Définir et catégoriser de façon aussi exhaustive que possible la nature des événements indésirables, i.e. les événements potentiels qui ne respecteraient pas les dispositifs réglementaires en vigueur et/ou les principes éthiques généraux définis. Définir - éventuellement de façon contractuelle - les motifs susceptibles d'engager la responsabilité de chaque acteur de la chaîne.		Dans le cas du transfer learning , la procédure doit permettre d'identifier les acteurs qui sont intervenus dans le développement et/ou dans le choix du modèle pré-entraîné et du corpus utilisé				De façon connexe à ce qui existe dans le système de santé, organiser et mettre en place une procédure de déclaration, d'enregistrement et de traitement des événements indésirables. Le système de déclaration doit pouvoir être accessible à tous les acteurs de la chaîne de production jusqu'aux utilisateurs.	La structure de gouvernance est responsable de la bonne prise en charge des événements indésirables mettre en place une veille pour intégrer les connaissances nouvelles sur les risques des SIA	
	Contrôle et auditabilité rendre possible l'examen des SIA par des tiers		Etablir une procédure d'audit interne et si possible externe Documenter notamment le besoin auquel le SIA répond, le processus de collecte des données d'entraînement et leur pertinence, le choix du modèle et les risques associés, les mesures de cybersécurité mises en oeuvre, les évolutions du modèle une fois le SIA déployé Assurer la protection de la propriété intellectuelle et du secret des affaires sans impacter la transparence.								

EXPLICABILITE	Transparence clarifier le fonctionnement technique du SIA			Enquêter et communiquer avec le plus de précision possible sur la nature et les tendances des données d'entraînement et réentraînement. Pour chaque critère, si la source des données est extérieure au concepteur, apporter la preuve qu'une démarche d'enquête a été initiée (Demande faite [date] / Informations collectées & éventuelle réponse [date]). Sur la nature des données collectées en input (géographie, sites, % de données personnelles)... Faire de même avec les inputs qui ont servi à fine-tuner le modèle (idem si le concepteur hérite d'un modèle de fondation déjà fine-tuned). Renseigner périodiquement sur les % de contenus erronés et potentiellement indésirables présents dans les corpus d'entraînement, et ceux générés par le modèle.		Privilégier l'algorithme le plus simple, reposant si possible sur des règles. Sinon, choisir celui qui répond le mieux au compromis performance (ou autre mesure) et complexité. Dans une notice synthétique, permettant de répondre à l'exigence de l'AI Act Article 12.2 sur la traçabilité du SIA et de son fonctionnement, documenter les méthodes employées, l'usage des données, l'historique des changements majeurs e.g. :Comment les données sont mises en forme ; Quel modèle a été choisi, suivant quelle méthode ; Quels hyper paramètres (i.e. fixés par le codeur) et comment ont-ils été choisis ;En quoi les paramètres (produits de l'entraînement du modèle) donnent des performances différentes ; Quelles sont les conditions de validité des résultats, etc.	si le concepteur arbitre entre plusieurs modèles de fondation à fine-tuner ou utiliser, en limiter le nombre de paramètres	Présenter l'outil d'explicabilité choisi, ses atouts et limites pour le cas d'usage traité. Afficher à l'utilisateur les limites techniques du modèle et de l'outil d'explicabilité, avec des exemples individuels. Notifier que l'humain interagit avec une IA, en cas d'interaction.	Auprès des utilisateurs, mettre en place des tests périodiques (fixer une régularité) et à grande échelle (fixer un % des données de test) pour écarter les contenus non pertinents, faux, et discriminants (e.g. avec IA générative). Prévenir l'utilisateur lorsque de nouvelles données sont utilisées pour ré-entraîner le modèle (e.g. IAG)	Donner des exemples de différences de prédictions avant / après, s'il y a évolution continue du programme. Prévenir l'utilisateur lorsque de nouvelles données sont utilisées pour ré-entraîner le modèle (e.g. IAG), en notant de la manière la plus précise leurs similitudes / différences avec les inputs de départ (mêmes types de phrases ; mêmes individus... Présence / absence de données personnelles...)		*AI Act (art. 12 et art. 13)*
	Justifiabilité permettre d'identifier une rationalité dans le fonctionnement du SIA	Expliciter et consigner par écrit les choix éthiques qui orientent l'IA (exemple : avant tout vie privée, anonymisation des exemples, les individus présentés dans l'explication ne sont pas réels).	Expliciter les sources.	Ex IA générative : durant le fine-tuning ou en surcouche d'explication, apporter des sources qui viennent en références et permettent de justifier la réponse de l'IA, tout en rendant intelligible le contenu généré (e.g. prompt) aux utilisateurs visés.	Comparer les opérations de l'IA au raisonnement d'un humain dans le même cas d'usage. Donner et documenter des exemples des limites de cette comparaison humain / IA (exemple : machine-learning vs raisonnement logique). Présenter l'outil d'explicabilité choisi, ses atouts et limites pour le cas d'usage traité.						
	Intelligibilité adapter l'information à l'utilisateur						Préparer des schémas de vulgarisation selon les profils auxquels expliquer l'IA (scenarii-types) Varier les supports : interface , démo, vocabulaire pour les différentes équipes Tester si le support est perçu comme explicatif chez les utilisateurs visés, en recourant à des méthodes interdisciplinaires pour limiter les biais Pouvoir expliquer les résultats du SIA à ses utilisateurs ainsi qu'aux patients éventuellement impactés.				
EQUITE	Mesure / gestion des biais		Mettre en place des mécanismes visant à corriger les biais existants dans les données d'entraînement et les erreurs du SIA (sur-apprentissage, sous-apprentissage).		Identifier le type de biais que l'entreprise souhaite corriger : entre individus, entre groupes ou sous-groupes. Une fois les biais analysés par variable, analyser les effets de variables croisées (intersectionnalité, e.g. « sexe » et « âge »). Choisir une métrique d'équité correspondant au biais qualifié comme injuste (régulation, culture d'entreprise). Choisir à quelle étape rendre l'IA plus « équitable » : au niveau des données (s'il est possible de les changer), pendant l'entraînement (pour une optimisation statistique et métrique d'équité), après, ou par une méthode hybride. Choisir et justifier le type de transformation équitable à adopter (ex : quelle méthode pré-, intra-, post-traitement ou de traitement hybride).				Suivre la mise à jour des données, pour vérifier l'impact du SIA sur les populations à protéger. Pour l'IA générative, renseigner périodiquement sur les % de contenus erronés et potentiellement indésirables présents dans les corpus d'entraînement, et ceux générés par le modèle.		
	Diversité des concepteurs	S'assurer que le recrutement tient compte de la diversité des équipes (profils, origines, formations académiques, visions, âge) Associer des intervenants internes ou externes spécialistes d'éthique pour former les data-scientists aux aspects éthiques et juridiques du projet Pratiquer et publier des auto-évaluations régulières de la conformité des travaux des data-scientists à la régulation et aux valeurs de l'organisation. S'assurer que le guide d'auto-évaluation interroge les choix et documente la méthode de collecte des données, l'étude des corrélations et biais, et les mesures prises face aux biais considérés comme inéquitables.									*Guide d'auto-évaluation des SIA. CNIL*
	Accès ouvert	Définir et publier les phases du SIA étant en libre accès (doit être géré par les acteurs de la stratégie et de la régulation). Définir ce qui est accessible avec / sans coût.			Sécuriser les parties de code à protéger						

RESPECT DE LA VIE PRIVEE	Respect de la vie privée											CNIL
	Anonymat			Analyser si l'anonymisation est la méthode permettant de protéger la vie privée par conception et par défaut la plus adaptée à l'usage qui doit être fait des données. Identifier les informations à conserver selon leur pertinence et les contraintes techniques de conception. Supprimer les éléments d'identification directe ainsi que les valeurs rares qui pourraient permettre une ré-identification avée des personnes. Distinguer les informations importantes des informations secondaires ou inutiles (c'est-à-dire supprimables). Définir la finesse idéale et acceptable pour chaque information conservée.					Effectuer une veille régulière pour préserver, dans le temps, le caractère anonyme des données produites. Cette veille doit prendre en compte les moyens techniques disponibles ainsi que les autres sources de données qui peuvent permettre de lever l'anonymat des informations.		
	Contrôle et consentement	Qualifier les traitements de données à caractère personnel opérés par le SIA (haut risque ou non) et, si nécessaire, réaliser une évaluation d'impact.	S'assurer de la conformité légale et réglementaire des données à caractère personnel traitées par le SIA - en présence d'un délégué à la protection des données au sein de la structure, auquel cette mission doit être confiée, notamment concernant la minimisation des données personnelles. Documenter la conformité et maintenir un registre des activités de traitement de données personnelles.	Ne pas utiliser les données personnelles traitées par le SIA de manière contraire aux intérêts des personnes concernées. Ne pas traiter de données sensibles (par exemple données de santé, données biométriques, données génétiques…) sauf cas exceptionnel autorisé par le RGPD				Faire apparaître de manière accessible, claire et intelligible aux personnes produisant les données qui sont traitées par le SIA les conditions légales et réglementaires de collecte, stockage, traitement et partage. Privilégier la base légale du consentement quand cela est possible dans le cas où le SIA collecte des données personnelles concernant ses utilisateurs, sinon informer l'utilisateur de l'utilisation de ses données pour qu'il puisse faire valoir ses droits. Recueillir ce consentement de manière libre (pas de conséquences négatives en cas de refus et surtout, pas de coercition), spécifique (si plusieurs finalités, demander le consentement pour chacune), éclairée (mentionner les informations exigées par le RGPD de manière compréhensible par toutes les personnes concernées, quel que soit leur âge ou leur niveau d'études) et univoque (acte positif clair. Ex : pas de case pré-cochées). Mettre en place des protocoles d'accès aux données personnelles traitées par le SIA, précisant qui peut avoir accès aux données et dans quelles circonstances. Permettre aux personnes concernées de retirer leur consentement d'une manière aussi simple que celle par laquelle elles y ont consenti, à tout moment. Garantir le droit à l'effacement permettant aux personnes concernées de supprimer les données les concernant. Favoriser l'exercice du droit à la portabilité des données personnelles par les personnes concernées. Anticiper le droit pour les destinataires, en présence d'un SIA utilisé dans le cadre d'un processus de décision, d'obtenir à la place l'intervention d'un humain . Faire comprendre à l'utilisateur ou au bénéficiaire du SIA qu'il interagit avec un SIA, notamment dans les cas suivants : interaction directe avec un SIA, utilisation d'un SIA pour reconnaître des émotions, utilisation d'un SIA ayant pour finalité de catégoriser des individus selon des critères biométriques, utilisation d'un SIA générant ou manipulant du contenu audio ou visuel .	Prévenir l'utilisateur lorsque de nouvelles données sont utilisées pour ré-entraîner le modèle (e.g. IAG)		Définir une durée de conservation des données adéquate aux finalités de traitement et à la durée légale de conservation. Mettre en place une gouvernance des données : optimiser le volume de données récupérées au cours du cycle de vie du SIA et mettre en place une politique d'archivage, d'expiration et de suppression des données.	*Se référer aux lignes directrices du Comité Européen pour la Protection des Données sur le consentement , et au Guide d'auto-évaluation pour les systèmes d'IA publié par la CNIL et à ses premières recommandations sur le développement des SIA .*

PRINCIPE DE SURETE

SURETE

DEFINITION DU PRINCIPE

Le principe de la sûreté couvre toutes les problématiques liées à la cybersécurité, aux risques, à l'imprévisibilité et à la volatilité de l'environnement réel où les SIA sont intégrés. Dans ce premier principe, nous distinguons trois sous-principes : la robustesse, la fiabilité des résultats fournis par le SIA et la protection physique contre les imprévus.

ROBUSTESSE

DEFINITION

Un SIA peut être considéré comme robuste, si et seulement si, son niveau de performance est stable face aux éventuels phénomènes à risque qui peuvent se présenter. Anticiper et contrer un phénomène à risque requiert l'identification des vulnérabilités du SIA (quantité et type de données d'entraînement, stockage...) *en rapport avec ses objectifs*[2].

Exemple : si un SIA a pour but de différencier des espèces de chiens, la dépendance des données à un motif de l'image en réalité indépendant de la cible (par exemple, les images de chiens-loups sont majoritairement sur fond de neige) est un phénomène à risque – risque intrinsèque.

Nous distinguons deux familles de risques :

- Les **risques intrinsèques** au système pouvant provenir d'anomalies des données d'entrée (*data drift, concept drift*, *etc*.), d'inférence ou de format de sortie. Ils peuvent aussi provenir d'un décalage entre ce qu'attend le concepteur (spécification fonctionnelle) et ce sur quoi le SIA fonde sa décision pour minimiser son erreur.
- Les **risques extrinsèques** provenant d'un acteur externe : cyberattaques, actions malfaisantes, *etc*.

Les SIA devraient donc être développés selon une approche de prévention des risques, de telle manière que ces systèmes se comportent de manière fiable et conformément aux attentes tout en réduisant le plus possible les atteintes involontaires et inattendues sur les systèmes (vulnérabilités qui pourraient permettre à des adversaires de les exploiter) ou l'infrastructure sous-jacente, tant matérielle que logicielle.[3]

Plusieurs vulnérabilités des grands modèles linguistiques (LLM) ont été traitées dans l'initiative https://llmtop10.com/. Ces vulnérabilités incluent des manipulations de prompt, une gestion insécurisée des sorties, des données d'entraînement altérées, , des divulgations d'informations sensibles, des conceptions de plugins insécurisées, une surdépendance et le vol de modèle.

[2] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 15 – Exactitude, Robustesse et cybersécurité

[3] Leike, J., Martic, M., Krakovna, V., Ortega, P. A., Everitt, T., Lefrancq, A., ... & Legg, S. (2017). *AI safety gridworlds*. arXiv preprint arXiv:1711.09883.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Indiquer des métriques d'évaluation de la robustesse technique d'une IA en amont de la conception du système.		
Identifier une *checklist* de robustesse technique, dans l'environnement de test / recette avant mise en production.		
Décrire les conditions attendues de fonctionnement et les types de perturbation auxquels les inputs peuvent être soumis		
Permettre au SIA de détecter par lui-même l'atteinte de ses limites de fonctionnement et d'en informer l'utilisateur		
Définir un degré de tolérance à l'erreur d'un système informatique.		
Maintenir la stabilité du système dans son fonctionnement nominal.		
Permettre au système de rendre la main lorsque les données sont compromises.		
Gérer les exceptions par des procédures adaptées communiquées en amont.		

FIABILITE DES RESULTATS

DEFINITION

Un SIA produit des résultats dits fiables si ces résultats permettent la réalisation de la finalité ou des finalités attendues s'ils sont reproductibles dans le cadre d'essais répétés dans les mêmes conditions[4]. Or, par nature, les SIA ne peuvent pas garantir que leurs résultats puissent être reproductibles, *a fortiori* s'ils apprennent en continu. Les résultats n'étant pas déterministes, la garantie de reproductibilité ne peut être que probabiliste autour d'intervalles de confiance. La reproductibilité est également difficile à appliquer pour les systèmes complexes qui manipulent un très grand nombre de paramètres. Les résultats peuvent varier en fonction des paramètres et de l'architecture du modèle, ce qui différencie les approches basées sur l'apprentissage automatique des approches symboliques.

[4] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 15 – Paragraphe 3

Reproduire les résultats nécessiterait donc une grande capacité de calcul due aux nombres de paramètres, et au nombre des combinaisons possibles[5].

Pour garantir la fiabilité des contenus générés par un SIA basé sur l'IA Générative, il est essentiel de conduire des tests réguliers et étendus. Ces tests permettent de filtrer les contenus inappropriés et d'intégrer des mécanismes préventifs pour assurer la pertinence des réponses.

Il est essentiel d'évaluer attentivement les réponses générées par l'IA à travers une diversité d'inputs (de prompts dans le cas de l'IAG), en analysant chaque contenu sous différents angles. Tout contenu inapproprié identifié doit être corrigé ou retiré immédiatement du système.

En parallèle, il est recommandé de prévenir la génération de contenus inappropriés en intégrant des mécanismes avancés. Ces mécanismes peuvent inclure la définition d'intervalles de confiance et l'ajout de règles supplémentaires pour guider la génération, particulièrement lorsque les questions ne correspondent pas aux demandes possibles.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Définir les résultats attendus, la référence ou les indicateurs utilisés pour mesurer le niveau de conformité des résultats.		
S'assurer que la référence est obtenue par des protocoles fiables et contrôlés : il faut garantir la fiabilité des données avec une annotation humaine soumise à une évaluation inter-juge et s'assurer de la neutralité des personnes et de leur objectivité.		
Définir des intervalles de confiance pour chaque usage / application.		
Faire apparaître l'intervalle de confiance aux utilisateurs.		
Mettre à contribution les statistiques pour définir ces méthodes / outils (ex : analyse de sensibilité) [6].		

[5] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv preprint arXiv:1802.07228
[6] de Rocquigny E., Modeling under Risk and Uncertainty, Wiley 2012

PROTECTION PHYSIQUE CONTRE LES IMPREVUS

DEFINITION

L'IA reste un outil qui peut ainsi être utilisé par les humains, aussi bien pour servir des fins socialement bénéfiques - superviser l'agriculture, protéger les espèces - que nuisibles – figure en *pole-position,* le score de crédit social de chaque résident chinois reposant sur un système de reconnaissance faciale [7]. Au même titre que l'énergie nucléaire dont l'ensemble des usages futurs n'était pas d'origine prévisible, la recherche sur les SIA peut (et nous le croyons, doit) certes être orientée, mais elle ne peut empêcher tous les potentiels mésusages à venir. Il n'est pas possible pour les chercheurs en IA d'éviter simplement de produire des recherches et des systèmes qui peuvent être dirigés vers des fins nuisibles[8,9].

Une phase importante de la conception d'un SIA réside dans l'anticipation des différentes modifications potentiellement dangereuses de l'environnement physique dues à l'utilisation du SIA[10]. Cette anticipation pourrait se concrétiser en développant un système d'autoprotection sous-jacent qui permettrait de mettre à l'arrêt l'outil dans le cas où l'intégrité physique d'un humain est menacée.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Identifier des secteurs dans lesquels la fiabilité des résultats est la plus critique. Notamment ceux où l'enjeu sécuritaire est prépondérant (centrale nucléaire, aéronautique...), et ceux qui sont catégorisés à « haut risque » dans l'annexe III du RIA.		
Permettre le passage en mode "*alternate*", dans lequel le système donne la main à l'humain.		
Identifier des exemples d'impact direct physique lorsque le SIA est intégré dans le *hardware* (exemple : robots, véhicules autonomes...) ou intégré dans un système de prise de décisions (exemple : situation où une application de microcrédit altère les conditions de vie d'un individu).		
Identifier des exemples d'impact indirect physique comme l'identification des situations de mise en danger physique potentielle de l'utilisateur.		

[7] SOUVERAIN T., MERIC N., *State of Ethical AI in 2021: Challenges posed by Ethics to companies developing AI*, DreamQuark, 2021, 110p.
[8] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv preprint arXiv:1606.06565.
[9] DIETTERICH, T. G., *Steps toward robust artificial intelligence*, Ai Magazine 2017, p. 3-24
[10] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 5 – Paragraphe 1 – Alinéa c

Dans le cas où cela est possible, lister et chiffrer le % d'applications bénéfiques / nuisibles prévisibles du SIA et fixer le seuil que la proportion doit dépasser.		
Qualifier et évaluer le SIA au regard des droits fondamentaux.		

PRINCIPE D'IMPACT DURABLE

IMPACT DURABLE

DEFINITION DU PRINCIPE

L'IA peut se distinguer d'autres technologies du fait qu'elle permet l'automatisation de nombreuses tâches jusqu'alors considérées comme réalisables uniquement par des êtres humains, ce qui est à la fois source d'opportunités (par ex. gains de productivité) et de risques (par ex. perte d'autonomie humaine dans la réalisation de certaines tâches). Par ailleurs, avec la généralisation de l'utilisation d'IA génératives de divers contenus (texte, image, son, vidéo …) au sein de notre société depuis la fin d'année 2022, de nouveaux risques émergent, susceptibles d'impacter de manière plus globale notre modèle social (par ex. manipulation de l'information, violation des droits d'auteur, impact plus important sur l'environnement…).

Afin de permettre la conception et le développement de SIA respectueux de nos valeurs et droits fondamentaux européens, il semble donc essentiel que les acteurs concernés prennent en considération l'impact sociétal que ces technologies pourraient avoir dès la conception de SIA, et imaginent des applications à impact durable contribuant à des futurs désirables par le plus grand nombre, y compris les générations futures[11]. Une première étape permettant de parvenir à un SIA à impact durable est donc de prendre en considération le plus en amont possible ses impacts potentiels sur l'environnement et sur la société, en vue d'en minimiser les effets négatifs et d'en maximiser les effets positifs pour les générations actuelles et futures. A titre d'orientation, les acteurs concernés peuvent dans un premier temps se référer à leurs objectifs stratégiques et de développement durable (mais aussi aux Objectifs de Développement Durable définis par les Nations Unies à réaliser à l'horizon 2030[12]).

EQUILIBRE ENVIRONNEMENTAL

DEFINITION

L'empreinte environnementale d'un SIA est liée à la somme des impacts environnementaux (directs, indirects, induits) dus à sa production, à son développement et à son utilisation, ce qui implique de prendre en compte les émissions directes de chaque acteur intervenant dans la chaîne de production.

Tout d'abord, il convient donc de minimiser le coût environnemental de la chaîne de production des matériaux nécessaires à la création de SIA[13]. Vient ensuite la nécessité de minimiser le coût environnemental du développement et du fonctionnement des SIA, en réduisant l'énergie nécessaire pour cela, par exemple en adoptant une approche frugale (éco-conception, des infrastructures écologiques, la sobriété des données..) visant à limiter la consommation énergétique à ce qui est strictement nécessaire pour assurer la performance

[11] 17 objectifs de développement durable ont été définis par les Nations Unies dans l'Agenda. Ils couvrent l'intégralité des enjeux de développement dans tous les pays tels que le climat, la biodiversité, l'énergie, l'eau, la pauvreté, l'égalité des genres, la prospérité économique ou encore la paix, l'agriculture, l'éducation, etc. Pour en savoir plus, voir notamment : https://www.agenda-2030.fr/17-objectifs-de-developpement-durable/

[12] 17 objectifs de développement durable ont été définis par les Nations Unies dans l'Agenda. Ils couvrent l'intégralité des enjeux de développement dans tous les pays tels que le climat, la biodiversité, l'énergie, l'eau, la pauvreté, l'égalité des genres, la prospérité économique ou encore la paix, l'agriculture, l'éducation, etc. Pour en savoir plus, voir notamment : https://www.agenda-2030.fr/17-objectifs-de-developpement-durable/

[13] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 95 – Paragraphe 2 – Alinéa b

attendue du modèle d'IA par rapport aux objectifs poursuivis. Enfin, afin de compenser le coût environnemental de la chaîne de production, du développement et du fonctionnement, une bonne pratique serait de favoriser des usages à impact positif ou neutre sur l'environnement..

BONNES PRATIQUES[14]

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Vérifier que le besoin auquel le SIA est censé répondre ne peut pas être adressé par une méthode à moindre impact environnemental (dans certaines circonstances il se peut que l'IA ne soit pas une solution optimale)		
Vérifier que le besoin auquel le SIA répond soit une action positive ou neutre pour l'environnement.		
Faire de la sobriété énergétique un critère de choix des fournisseurs, prestataires et sous-traitants impliqués dans la production, le développement et l'utilisation du SIA (par exemple en privilégiant des centres de données situées dans des zones nordiques).		
Effectuer les traitements de données en veillant à minimiser leur impact environnemental		
Se renseigner sur les meilleurs arbitrages "coût-efficacité environnementale" pour opérer un calcul (cela peut être l'achat, la location de GPU, la sous-traitance...)		
Mesurer régulièrement l'impact environnemental du SIA, par exemple à l'aide de 4 indicateurs fréquemment utilisés (les ressources abiotiques, le GWP (réchauffement global et émissions de GES), l'eau et l'énergie. Pour ce qui est des algorithmes, il existe quelques outils (Green Algorithms, Carbon Tracker, Code Carbon, MLCO2 Impact ...)		
Réduire le coût environnemental lié à l'entraînement d'un modèle en trouvant, au cas par cas, la bonne option, par exemple en utilisant des modules pré-entraînés, des réseaux de neurones « one-for-all », des algorithmes comprimés, en appliquant des méthodes telles que l'apprentissage fédéré[15]		
Établir une stratégie de choix des modèles visant à optimiser la consommation énergétique globale, par ex. dans certains		

[14] AFNOR SPEC 2201. Ecoconception des services numériques. Z77-102-0. Avril 2022, 58 p.
[15] Se référer aux bonnes pratiques et aux recommandations de l'AFNOR SPEC « Référentiel général sur l'IA frugale » (paru le 28 juin et téléchargeable ici : https://www.boutique.afnor.org/fr-fr/norme/afnor-spec-2314/referentiel-general-pour-lia-frugale-mesurer-et-reduire-limpact-environneme/fa208976/421140

cas un modèle à 7 milliards de paramètres peut être optimal pour répondre au besoin caractérisé		
Planifier la fin de vie de tout ou partie du SIA en prêtant une attention particulière à l'avenir des données et des composants logiciels spécifiques au service numérique et à l'avenir des matériels et des ressources libérés.		
Considérer la possibilité d'optimiser le SIA au fil de son évolution, par exemple en supprimant les fonctionnalités non utilisées..		
Mettre en place une gouvernance des données : optimiser le volume de données récupérées au cours du cycle de vie du SIA et mettre en place une politique d'archivage, d'expiration et de suppression des données.		
Adapter en continu la quantité d'information visualisée au besoin, afin de limiter la consommation d'énergie.		
Faire de la veille active sur les pratiques d'écoconception de services numériques.		
Diffuser ses connaissances en matière d'impact environnemental du numérique auprès de ses équipes et des utilisateurs.		

APPORT SOCIO-ECONOMIQUE

DEFINITION

L'empreinte socio-économique d'un SIA peut être appréciée par rapport à la somme de ses répercussions (directes, indirectes ou induites) sur l'économie, sur l'emploi et sur les interactions sociales.

Plusieurs critères peuvent être pris en considération afin d'évaluer l'apport socio-économique d'un SIA. Tout d'abord, les usages d'un SIA devrait contribuer à un bien-être collectif, par exemple en favorisant la prospérité des individus et une société plus juste. Par ailleurs, le(s) besoin(s) au(x)quel(s) il répond ainsi que ses fonctionnalités devraient toujours être pensés dans un objectif de complémentarité avec les êtres humains et ce, qu'il ait vocation à être utilisé dans la sphère professionnelle, sociale ou privée. Enfin, les conditions d'accès aux SIA devraient toujours être pensées d'une manière égalitaire de sorte que le plus grand nombre de personnes parmi le public cible puisse l'utiliser, quelles que soient leur situation et leur condition (cf. Principe d'équité).

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Définir, le plus en amont possible de la conception du SIA, les valeurs socio-économiques à respecter dans le cadre de son développement et de son utilisation, par exemple en prenant en considération son domaine d'application, le nombre de personnes potentiellement impactées, les utilisateurs cibles.		
Réaliser une évaluation bénéfices-risques concernant l'impact socio-économique du SIA afin de s'assurer que cet impact soit positif par rapport aux valeurs prédéfinies.		
Encadrer les usages du SIA afin qu'il ne puisse pas être légalement exploité à des fins préjudiciables pour les individus, en leur qualité de salarié ou de citoyen.		
Inclure le parties prenantes directes et indirectes au SIA dans sa conception et dans son développement et/ou son suivi.		
Évaluer l'accessibilité de l'interface utilisateur du SIA [16] et l'améliorer en continu.		
Mettre à disposition des utilisateurs du SIA du matériel de formation facilement accessible et compréhensible leur permettant d'en faire une utilisation optimale, en toute conscience de capacités et des limites du SIA, notamment lorsque celui-ci a vocation à être intégré dans un métier existant		
Mettre en place une stratégie de compatibilité du SIA avec les terminaux et versions logicielles obsolètes afin d'éviter de creuser la fracture numérique et de lutter contre l'obsolescence logicielle.		

IMPACT SUR LA SANTE

DEFINITION

Les SIA peuvent avoir un impact sur la santé des individus. Il peut être direct, par exemple dans le cadre d'un usage ayant une finalité de soin, de recherche en santé ou encore de mise en œuvre d'une politique de santé publique. Il est dans ce cas important que leurs usages dans le domaine de la santé respectent les valeurs éthiques phares de ce secteur ainsi que les

[16] DINUM, *Référentiel Général d'Amélioration de l'Accessibilité*, version 4.1, Décembre 2020, 128 p.

normes en vigueur, notamment afin de permettre une protection effective des données personnelles[17].

Cependant, des SIA n'ayant pas directement pour objet d'améliorer la santé des individus peuvent eux aussi avoir un effet sur celle-ci et notamment sur leurs fonctions cognitives et sur leur mental. Il est alors important de prévenir de potentiels effets délétères, à court terme comme à long terme, sur la psychologie des utilisateurs.

→ LORSQUE LE SIA N'A PAS VOCATION A ETRE UTILISE DIRECTEMENT AU SERVICE DE LA SANTE

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Identifier le plus en amont possible les risques potentiels du SIA sur les fonctions cognitives et la santé mentale des utilisateurs et les éliminer ou les réduire au minimum		
Mesurer l'impact que le SIA pourrait avoir sur la santé et la psyché (pour l'instant, en dehors de l'utilisation de modèles fondés sur une économie de l'attention, il n'y a pas à notre connaissance d'outils ou de méthodes permettant d'anticiper pleinement les impacts, il apparaît plus simple de monitorer la portée de la solution et de réagir en conséquence lorsqu'un problème est signalé).		

→ LORSQUE LE SIA A VOCATION A ETRE UTILISE AU SERVICE DE LA SANTE[18]

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Prendre en compte les principes de la bioéthique la bienfaisance, la non-malfaisance, la justice et l'autonomie dès la conception du SIA[19].		
Prioriser les cas d'usage du SIA visant à favoriser un état de bien-être physique, mental et social, à réduire les inégalités d'accès au système de santé, à aider à identifier, évaluer, surveiller, prévenir ou résoudre les problèmes de santé présents ou à venir (virus, maladies chroniques, perte d'autonomie, vieillissement...).		

[17] Règlement - 2016/679 - EN - rgdp - EUR-Lex (europa.eu) – Article 9
[18] Recommandations de bonnes pratiques pour intégrer l'éthique dès le développement des solutions d'intelligence artificielle en santé : *mise en œuvre de "l'éthique by design"*, publiées par la Délégation ministérielle au numérique en santé en avril 2022, ANFH [en ligne], Disponible à cette adresse : https://www.anfh.fr/actualites/recommandations-de-bonnes-pratiques-pour-integrer-l-ethique-des-le-developpement-des-solutions-d
[19] Afin de faciliter l''acculturation des entrepreneurs du numérique à ces principes spécifiques au secteur de la santé, la plateforme G-Nius propose une boîte à outils éthiques dédiée : https://gnius.esante.gouv.fr/fr/reglementation/fiches-reglementation/boite-a-outils-ethique.

Inclure les utilisateurs / les bénéficiaires finaux dans la conception du SIA, dans son développement et dans son amélioration continue.		
S'assurer que les données personnelles traitées par le SIA ont été / sont collectées dans le respect des lois et réglementations applicables (notamment lorsque ces données sont collectées auprès de tiers)[20].		
S'assurer que les données servant à l'entraînement du SIA sont bien représentatives de la population cible (à noter toutefois que cette bonne pratique doit être mise en balance avec le principe de minimisation des données à caractère personnel)[21].		
Mettre en place des mécanismes visant à corriger les biais existant dans les données d'entraînement et les erreurs du SIA (sur-apprentissage, sous-apprentissage).		
Pouvoir expliquer les résultats du SIA à ses utilisateurs ainsi qu'aux patients éventuellement impactés.		
Héberger les données de santé traitées par le SIA dans des centres de données certifiés "Hébergeurs de données de santé"[22].		
Assurer une supervision humaine du fonctionnement du SIA et de ses résultats.		

[20] Cf. 7. Respect de la vie privée.
[21] Cf. 6. Équité
[22] Pour en savoir plus sur cette certification, voir : https://esante.gouv.fr/labels-certifications/hds/certification-des-hebergeurs-de-donnees-de-sante

PRINCIPE D'AUTONOMIE

AUTONOMIE

DEFINITION DU PRINCIPE

Puisant ses racines en Grèce antique, actualisée par les penseurs des Lumières [23] pour devenir une notion centrale en droit, '"auto"-"nomie" est une qualité et une capacité dont l'homme est doté qui lui permet de prendre des décisions librement, c'est à dire, en ayant conscience et en acceptant les conséquences de cette prise de décision. Ainsi, l'humain est capable de se donner sa propre loi et de régler sa conduite sans qu'un autre (maître, tuteur, ordre...) le fasse pour lui.

C'est donc la capacité de **contrôle** qui est en jeu dans l'**autonomie *de l'humain***. Ici, étant donné le processus d'automatisation croissante de l'IA, des tâches auparavant réservées à l'humain sont accomplies par la machine (conception, mémorisation, mécanisation...). C'est la raison pour laquelle la notion de garantie humaine est centrale dans la réglementation sur l'IA et qu'il est demandé aux organisations de mettre en place une chaîne de responsabilité, une gouvernance claire qui facilitent la traçabilité nécessaire pour protéger l'autonomie humaine. Qu'elle soit du côté du fournisseur de la solution qui doit rendre accessible toutes informations sur sa démarche IA que du côté de l'utilisateur qui doit pouvoir prendre connaissance du fonctionnement de l'Outil d'IA avec lequel il interagit.

Selon l'application d'IA développée, il faut ainsi assumer et d'abord questionner, quel degré de délégation permet à l'humain de rester maître et non esclave, des opérations de l'IA : pour en répondre et garder le contrôle, l'IA doit-elle opérer avec une validation humaine ; si oui, comment et à quelles étapes ?

Le choix de ce degré s'effectue au regard de plusieurs critères :

- La finalité de la tâche ou de la mission à laquelle le SIA est soumis ;
- La place de l'opérateur qui interagit avec ce système. Ici, des contraintes légales existent : aucune décision automatique ne peut être réalisée si elle produit un effet juridique (RGPD, loi CNIL notamment) ;
- Le contexte de son déploiement, à risque ou pas à risque. Ce risque s'évalue à travers l'impact que ce SIA a sur le respect des droits fondamentaux[24]. Ici l'autonomie se définit par la négative, par exemple comme faible risque pour la liberté d'expression, contre les particularités religieuses, ou même l'intégrité physique des individus (Cf. Principe 1. Sûreté).

Ce degré de délégation dépend du cas d'usage de l'IA, où les valeurs de l'entreprise mais aussi les exigences de contrôle des consommateurs s'appliqueront avec plus ou moins de force. Si le contexte est trop à risques, il faudra réaliser un compromis quant à la performance du SIA afin de privilégier l'action de l'homme. *A contrario* si le contexte est fermé, plus encadré, donc moins à risque, la performance du SIA pourra être le point majeur du projet IA.

[23] KANT E., *Qu'est-ce que les Lumières*, éd. Flammarion, trad. POIRIER J.F & PROUST F., 2020 (1784), 128 p.
[24] Union Européenne, *Charte des droits fondamentaux*, 18 décembre 2000, ELI : https://www.europarl.europa.eu/charter/pdf/text_fr.pdf

Les questions centrales inhérentes à tous projets d'IA, qui devront être posées avant tout développement sont donc :

Quel niveau de délégation l'homme donne-t-il au SIA ? Et selon quels critères ce niveau est-il déterminé ?

HUMAN IN COMMAND : L'HUMAIN CONÇOIT ET DECIDE

DEFINITION

Le niveau de délégation **est réfléchi** dès la phase de qualification du projet IA **et déterminé au regard de la finalité attendue** du développement et déploiement d'un SIA.

L'approche « *human-in-command* » (HIC) désigne une capacité de mise en place et de contrôle / maîtrise de l'activité globale du SIA (y compris ses incidences économiques, sociales, juridiques et éthiques au sens large) ainsi qu'une faculté de décider quand et comment utiliser le système dans une situation donnée. La décision est humaine tout le long du processus.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Mettre en place une modalité de gestion de projet adaptée aux spécificités de l'IA		
Encadrer ce projet juridiquement (exemple : valider que les interdictions quant à la vie privée et aux droits fondamentaux sont respectées).		
Définir la marge de manœuvre "idéale" de l'humain dans le SIA en posant un cadre éthique : le publier auprès des utilisateurs.		
Définir un processus de reprise en main du SIA en cas de défaillance.		
Mettre en place un processus de détection, d'alerte et de correction des défaillances (non-respect des droits fondamentaux, situation à risque).		
Assister l'utilisateur pour gérer la correction, en cas de détection de défaillance(s).		

HUMAN IN THE LOOP : L'HUMAIN INTERVIENT A DES ETAPES CHOISIES

DEFINITION

Le niveau de délégation est déterminé **au regard du contexte de déploiement du SIA.**

L'approche « *human-in-the-loop* » (HITL) désigne ainsi la capacité d'intervention humaine à une ou plusieurs des étapes du cycle de vie d'un SIA. Ce sont des **points de contrôle humain** dans le cycle de vie d'une IA qui garantissent l'implication et la responsabilisation de l'homme vis-à-vis du SIA. **L'action humaine est donc requise** dans le fonctionnement du SIA.

Cette faculté peut comprendre la décision de ne pas utiliser un SIA dans une situation donnée, de définir des marges d'appréciation pour les interventions humaines lors de l'utilisation du système ou d'ignorer une décision prise par un système. Il convient en outre de veiller à ce que les autorités publiques soient en mesure d'exercer un contrôle conformément à leur mandat.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Étudier les impacts de la mise en œuvre et l'exploitation du SIA.		
Identifier les points de contrôle humain.		
Valider les actions humaines aux points de contrôle par rapport au risque de biais (garder un suivi / historique).		
Garantir la journalisation du suivi du SIA.		
Mettre en place un cahier de test & de recettes permettant de vérifier le niveau d'autonomie attendu.		
Mettre en place une boucle de *feedback* du comportement du SIA pour remonter à l'utilisateur final ce qu'il se passe en vue de corriger le modèle. Exemple : permettre le recours aux utilisateurs lorsque le système prend une décision ou de faire remonter des éventuelles défaillances.		

HUMAN ON THE LOOP : L'HUMAIN SUPERVISE ET N'INTERVIENT PAS

DEFINITION

Le niveau de délégation est déterminé **au regard des tâches remplies par le SIA.**

L'approche « *Human on the Loop* » (HOTL) permet à un SIA **d'opérer sans intervention humaine**. Il existe des situations de rythme extrêmement élevé où l'homme n'a pas la capacité

d'intervenir, au regard de la vitesse d'exécution du SIA. Il existe également des situations où le risque de ne pas respecter les droits fondamentaux est faible.

Cependant, même pour un *chatbot* qui serait *a priori* sans danger pour l'intégrité humaine et les droits fondamentaux, les risques de dérive sont toujours possibles (Cf. 6. Équité). **"Human on the Loop" ne signifie pas "Human OUT of the Loop"** : des comptes-rendus synthétiques du SIA, ou possibilités d'intervenir en cas de défaillance, doivent être réservés comme entrées de validation ou d'urgence dans le SIA.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Qualifier le contexte et les conditions attendues de fonctionnement du SIA.		
Qualifier et évaluer le SIA au regard des droits fondamentaux.		
Garantir le suivi du SIA par des protocoles de conception, d'essai et de suivi approuvés par un processus normatif très strict.		
Expérimenter le SIA afin d'évaluer son fonctionnement et ses impacts. Exemple : s'assurer que l'algorithme ne prend pas de décision automatique ayant des effets juridiques.		
S'assurer que l'humain n'est pas "OUT of the Loop" : synthèses et portes d'entrée d'urgence dans le SIA pour corriger, surveiller, émettre des rapports notamment vers les utilisateurs (Cf. 1. Sûreté).		

L'HUMAIN CONSENT

DEFINITION

Avant de déployer son SIA, l'entreprise doit s'assurer que les consommateurs et plus généralement les citoyens, **consentent à cette automatisation et à sa finalité**. Les obligations du RIA dans le cas spécifique des technologies biométriques s'appliquent en réalité à tous les systèmes, en ce sens qu'elles sous-entendent qu'il y a acceptabilité des personnes impactées par le SIA. Une première condition pour le consentement est d'informer sur les finalités et capacités du SIA (Cf. 5. Explicabilité), aussi bien les concepteurs que les sociétés civiles, afin de s'assurer de leur consentement éclairé (soit au moment de l'utilisation soit au préalable au moment de l'acquisition du SIA par l'utilisateur final).

Les utilisateurs devraient être en mesure de prendre des décisions autonomes éclairées à l'égard des SIA. Ils devraient recevoir les connaissances et les outils pour comprendre les SIA et

interagir avec eux dans une mesure satisfaisante et, autant que possible, être à même de procéder à une évaluation du système ou de le contester. Les SIA devraient aider les individus à prendre de meilleures décisions et à faire des choix plus éclairés en rapport avec leurs objectifs.

Pour cela, il est opportun de remplir un guide d'évaluation afin de :

- Davantage sensibiliser les chefs de projets, les *product owers* et les *data scientists* sur ce sujet central du consentement ;
- Mieux faire saisir la finalité du SIA au consommateur, et plus largement au citoyen ;
- Utilement former tous les salariés sur ce sujet.

L'utilisateur peut être informé sur les conditions du calcul des recommandations du SIA, ce qui renvoie à la communication des conditions de la fiabilité générale du système, à la communication d'informations permettant d'expliquer, interpréter ou justifier une recommandation spécifique, voire de la description des conséquences de la réalisation de la recommandation.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Réaliser un guide d'évaluation du SIA pour l'utilisateur final (consommateur du service). Pour les utilisateurs réfractaires, laisser la possibilité de ne pas être traité par le SIA.		
Monter une action de communication au sujet du guide d'auto-évaluation.		
Définir les conditions de l'accès à ce guide d'auto-évaluation.		
Mettre à disposition une méthode de contestation de l'action du SIA par l'utilisateur.		
Mettre en place un message d'information à destination de l'utilisateur final qui soit visible, intelligible et paramétrable.		

PRINCIPE DE RESPONSABILITE HUMAINE

RESPONSABILITE HUMAINE

DEFINITION DU PRINCIPE

Respondere, répondre, la responsabilité exprime le devoir de répondre de ses actes, toutes circonstances et conséquences comprises, c'est-à-dire d'en assumer l'énonciation, l'effectuation et par suite la réparation voire la sanction lorsque l'attendu n'est pas obtenu[25].

Différents types de responsabilité sont envisageables : responsabilité légale, assurantielle, contractuelle, ou encore éthique.

Si les responsabilités légale, assurantielle et contractuelle s'appliquent en vertu de lois et/ ou de contrats, la responsabilité éthique ou morale est d'abord l'expression de notre capacité à agir et prendre des décisions en autonomie et à motiver ces dernières en se fondant sur des principes moraux.

La responsabilité éthique est engagée dès l'émergence d'un événement indésirable/non-éthique qui serait la conséquence de la mise en œuvre d'un SIA. Elle peut dès lors être engagée - pour la chaîne des acteurs humains – par le non-respect (1) des dispositifs législatifs et réglementaires, (2) de principes éthiques généraux, (3) de méthodologies et procédures réputées/certifiées comme étant éthiques. Par voie de conséquence, la prise en compte du règlement européen sur l'IA a pour effet de modifier le statut même de la responsabilité éthique, la faisant passer de l'expression d'un devoir moral librement consenti, à l'application de principes formellement et légalement définis. Dans ce qui suit, la responsabilité éthique de l'IA s'entend selon ces deux facettes : fondée selon des principes éthiques et structurée légalement.

ASSUMER LES CONSEQUENCES

DEFINITION

La responsabilité d'un acteur de la chaîne de production d'un SIA entraîne la nécessité pour celui-ci d'en assumer les conséquences[26]. Il ne saurait être dégagé de sa responsabilité quand bien même le SIA aurait été certifié par la norme Afnor 42001.

Pour que chacun puisse « assumer les conséquences », il faut donc, dans un premier temps, être en mesure d'établir les responsabilités de chaque acteur de la chaîne de production et d'utilisation d'un SIA.

Le « fournisseur » légal d'un SIA est responsable de sa mise sur le marché[27]. Il est en particulier responsable de la conformité réglementaire, et éthique, du produit qu'il met sur le marché vis-à-vis des dispositifs réglementaires. Dans certains cas comme dans celui d'un SIA embarqué au sein d'un robot, ou d'un SIA collaboratif, la responsabilité du déployeur pourra être engagée et le poste de travail devra lui aussi être conforme.

---

[25] FERGUSON Y., PECTOSE C., LEBLANC A., CRESPIN P., *« L'IA au travail : propositions pour outiller la confiance »*, in Zied Bouraoui, Valérie Camps, Éric Gaussier, Maxime Guériau, Christelle Launois, et al., Actes CNIA 2022 : Conférence Nationale d'Intelligence Artificielle, 2022, hal-03777860v2, p.75 & BACACHE M., *« Intelligence artificielle et droits de la responsabilité et des assurances »*, in Droit de l'intelligence artificielle, dir.BENSAMOUN A. et LOISEAU G., 2ème éd., LGDJ, 2022, p. 79 et s.

[26] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 25 – Responsabilités tout au long de la chaîne de valeur de l'IA

[27] Règlement (UE) 2024/1689 du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle (règlement sur l'intelligence artificielle), cond. 79.2

Au-delà de la conformité, seul l'humain, dans les différentes étapes, de la conception à l'utilisation, est responsable de l'usage d'un résultat issu d'un SIA. Dans un cas de *transfer learning,* le fournisseur du modèle pré-entraîné fait partie de la chaîne des acteurs potentiellement responsables. Dans le cas particulier de l'*Auto-ML*, la responsabilité du fournisseur de la solution auto-apprenante sera engagée.

Il faut alors distinguer les conséquences indésirables prévisibles de celles qui ne le sont pas. Les conséquences prévisibles doivent pouvoir être envisagées et traitées via la mise en place de mesures visant à empêcher leur émergence. Afin d'assumer les conséquences indésirables non prévues, il faut avoir préalablement mis en place l'organisation qui doit permettre, *a posteriori*, de les réparer et de corriger le SIA.

BONNES PRATIQUES

→ PREALABLEMENT A LA MISE EN ŒUVRE D'UN CAS D'USAGE

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Définir une procédure d'identification des responsabilités dans la chaîne de conception-mise en œuvre d'un SIA. Dans le cas du transfer learning, la procédure doit permettre d'identifier les acteurs qui sont intervenus dans le développement et/ou dans le choix du modèle pré-entraîné et du corpus utilisé		
Mettre en place une structure référente de gouvernance des SIA au sein de l'organisation.		
De façon connexe à ce qui existe dans le système de santé, organiser et mettre en place une procédure de déclaration, d'enregistrement et de traitement des événements indésirables. Le système de déclaration doit pouvoir être accessible à tous les acteurs de la chaîne de production jusqu'aux utilisateurs. La structure de gouvernance est responsable de la bonne prise en charge des événements indésirables		

→ POUR CHAQUE CAS D'USAGE

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Nommer un *model owner* qui sera le représentant technique unique des différents acteurs humains impliqués dans le cycle de développement du modèle, en relation avec la structure de gouvernance.		

Définir et catégoriser de façon aussi exhaustive que possible la nature des événements indésirables pouvant être envisagés, i.e. les événements potentiels qui ne respecteraient pas les dispositifs réglementaires en vigueur et/ou les principes éthiques généraux définis.		
Définir - éventuellement de façon contractuelle - les motifs susceptibles d'engager la responsabilité de chaque acteur de la chaîne : du concepteur / du développeur / du fournisseur / de l'hébergeur / des autres sous-traitants / de l'utilisateur final.Il est entendu que cette liste de motifs ne pourra envisager que les conséquences prévisibles.		

CONTROLE ET AUDITABILITE

DEFINITION

L'auditabilité d'un SIA consiste à donner la possibilité à des tiers de réaliser une étude approfondie sur les différentes étapes de conception, le parcours des données utilisées et les configurations de l'algorithme. Ces aspects se doivent d'être totalement transparents et accessibles pour que les auditeurs puissent analyser le fonctionnement, les résultats et les effets, même inattendus[28].

L'audit peut être technique ou fonctionnel. Le volet technique consiste à évaluer la performance du système selon plusieurs critères (fiabilité, exactitude des résultats, etc.). Le volet fonctionnel consiste à étudier ses impacts sur les utilisateurs et à vérifier le respect des différentes valeurs éthiques (équité, souveraineté, etc.).

Ces audits peuvent être réalisés en amont du développement et durant toute la phase de conception, de manière à pouvoir intervenir de façon proactive, en anticipant les anomalies et risques potentiels, et/ou après le développement, requérant d'être réactif, face à des anomalies avérées.

[28] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 10 – Données et gouvernance de données

BONNES PRATIQUES

→ EN AMONT DU DEVELOPPEMENT ET DURANT TOUTE LA PHASE DE CONCEPTION

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Établir une procédure d'audit interne, et si possible externe.		
Préparer le terrain pour la conduite des audits externes, en documentant (par exemple, à l'image du rapport d'audit fondé par la société de Cathy O'Neil [29]) : - le besoin auquel le SIA répond ; - le processus de collecte des données d'entraînement et leur pertinence ; - le choix du modèle et les risques associés ; - les mesures de cybersécurité mises en œuvre ; - les évolutions du modèle une fois le SIA déployé.		
Assurer la protection de la propriété intellectuelle et du secret des affaires sans impacter la transparence.		

[29] O'NEIL C., *Description of Algorithmic Audit: Pre-built Assessments*, https://techinquiry.org/HireVue-ORCAA.pdf, 2020, 8 p.

PRINCIPE D'EXPLICABILITE

EXPLICABILITE

DEFINITION DU PRINCIPE

Par « explicabilité », on entend la capacité à rendre compte d'un programme informatique d'IA et de la décision qui y est associée (diagnostic d'une rétinopathie, détection d'une image d'animal, octroi d'un prêt...) ; c'est-à-dire, à rendre accessible la manière dont il s'est entraîné et les résultats qu'il délivre au moment-même.

Trois étapes successives nous semblent nécessaires pour expliquer l'IA : clarifier en quoi consiste techniquement la « boîte noire » de l'IA, identifier une rationalité dans ce fonctionnement[30], puis adapter l'information à l'utilisateur qui la demande.

Nous avons donc structuré ce principe en trois axes : comment l'IA fonctionne techniquement (transparence), quels sont les possibles arguments qui rapprochent son fonctionnement d'un raisonnement humain (justifiabilité), et les manières de présenter le SIA sous une forme adaptée à un profil d'utilisateur (intelligibilité).

TRANSPARENCE

DEFINITION

La transparence peut être définie par opposition à l'opacité ou « boîte noire » de l'IA dans laquelle « l'expérimentateur a la méconnaissance d'une partie ou de la totalité des processus à l'œuvre qui produisent le phénomène observé »[31].

Pour rendre le cheminement et les résultats de l'IA justifiables (5.2) et intelligibles (5.3), il s'agit d'abord de *connaître* (dans la mesure du possible) et synthétiser les processus informatiques à l'œuvre [32]. L'accès à cette connaissance est nécessaire au développeur pour résoudre les bugs et améliorer son programme, à l'utilisateur pour connaître les conditions de production d'un résultat, à la société pour appréhender les conditions de validité et les limitations d'un programme[33].

Le concepteur doit ainsi remplir un ensemble d'obligations techniques pour rendre son programme plus transparent : exposer et assumer ses choix de conception et de paramétrage ; noter et diffuser les conditions de la validité de ses résultats, les limitations du modèle et ses conditions de validité (capacité de généralisation, types de données...)[34].

Le concepteur doit ainsi remplir un ensemble d'obligations techniques pour rendre son programme plus transparent : exposer et assumer ses choix de conception et de paramétrage ; noter et diffuser les conditions de la validité de ses résultats, les limitations du modèle et ses conditions de validité (capacité de généralisation, types de données...).

---

[30] DENIS C., VARENNE F., *Interprétabilité et explicabilité de phénomènes prédits par de l'apprentissage machine*. Revue Ouverte d'Intelligence Artificielle, Association pour la diffusion de la recherche francophone en intelligence artificielle, 2022, 3 (3-4), pp.287-310

[31] JEBEILE J., *Épistémologie des Modèles et des Simulations numériques – De la Représentation à la Compréhension scientifique*, CNRS Éditions, 2019, p.127

[32] FELZMANN H., FOSCH-VILLARONGA E., LUTZ C., TAMÒ-LARRIEUX A., *Towards transparency by design for artificial intelligence*, Science and Engineering Ethics, 2020, p. 3333-3361

[33] WELLER A., *Challenges for Transparency*, Arxiv :1708.01870v1 [cs.CY] 29 Jul 2017, consulté le 29/07/2022

[34] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 13 – Transparence et fourniture d'informations aux déployeurs

Le RIA ajoute à cette transparence technique une transparence sociale renvoyant à la notification à l'utilisateur humain de son interaction avec une IA (lors d'une interaction directe). Ce souci de transparence sociale s'applique aussi aux SIA permettant de générer des contenus de synthèse de type audio, image, vidéo ou texte, qui doivent rendre ceux-ci « identifiables comme ayant été générés ou manipulés par une IA ». Ces notifications doivent par ailleurs être adressées aux personnes physiques concernées « au plus tard au moment de la première interaction ou de la première exposition ». Ce faire-savoir à l'utilisateur que l'IA est utilisée est évidemment un prérequis à une demande d'explicabilité.

→ CONCERNANT LA CONCEPTION D'UN SIA

Les choix de structures, de paramètres, d'entraînement, doivent être documentés lorsqu'ils sont décisifs pour l'entraînement du modèle [35] (excepté quand les données sont confidentielles, mais dans ce cas leur absence doit être exprimée)

- Au moment de la conception, privilégier les modèles plus simples et plus explicables (e.g. si une régression logistique peut être employée) ou reposant sur des règles (e.g. pense à concevoir des systèmes hybrides [36], où les règles métier introduisent des contraintes intelligibles en IA). Si cela n'est pas possible, choisir celui qui répond le mieux au compromis performance-complexité, c'est-à-dire, analyser le rapport entre la performance et la complexité des modèles possibles pour retenir celui qui minimise la complexité (le fonctionnement interne est plus simple) et qui maximise la performance. Sinon, expliquer le choix par la performance (ou une autre mesure) au détriment de la simplicité du modèle : pourquoi un modèle plus complexe doit être utilisé à la place d'un modèle plus simple. Éviter l'utilisation d'apprentissage-machine, et surtout des modèles plus complexes, pour le seul défi technique ou le seul argument « IA ».
- Documenter les méthodes employées (collecte et étiquetage des données, types d'algorithme, paramètres et hyperparamètres) et l'historique des révisions et changements majeurs qui ont eu lieu (la « traçabilité » [37]). Documenter prépare ainsi le caractère auditable des systèmes et ouvre sur une possible critique et amélioration des modèles en cours.

Conseils pour construire la « notice » du modèle d'IA[38] :

Même si la transparence est la partie la plus technique de l'IA explicable, un tri des informations-clefs est ici nécessaire. Il est donc du devoir éthique du concepteur de présenter ses modèles sous un format clair, à la fois technique et accessible à un agent extérieur disposant de bagages mathématiques et informatiques en IA [39], et lui fournissant les informations spécifiques aux opérations de cette IA et à ses performances.

Néanmoins, comme le fait d'exposer un catalogue complet de moteurs n'informe pas des spécificités d'un moteur particulier, mieux vaut une synthèse des données et évolutions-clefs sur ce qui différencie une IA d'autres modèles et ce qui lui assure des performances optimales dans des limites bien définies.

---

[35] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 11 – Documentation technique
[36] BARREDO ARRIETA A. et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion 58, 2020, p. 100-103
[37] Numeum, *Behind the Codes and the Date, Guide Pratique pour des IA éthiques*, 2021, p.44
[38] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 13 – Paragraphe 3
[39] SOUVERAIN T., *L'IA explicable pour différents utilisateurs – note de lecture*, 2022 (à paraître), Cf. EHSAN U., PASSI S., LIAO Q. V., CHAN L., LEE I., MULLER M., RIEDL M. O., *The who in explainable AI: How ai background shapes perceptions of AI explanations*, arXiv preprint arXiv:2107.13509, 2021

→ LORSQUE LES RESULTATS DE L'IA DOIVENT ETRE SIMPLIFIES

Pour interpréter plus facilement les résultats, notamment à destination des parties commerciales ou utilisateurs finaux, les *data-scientists* empruntent souvent au large champ d'outils de l'"*XAI*"[40]. Si l'IA est un modèle d'apprentissage-machine complexe[41], dont le cheminement complet et détaillé est inaccessible, des outils tels qu'Anchor[42] ou SHAP[43] ou global (SHAP) peuvent ainsi être utilisés pour en reconstituer les traits majeurs, mais *en admettant leurs limites*[44].

Exemple : les utilisateurs doivent être avertis que même si, avec SHAP, il est tentant d'interpréter un score de 0.21 comme une corrélation entre le salaire de Monsieur (u) et un risque de crédit important, il ne permet pas de dire que son salaire cause le refus du prêt à Monsieur (u)[45].

→ SI LE MODELE EVOLUE AVEC SON ENVIRONNEMENT

Informer au moment du changement tout changement de l'IA portant sur ses opérations, sa finalité, ses limitations.

Exemple : "le *chatbot* utilise dorénavant des données de 2022 récoltées dans des États ouest-européens et asiatiques ; vos recommandations d'offres en lignes sont susceptibles d'être modifiées en conséquence"

Si de plus l'IA est en interaction avec un utilisateur humain, celui-ci doit se voir notifier qu'il interagit bien avec une IA[46].

Exemple de la « mise à jour » (Cf. Partie Intelligibilité) : "21 avril, 13:34 – nouvelle version - pour des raisons de performance, votre système de recommandation d'assurances passe d'un modèle de règles logiques à un réseau de neurones"

En cas de réentraînement régulier de l'IA (*continous learning* dans la reconnaissance d'images), prévenir l'utilisateur du moment du changement en fournissant un exemple d'avant / après, et renseigner quelles sont ses nouvelles limites.

Exemple : meilleurs résultats, mais moins bonne précision sur les recherches en ligne de « femmes » occupant des fonctions de direction.

---

[40] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022, April). Explainable AI methods-a brief overview. In xx*AI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers* (pp. 13-38). Cham: Springer International Publishing.

[41] Local Interpretable Model-agnostic Explanations, technique de perturbation du modèle pour approcher de manière linéaire la limite entre les individus classés « 0 » et « 1 ». Cf. RIBEIRO M.T., SINGH S., GUESTRIN C., "*Why should I trust you? : Explaining the predictions of any classifier*", *in*: *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2016, p. 1135–1144.

[42] Cette technique de perturbation du modèle revient à trouver des règles « si… alors » à partir d'une sélection de variables et des groupes d'individus. Cf. RIBEIRO M.T., SINGH S., GUESTRIN C., *Anchors: high-precision model-agnostic explanations*, AAAI Conference on Artificial Intelligence (AAAI), 2018

[43] SHapley Additive exPlanations. Outil construit sur les valeurs de Shapley, technique d'attribution qui quantifie pour chaque variable une influence sur le résultat. Ces valeurs sont estimées pour chaque individu, et peuvent être agrégées au niveau de groupes. Cf. MERRICK L., TALY A., *The explanation game: Explaining machine learning models using shapley values*, International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer, Cham, 2020, pp.17-38. Pour le package, Cf. https://shap.readthedocs.io/en/latest/index.html.

[44] DUCHEMANN B., *Intelligence artificielle et santé publique : normes, savoirs, appropriations et sécurité*, Université de Paris, 2021, p.294-297

[45] KUMAR I., SCHEIDEGGER C., VENKATASUBRAMANIAN S., FRIEDLER S., *Shapley Residuals: Quantifying the limits of the Shapley value for explanations*, Advances in Neural Information Processing Systems, 34, 2021, p. 26598-26608

[46] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 50 – Paragraphe 1

BONNES PRATIQUES

NB : ces bonnes pratiques ouvrent des pistes concrètes pour atteindre des exigences du RIA (Cf. Article 12 « Record keeping » & Article 13 « Transparency and provision of information to the users »)

→ CONCERNANT LA CONCEPTION D'UN SIA

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Privilégier l'algorithme le plus simple statistiquement, reposant si possible sur des règles. Sinon, choisir celui qui répond le mieux au compromis performance (ou autre mesure) et complexité. Ex IAG : si le concepteur arbitre entre plusieurs modèles de fondation à *fine-tuner* ou utiliser, en limiter le nombre de paramètres		
Dans une notice synthétique, permettant de répondre à l'exigence du RIA Article 12.2 sur la traçabilité du SIA et de son fonctionnement, documenter les méthodes employées, l'usage des données, l'historique des changements majeurs e.g. : - Comment les données sont mises en forme ; - Quel modèle a été choisi, suivant quelle méthode ; - Quels hyper paramètres (i.e. fixés par le codeur) et comment ont-ils été choisis ; - En quoi les paramètres (produits de l'entraînement du modèle) donnent des performances différentes ; Quelles sont les conditions de validité des résultats, etc.		
Conception d'IAG et Transfer Learning : si un modèle pré-entraîné est réemployé ou *fine-tuned* (e.g. dans la reconnaissance d'images, pour analyser le risque de mélanome) **Enquêter et communiquer avec le plus de précision possible** sur la nature et les tendances des données d'entraînement et réentraînement. Pour chaque critère, si la source des données est extérieure au concepteur **COCHER** « Demande faite [date] / Informations collectées & éventuelle réponse [date] » : Sur la nature des données collectées en input (géographie, sites, % de données personnelles)... Faire de même avec les inputs qui ont servi à *fine-tuner* le modèle (idem si le concepteur hérite d'un modèle de fondation *déjà fine-tuned*).		

Renseigner périodiquement sur les % de contenus erronés et potentiellement indésirables présents dans les corpus d'entraînement, et ceux générés par le modèle.		

→ LORSQUE LES RESULTATS DE L'IA DOIVENT ETRE SIMPLIFIES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Présenter l'outil d'explicabilité choisi, ses atouts et limites pour le cas d'usage traité.		
Afficher à l'utilisateur les limites techniques du modèle et de l'outil d'explicabilité, avec des exemples individuels.		

→ SI LE MODELE EVOLUE AVEC SON ENVIRONNEMENT

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Donner des exemples de différences de prédictions avant / après, s'il y a évolution continue du programme.		
Notifier que l'humain interagit avec une IA, en cas d'interaction.		
Auprès des utilisateurs, mettre en place des tests périodiques (fixer une régularité) et à grande échelle (fixer un % des données de test) pour écarter les contenus non pertinents, faux, et discriminants (e.g. avec IA générative). Exemple de « chain of thought » : ex de test périodique sur prompt-engineering, questionner régulièrement le modèle sur les données d'entraînement (et examiner leur stabilité) -> tester raisonnement et robustesse de la réponse...		
Prévenir l'utilisateur lorsque de nouvelles données sont utilisées pour ré-entraîner le modèle (e.g. IAG), en notant de la manière la plus précise leurs similitudes / différences avec les inputs de départ (mêmes types de phrases ; mêmes individus... Présence / absence de données personnelles...)		

JUSTIFIABILITE

DEFINITION

Il est possible de comprendre le fonctionnement du modèle suivant une logique (arguments). Le fonctionnement de l'IA est mis en relation avec le cas d'usage où elle opère.

Image du moteur : si la transparence donne accès à l'architecture du moteur, la justifiabilité exige de comprendre comment cette architecture permet à la voiture de rouler.

La justifiabilité suppose un minimum de transparence, c'est-à-dire de compréhension technique du fonctionnement de l'IA. Elle y ajoute une étape supplémentaire : lire dans ce fonctionnement technique des formes de raisonnement. L'humain peut ainsi raisonner pour contester ou valider les résultats de l'IA, gage d'une acceptation fondée en raison[47].

Dès lors, lorsque c'est possible, il s'agit de transposer ce que l'IA fait dans le cas d'usage où un acteur métier le ferait[48]. Par exemple, à partir d'un système hybride associant règles et réseau de neurones, l'IA est spécialisée en plusieurs modules de détection des formes, des coins des oreilles d'un côté, de l'autre du nez ou des formes de la bouche, ce qui lui permet comme un humain le ferait par parcelles de valider la forme d'un visage. Ainsi un fonctionnement transparent, ou hybride, de l'IA, permet de justifier son fonctionnement en l'apparentant à nos catégories humaines[49].

Il s'agit aussi d'être en mesure d'expliciter les limites de cette justification (exemple : si l'outil n'est pas causal, avertir qu'on ne peut pas dire que les liens détectés par l'IA sont causaux[50] ; exemple : si l'outil ne fonctionne pas à partir de règles, admettre qu'une interprétation humaine des liens prédits par l'IA est possible mais non logiquement fondée). Il faut ainsi savoir interpréter les outils techniques à leur juste valeur et s'assurer qu'on ne diffuse pas d'information abusive sur ce que fait le modèle.

Mise en contexte, l'information technique sur l'IA doit aussi montrer en quoi d'éventuels dilemmes éthiques ont été identifiés et traités - et le diffuser ouvertement à tout utilisateur[51]. L'utilisateur sera ainsi en mesure de choisir d'adhérer à l'offre d'IA, et y adhérera d'autant plus que ces choix font écho en lui.

Exemple : ouverture – vie privée, « nous avons choisi le respect de la vie privée en changeant certaines valeurs de variables pour anonymiser les clients, ce pourquoi au niveau individuel de l'explicabilité certains exemples vous paraîtront moins réalistes ».

Exemple : précision statistique – transparence, « nous avons choisi un modèle plus transparent pour éviter d'identifier par erreur des fraudeurs sur des caractéristiques opaques ou discriminantes ».

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?

[47] HÉNIN C., LE MÉTAYER D., *Beyond explainability: justifiability and contestability of algorithmic decision systems,* AI & SOCIETY, 2021, p. 1-14.
[48] MITCHELL M., *Abstraction and analogy-making in artificial intelligence,* Annals of the New York Academy of Sciences, 1505(1), 2021, p. 79-101.
[49] BARREDO ARRIETA A. et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion 58, 2020.
[50] KUMAR I., SCHEIDEGGER C., VENKATASUBRAMANIAN S., FRIEDLER S., *Shapley Residuals: Quantifying the limits of the Shapley value for explanations,* Advances in Neural Information Processing Systems, 2021, 34, p. 26598-26608.
[51] BONNEFON J-F., SHARIFF A., RAHWAN I., *The Moral Psychology of AI and the Ethical Opt-Out Problem*, Oxford University Press, Oxford, UK, 2020, p. 109-126.

Comparer les opérations de l'IA au raisonnement d'un humain dans le même cas d'usage.		
Donner et documenter des exemples des limites de cette comparaison humain / IA (exemple : *machine-learning* vs raisonnement logique).		
Écrire les choix éthiques qui orientent l'IA (exemple : avant tout vie privée, anonymisation des exemples, les individus présentés dans l'explication ne sont pas réels).		
Expliciter les sources. Ex IA générative : durant le *fine-tuning* ou en surcouche d'explication, apporter des sources qui viennent en références et permettent de justifier la réponse de l'IA, tout en rendant intelligible le contenu généré (e.g. prompt) aux utilisateurs visés.		
En particulier pour ces cas de justifiabilité avec une transparence minimale (méthodes post-hoc sur des modèles accessibles seulement via API, modèles pré-entraînés, comptant potentiellement des milliards de paramètres...), ajouter une exigence sur la fiabilité de la technique d'explication ; intervalle de confiance ?		
Dans les cas où la transparence est limitée (e.g. IA générative), donner des preuves de fiabilité de la technique d'explicabilité mobilisée (e.g. robustesse, intervalle de confiance, solutions de preuves *post-hoc*...)		

INTELLIGIBILITE

DEFINITION

Prérequis : comme la justifiabilité (5.2), l'intelligibilité requiert *a minima* la transparence (5.1) du modèle.

L'intelligibilité d'un SIA engage la capacité de rendre son modèle accessible à ses utilisateurs. Elle s'appuie sur un discours de clarification rationnelle du modèle de SIA à destination des parties prenantes. Pour être intelligible, la présentation du SIA doit donc être adaptable : elle doit proposer une famille de réponses qui lui permette de s'adapter à chacun des profils d'utilisateurs : parmi les raisonnements possibles, sont « intelligibles » à un utilisateur donné ceux qui permettent à celui-ci, à partir de ses connaissances économique, technique, juridique, d'acquérir une compréhension des opérations du SIA considéré[52].

L'intelligibilité doit donc être considérée relativement au profil des utilisateurs : équipe de *data-scientists* participant à un projet, donneur d'ordre chez le client, utilisateur, régulateur, etc. A chaque profil d'utilisateur doit correspondre un discours explicatif :

Exemple : si le concepteur technique doit développer une explication de recommandation de produits à partir d'un réseau de neurones à un commercial chargé de vendre l'IA, il

[52] SOUVERAIN T., *Les Enjeux de l'explication en Intelligence Artificielle, appliquée à l'octroi de prêt financier*, Paris, ENS Ulm, 2020.

sélectionnera[53] le vocabulaire (non-technique mais adapté à l'offre commerciale) et les schémas.

Exemple : si l'ingénieur commercial a besoin de convaincre des prospects, le *data-scientist* peut lui soumettre d'utiliser SHAP[54] pour expliquer des résultats individuels. S'il s'agit pour lui de convaincre un régulateur, l'explicabilité globale peut être présentée comme influence de variables non sur toute la population, mais seulement sur le groupe qui a servi de base de calcul.

À noter que lorsqu'il s'agit de comprendre le traitement de données protégées (genre, localisation, …) d'un individu par un site ou moteur recourant à l'IA, l'intelligibilité est une exigence légale en Europe[55] : l'utilisateur est en droit d'accepter et le cas échéant de comprendre l'usage et le traitement qui peut être fait de ses données « privées », transparence accessible sous forme de raisonnement à différents types d'utilisateurs.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Préparer des schémas de vulgarisation selon les profils auxquels expliquer l'IA (*scenarii*-types)[56]. **Exemple** de la « mise à jour », adapté pour un profil non *data-scientist* (Cf. Partie Transparence) : « 21 avril, 13:34 – nouvelle version – pour recommander avec plus de finesse des assurances aux clients, votre système passe de règles métier (prix fixé selon les profils de risque, les niveaux de garanties...) à un réseau de neurones (prix estimé automatiquement par SIA, selon le succès des recommandations passées) »		
Avoir en stock différents supports possibles : interface[57], démo, vocabulaire pour les différentes équipes (formation au métier pour les équipes techniques, et aux techniques pour les équipes métier).		
Tester si le support est perçu comme explicatif chez les utilisateurs visés, en recourant à des méthodes interdisciplinaires pour limiter les biais : philosophie de la connaissance, sociologie, sciences cognitives, spécialistes de l'ergonomie ou de l'UX...[58]		

[53] MILLER T., *Explanation in artificial intelligence: Insights from the social sciences*, Artificial Intelligence n°267, 2018.
[54] Voir le site de l'outil, détectant les variables qui influent sur le résultat : https://shap.readthedocs.io/en/latest/
[55] Règl. (UE) 2016/679 du PE et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données) (Texte présentant de l'intérêt pour l'EEE). ELI: http://data.europa.eu/eli/reg/2016/679/oj
[56] Cf. Exemples dans la partie 5.3.1. Définition
[57] TRIBE D. U. C., *Model reports, a supervision tool for Machine Learning engineers and users*, 2021. Cf. https://pypi.org/project/shapash/
[58] BOVE C., AIGRAIN J., LESOT M.-J., TIJUS C., DETYNIECKI M., *Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users,* 2022, 807-819. 10.1145/3490099.3511139.

Expliciter les sources. Ex IA générative : durant le *fine-tuning* ou en surcouche d'explication, apporter des sources qui viennent en références et permettent de justifier la réponse de l'IA, tout en rendant intelligible le contenu généré (e.g. prompt) aux utilisateurs visés.		

PRINCIPE D'EQUITE

EQUITE

DEFINITION DU PRINCIPE

Nous voyons d'abord l'équité de l'IA sous un angle pragmatique : faire en sorte que l'IA ne soit pas la source de prédictions ou de décisions injustes.

L'iniquité d'un SIA peut être définie de manières très différentes – par exemple : faut-il traiter les individus de la même sorte ? Se focaliser sur une minorité à protéger ? Statistiquement, rendre les résultats indépendants ou non de la caractéristique à protéger comme le genre ?

Nous commençons à cet effet par une rapide synthèse des méthodes de mesure et de correction de l' « injustice » du SIA - i.e. biais que l'utilisateur juge inéquitables, ou par raccourci « biais » (partie 6.1).

L'équité étant portée par l'entreprise d'IA en dehors de la conception seule du système, deux autres aspects motivent à notre sens des mesures concrètes des entreprises d'IA. L'inclusion d'équipes et de visions différentes au sein des équipes d'IA, garantit un socle commun autour des valeurs d'équité (partie 6.2). Si elle ne se pose pas toujours, la question du libre accès du traitement des données, de certains codes et des procédés techniques peut aussi tendre vers une plus large inclusion des utilisateurs comme citoyens (partie 6.3). Si une partie de l'équité en IA est à l'initiative des entreprises, le droit à la transparence et à l'égalité de traitement est également une obligation légale dans certains cas d'usage comme l'accès aux plateformes en ligne[59] - exigence que le RIA tendra probablement à renforcer.

MESURE ET GESTION DES BIAIS

DEFINITION

Même si des outils existent pour éliminer la subjectivité propre à l'humain, ils ne garantissent pas un fonctionnement sans biais (ici, écarts de traitement illégaux ou non désirés) de l'IA. C'est la raison pour laquelle il est essentiel de bien déterminer en amont la finalité du SIA dans son contexte de déploiement, ajustant la place de la décision de l'humain dans son fonctionnement[60]. Penser un SIA comme un système humain-machine permet de prendre en compte les biais et ainsi, de compenser plus justement les dommages qui en découleraient grâce à une interaction en continu entre le SIA et son opérateur. La « garantie humaine », issue du principe d'autonomie, est centrale et structure le RIA.

Partir de l'interaction humain-machine permet de mieux encadrer 2 sortes de biais : les biais mathématiques (composition et complétude du jeu de données) et cognitifs (lors de la conception du système par le programmeur). Les bonnes pratiques ici proposées évoquent ces 2 aspects. Il existe plusieurs mesures pour quantifier les différents types de biais :

- (i) Par individu : des individus similaires doivent être traités de manière similaire. La proximité des individus et de leur traitement est évaluée par des mesures de distance[61,62] ;

[59] Regl. (EU) 2019/1150 du Parlement Européen et du Conseil du 20 juin 2019 sur la promotion de l'équité et de la transparence pour les utilisateurs métiers de services d'intermédiation en ligne, OJ L186/57, 11.07.2019 ('P2B Réglement).
[60] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 10 – Paragraphe 2 – Alinéa f
[61] PHILIPS G. J., DEEPAK V., DIPTIKALYAN S., *Verifying individual fairness in machine learning models*, UAI, volume 124 of Proceedings of Machine Learning Research, AUAI Press, 2020, p. 749–758.
[62] BIEGA A.J., GUMMADI K.P., WEIKUM G., *Equity of attention: amortizing individual fairness in rankings,* SIGIR, 405–414. ACM, 2018.

- (ii) Par groupes (ou sous-groupes[63]) : lorsque la loi ou la sensibilité des utilisateurs définit une variable (« sexe ») sur laquelle les groupes doivent être traités d'une manière semblable.
- (iii) En croisant les appartenances individuelles et à des groupes. Le fait d'être « femme » et d'habiter dans telle région peut, par exemple, renforcer de manière non linéaire l'écart entre genres pour établir des scores de crédit (intersectionnalité[64]).

Contre la discrimination à l'embauche, la loi peut par exemple exiger que la prédiction soit statistiquement indépendante de la caractéristique à protéger (métrique de parité statistique[65]) ; ou bien l'entreprise peut se focaliser sur un type d'erreurs à éviter comme les faux positifs et les faux négatifs (égalité des chances [66]), ou leur ratio (égalité de traitement) parmi nombres de métriques. Des mesures d'équité non statistiques ont été aussi proposées et diffèrent des précédentes en ce qu'elles ne reposent pas entièrement sur des considérations statistiques et prennent en compte des connaissances-métier, ou raisonnent sur des situations hypothétiques en mobilisant par exemple les contrefactuels [67].

Dans l'équité par groupes fréquemment utilisée, les métriques sont souvent incompatibles entre elles [68] ce qui implique pour le praticien d'arbitrer pour en cibler certaines en fonction du cas d'usage.

Pour l'IAG, plus précisément les LLMs (grands modèles de langage), certaines méthodes d'analyse des biais sociaux s'inspirent de l'équité par groupes [69]. Cependant, l'évaluation dépend du type de tâche, du type de biais et, surtout du langage, car ces variables ont un impact sur le choix du corpus et des mesures d'évaluation à utiliser. Par exemple, « CrowS-Pairs » et « StereoSet » sont deux corpus d'évaluation des biais bien connue dans la littérature qui se sont construits sur le paradigme de la paire minimale et qui permettent de quantifier l'équité vis-à-vis le genre, profession et religion, mais sont tous les deux de corpus en anglais. Une version du corpus CrowS-Pairs traduite, adaptée et étendue en français est disponible [70].

Les mesures d'évaluation pour les modèles de langage peuvent être organisées en fonction de leur nécessité (ou non) d'accéder à certaines structures internes du système. Dans ce cas, il existe trois types de métriques : les métriques fondées sur (1) les représentations vectorielles (e.g. *Sentence Encoder Association Test* (SEAT) et *Contextualized Embedding Association Test* (CEAT)), (2) les probabilités (e.g. *Discovery of Correlations* (DisCo) et *Log-Probability Bias Score* (LPBS)) et (3) les sorties ou extrinsèques (e.g. HONEST).

Par exemple, le corpus d'évaluation peut contenir des jetons cachés, obligeant le modèle à prédire quel mot doit être utilisé pour compléter la phrase, puis la présence ou l'absence de mots caractérisant certains groupes sociaux est utilisée pour calculer la métrique d'équité[71].

---

[63] KEARNS M.J., NEEL S., Roth A., WU Z.S., *Preventing fairness gerrymandering: auditing and learning for subgroup fairness,* ICML, volume 80 of Proceedings of Machine Learning Research, PMLR, 2018, p. 2569–2577

[64] BUOLAMWINI, Joy et GEBRU, Timnit. Gender shades: Intersectional accuracy disparities in commercial gender classification. In : Conference on fairness, accountability and transparency. PMLR, 2018. p. 77-91.

[65] DWORK C., HARDT M., PITASSI T., REINGOLD O., ZEMEL R. *Fairness through awareness,* Proceedings of the 3rd innovations in theoretical computer science conference, 2012, p. 214–226

[66] HARDT M., PRICE E., SREBRO N. *Equality of opportunity in supervised learning,* NIPS, 2016,p. 3315–3323

[67] PEARL J., *Causality,* Cambridge university press, 2009

[68] KLEINBERG J.M., MULLAINATHAN S., RAGHAVAN M.: *Inherent Trade-Offs in the Fair Determination of Risk Scores*. ITCS 2017: 43:1-43:23

[69] CHANG, Yupeng, WANG, Xu, WANG, Jindong, et al. A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology, 2023

[70] NÉVÉOL, Aurélie, DUPONT, Yoann, BEZANÇON, Julien, et al. French CrowS-pairs: *Extending a challenge dataset for measuring social bias in masked language models to a language other than English*. In : Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1). 2022. p. 8521-8531.

[71] GALLEGOS, Isabel O., ROSSI, Ryan A., BARROW, Joe, et al. *Bias and fairness in large language models: A survey*. arXiv preprint arXiv:2309.00770, 2023.

La correction du biais peut ensuite intervenir avant, pendant et après entraînement du modèle, c'est-à-dire au niveau :

- (1) Des données (e.g. *Reweighing*[72] et *FairBatch*[73]) - présentant l'avantage de ne pas modifier le modèle, mais pouvant nuire à la transparence sur les données ou aux performances d'un classifieur ;
- (2) De l'entraînement du modèle (e.g. *debiasing with adversarial learning*[74] et *AdaFair*[75]) : moyen aisé d'imposer des contraintes d'équité, ces processeurs tiennent généralement compte de la tension entre l'équité et les performances de classification. Cependant, ils ne peuvent pas être appliqués à n'importe quel SIA puisqu'ils sont généralement spécifiques au modèle ;
- (3) Après entraînement (e.g. *Threshold Optimizer*[76] et SBD (*Shifted Decision Boundary*)[77])- corrigeant le modèle à partir de ses données d'entrée, ce trait les rend plus faciles à mettre en œuvre. Cependant, ces méthodes sont souvent moins performantes que (1) et (2).

Les approches de traitement hybride combinent ces familles de méthodes pour corriger les défauts de chacun (e.g. LRF (*Learning Fair Representations*)[78] et FixOut (*FaIrness through eXplanations and feature dropOut*)[79]). Si l'environnement du cas d'usage évolue, l'impact du SIA sur les nouvelles données doit faire l'objet d'une évaluation attentive.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Identifier le type de biais que l'entreprise souhaite corriger : entre individus, entre groupes ou sous-groupes.		
Une fois les biais analysés par variable, analyser les effets de variables croisées (intersectionnalité, e.g. « sexe » et « âge »).		
Choisir une métrique d'équité correspondant au biais soulevé comme injuste (régulation, culture d'entreprise).		
Choisir à quelle étape rendre l'IA plus « équitable » : au niveau des données (s'il est possible de les changer), pendant l'entraînement (pour une optimisation statistique et métrique		

[72] CALDERS T., KAMIRAN F., PECHENIZKIY M., *Building classifiers with independency constraints,* IEEE International Conference on Data Mining Workshops, IEEE, 2009, p. 13–18
[73] ROH Y., LEE K., WHANG S.E., SUH C., *Fairbatch: Batch selection for model fairness,* International Conference on Learning Representations, 2020
[74] ZHANG B.H., LEMOINE B., MITCHELL M., *Mitigating unwanted biases with adversarial learning,* Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 2018, p. 335–340
[75] IOSIFIDIS V., NTOUTSI E., *Adafair: Cumulative fairness adaptive boosting,* Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, p. 781–790
[76] HARDT M., PRICE E., SREBRO N., *Equality of opportunity in supervised learning,* Advances in neural information processing systems 29, 2016, p. 3315–3323
[77] FISH B., KUN J., LELKES Á. D., *A confidence-based approach for balancing fairness and accuracy*, Proceedings of the 2016 SIAM international conference on data mining, 2016, p. 144-152
[78] ZEMEL R., WU Y., SWERSKY K., PITASSI T., DWORK C., *Learning fair representations,* International Conference on Machine Learning, 2013, p. 325–333
[79] ALVES G., AMBLARD M., BERNIER F., COUCEIRO M., NAPOLI A., *Reducing unintended bias of ML models on tabular and textual data,* The 8th IEEE International Conference on Data Science and Advanced Analytics, 2021

d'équité), après (exemple : s'il n'est ni possible de changer les données, ni le modèle), ou méthode hybride si possible.		
Choisir et justifier le type de transformation équitable à adopter (exemple : quelle méthode pré-, intra-, post-traitement ou de traitement hybride).		
Suivre la mise à jour des données, pour vérifier l'impact du SIA sur les populations à protéger. Pour l'IA générative, renseigner périodiquement sur les % de contenus erronés et potentiellement indésirables présents dans les corpus d'entraînement, et ceux générés par le modèle.		
IAG : auprès des utilisateurs, mettre en place des tests réguliers et à grande échelle pour écarter les contenus non désirés, vecteurs de toxicité, stéréotypes ou de sous-représentation (e.g. dans la manière de parler) de groupes.		
Cibler et mesurer la production de fausses informations selon les groupes ou individus.		

DIVERSITE DES CONCEPTEURS

DEFINITION

Ceux qui conçoivent l'IA (construisent, mais aussi pensent les conséquences pratiques de court et long terme) représentent les différents pôles et minorités de la société[80]. Dans la phase à proprement parler de modélisation, le *data-scientist* devra également s'assurer de la représentativité de ces groupes dans les données (en particulier les minorités de nationalités, âges, sexes, etc.) et du fait que l'IA a un haut niveau de performance sur chacun des groupes. Cet échantillonnage servant de base à l'apprentissage de l'IA fera partie des éléments d'explicabilité et de redevabilité en cas de demande.

Il s'agit donc, à la fois de représenter différents profils dans les équipes de conception, mais aussi que les concepteurs intègrent les attentes d'inclusion des utilisateurs finaux et les enjeux de diversité tout au long de la construction du SIA. Cet aspect de formation répond à une obligation présente dans le projet européen de RIA[81]. A cet effet, sur le modèle de ce que propose la CNIL, un guide d'auto-évaluation des SIA aide à se poser les bonnes questions avant de développer et d'utiliser un SIA[82]. Par exemple, le jeu de données utilisés pour l'entraînement répond-il au principe de minimisation ? Les données utilisées sont-elles représentatives des données observées en environnement réel ?

Exemple de réponse à la question de la diversité des données : si basé sur un échantillonnage, l'outil commode et paramétrable « *submodular pick* » permet de garder une diversité dans

[80] CHOU J., IBARS R., MURILLO O., *In pursuit of inclusive AI*, Inclusive Design, 2018
[81] Règlement - UE - 2024/1689 - EN - EUR-Lex (europa.eu) – Article 4 – Maîtrise de l'IA
[82] Cf. Guide d'auto-évaluation de la CNIL joint dans la partie 6.2.2 sur les bonnes pratiques (à adapter selon le cas d'usage du SIA développé)

l'échantillonnage pour construire un échantillon représentatif du groupe, de ses minorités et de ses différentes visions.

Outil de réflexion sur l'équité d'un SIA appliqué à un cas d'usage précis, un guide intervient comme étape indispensable :

- À la soumission finale des SIA au droit (RGPD, RIA, entre autres) ;
- Une étape de bonne foi, i.e. l'expression d'une vigilance certaine de l'entreprise ou de l'institution aux risques inhérents au développement de ces SIA sur les droits et libertés fondamentaux ;
- Une condition à une responsabilisation concrète et réelle de l'organisation sur ce sujet ;
- Un moyen de valorisation de l'entreprise ou de l'institution vis-à-vis du consommateur final et du citoyen, en ce qu'elle tente de répondre à leurs exigences éthiques.

BONNES PRATIQUES

GUIDE D'AUTOÉVALUATION DES SIA. CNIL[1]

Identifier les risques de biais et les corriger efficacement

Les risques de discriminations liées à l'utilisation d'un algorithme entraîné sur des données biaisées sont largement connus aujourd'hui. En revanche, les facteurs contribuant à ces risques restent mal identifiés et les méthodes pour les corriger sont encore expérimentales.

Il est donc nécessaire d'inspecter rigoureusement l'ensemble des données d'entraînement afin d'y déceler les indices de potentiels biais.

- La méthode utilisée pour la collecte des données d'entraînement est-elle suffisamment connue ?
- Des biais peuvent-ils exister du fait de la méthode utilisée ou des conditions particulières de la collecte ?
- Les données d'entraînement comportent-t-elles des données liées aux caractéristiques particulières des personnes telles que leur sexe, leur âge, leurs caractéristiques physiques, des données sensibles, etc. ?
- Lesquelles ?
- Les hypothèses effectuées sur les données d'entraînement ont-elles été discutées, clairement documentées et confrontées à la réalité ?
- Une étude des corrélations entre ces caractéristiques particulières et le reste des données d'entraînement a-t-elle été effectuée afin d'identifier de possibles proxys ?
- Une étude des biais a-t-elle été effectuée ?
- Selon quelle méthode ?
- Si un biais a été identifié, quelles mesures ont été prises pour le réduire ?

IAG : Enquêter en suivant une démarche analogue au guide d'auto-évaluation de la CNIL, pour communiquer avec la plus grande précision possible sur la qualification du corpus de (ré)entraînement et de ses sources (Cf. Explicabilité -> Transparence).

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
S'assurer que le recrutement tient la diversité des équipes (profils, origines, formations académiques, visions, âge)[83].		
Associer des intervenants internes ou externes spécialistes d'éthique pour former les *datascientists, date engineers et autres data managers* aux aspects éthiques et juridiques du projet[84].		
Pratiquer et publier des auto-évaluations régulières (par exemple semestrielles) de la conformité des travaux des *data-scientists* à la régulation et aux valeurs de l'organisation (exemple : formulaire de la CNIL).		
S'assurer que le guide d'auto-évaluation interroge les choix et documente la méthode de collecte des données, l'étude des corrélations et biais, et les mesures prises face aux biais considérés comme inéquitables (Cf. partie 6.1).		

ACCES OUVERT

DEFINITION

Afin de faciliter l'accès et le développement de l'IA, les acteurs sont invités à partager l'écosystème qui permet de mettre en œuvre la technologie. Ceci est particulièrement important pour permettre aux concepteurs et aux utilisateurs de SIA de mettre en commun les ressources dans le respect des principes éthiques. L'accès est réputé ouvert si, par consultation ou sur simple demande, l'accès à une version de production du SIA, à ses données non confidentielles ou aux parties permettant une meilleure compréhension de celui-ci est accordé.

Une documentation précise sur la finalité du SIA est à cet effet nécessaire au préalable de sa construction, afin de :

- Définir ce qui est acceptable en termes d'ouverture, pendant les phases de conception, de tests, d'utilisation ou de révision du SIA ;
- S'aligner autant que possible avec les bonnes pratiques de l'*open source ;*
- Définir le service d'accès, ce qui peut ou ne peut pas être facturé lors de cet accès ;
- Mettre en œuvre des mesures de protection et de traçabilité des accès.

[83] LI F., DONG H., LIU L., *Using AI to Enable Design for Diversity: A Perspective*, International Conference on Applied Human Factors and Ergonomics, 2020, p. 77-84
[84] MILLER T., *Explanation in artificial intelligence: Insights from the social sciences*. Artificial intelligence, 267, 2019, p. 1-38

BONNES PRATIQUES

Description	Cela concerne mon projet ?	Ai-je pris en compte ce point pour mon projet ? (Si cela concerne mon projet)
Définir et publier les phases du SIA étant en libre accès (doit être géré par les acteurs de la stratégie et de la régulation).		
Définir ce qui est accessible avec / sans coût.		
Sécuriser les parties de code à protéger.		
Donner des accès externes temporaires pour auditer le modèle sur ses parties sensibles (e.g. pour les effets croisés des attributs protégés, une analyse collaborative avec les auditeurs pourrait élargir, consolider et certifier les mesures de biais).		

PRINCIPE DU RESPECT DE LA VIE PRIVEE

RESPECT DE LA VIE PRIVEE

DEFINITION DU PRINCIPE

Le droit au respect de la vie privée et le droit à la protection des données à caractère personnel sont deux droit fondamentaux consacrés, notamment, au niveau européen [85] susceptibles d'être mis en cause par les SIA dès lors que ceux-ci traitent des données à caractère personnel (aussi bien à la conception pour l'entraînement des modèles que lors de l'utilisation des SIA). Les modalités d'application de ces droits fondamentaux sont définies au niveau européen par le RGPD[86], au niveau français par la Loi Informatique et Libertés[87] ainsi que par des dispositions sectorielles spécifiques (telles que le secret médical et le secret des correspondances). Juridiquement, il est donc nécessaire que toutes les étapes du cycle de vie d'un SIA respectent ces textes.

Le droit européen prévoit notamment que les données personnelles doivent être minimisées[88], traitées à des fins déterminées et sur la base du consentement de la personne concernée ou en vertu d'un autre fondement légitime prévu par la loi[89], et que toute personne a le droit d'accéder aux données collectées la concernant[90] et d'en obtenir la rectification[91] ou la suppression[92]. Il est par ailleurs à noter la distinction entre « données à caractère personnel »[93] et « catégories particulières de données à caractère personnel » (ou « données sensibles »)[94]. Le traitement de ces dernières est interdit, sauf exceptions.

De plus, certains SIA emportent des traitements de données personnelles qualifiés juridiquement de traitement à haut risques qui exigent de réaliser une analyse d'impact sur la protection des données [95] - exigences qui seront renforcées par le RIA.

Au-delà des exigences juridiques, le traitement de données à caractère personnel par un SIA soulève des enjeux de confiance. Il est en ce sens important que **les personnes concernées** par le traitement de leurs données aient la **garantie** que celles-ci ne seront pas utilisées à leur encontre à des fins qu'elles n'auraient pas souhaitées. De manière complémentaire, une gouvernance appropriée (c'est-à-dire des mesures techniques et organisationnelles) doit être mise en place afin d'assurer leur qualité, leur intégrité et leur confidentialité.

---

[85] Union Européenne, *Charte des droits fondamentaux*, Articles 7 et 8, 18 décembre 2000, ELI : https://fra.europa.eu/fr/eu-charter
[86] Règl. (UE) n° 2016/679 du PE et du Conseil, 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données (règlement général sur la protection des données), (Texte présentant de l'intérêt pour l'EEE), ELI : https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679&qid=1672317923807&from=EN
[87] L. n° 78-17, 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés, ELI : https://www.legifrance.gouv.fr/loda/id/LEGITEXT000006068624
[88] Article 5(c) du RGPD : Le principe de minimisation des données personnelles impose que seules les données adéquates et nécessaires à la finalité du traitement ne soient collectées/conservées/utilisées.
[89] Article 5 du RGPD.
[90] Article 15 du RGPD.
[91] Article 16 du RGPD.
[92] Article 17 du RGPD.
[93] Article 4 du RGPD : " Toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée «personne concernée») ; est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale"
[94] Article 9 du RGPD : " Donnée à caractère personnel qui révèle l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique"
[95] Article 35 du RGPD (et pour plus de détails, voir EDPB Guidelines https://ec.europa.eu/newsroom/article29/items/611236/en et les fiches de la CNIL https://www.cnil.fr/sites/default/files/atoms/files/liste-traitements-aipd-requise.pdf et https://www.cnil.fr/sites/default/files/atoms/files/liste-traitements-aipd-non-requise.pdf)

ANONYMAT

DEFINITION

Lorsqu'un SIA traite des données à caractère personnel, la question d'anonymiser celles-ci se pose, afin notamment de respecter le principe essentiel de minimisation des données personnelles. Selon la CNIL, l'autorité compétente en matière de protection des données personnelles en France, « l'anonymisation [...] est un traitement de données à caractère personnel qui consiste à utiliser un ensemble de techniques de manière à rendre impossible, en pratique, toute identification de la personne concernée par quelque moyen que ce soit et de manière irréversible »[96]. Il convient de préciser que l'anonymisation est en tant que telle un traitement de données soumis au RGPD et que seules les données anonymisées résultant de ce traitement y échappent.

Dans le contexte d'un traitement ou d'une série de traitements poursuivant une finalité déterminée, en tenant compte des technologies utilisées ou disponibles et raisonnablement susceptibles d'être utilisées, trois critères cumulatifs permettraient de s'assurer qu'un jeu de données est véritablement anonyme :

- Il ne doit pas être possible d'isoler un individu dans le jeu de données (impossibilité d'individualisation) ;
- Il ne doit pas être possible de relier entre eux des ensembles de données distincts concernant un même individu (impossibilité de corrélation) ;
- Il ne doit pas être possible de déduire, de façon quasi certaine, de nouvelles informations sur un individu (impossibilité d'inférence)

Remarque : L'anonymisation doit être bien distinguée de la pseudonymisation qui est un traitement de données personnelles consistant à remplacer les données directement identifiantes (nom, prénom, etc.) d'un jeu de données par des données indirectement identifiantes (alias, numéro séquentiel, etc.), l'opération étant réversible (contrairement à l'anonymisation). **Contrairement aux données anonymisées, les données pseudonymisées restent donc soumises au RGPD.**

Il est à noter que le processus d'anonymisation, éliminant toute possibilité de réidentification, peut parfois conduire à une perte d'informations exploitables. Aujourd'hui d'autres méthodes prometteuses, rassemblées sous le nom de *privacy enhancing technologies*, permettant le respect de la vie privée **dès la conception et par défaut** (ou *privacy by design and by default*), sont en train d'émerger[97].

Il est également à noter que si les techniques d'anonymisation concernent spécifiquement les données à caractère personnel, des techniques similaires peuvent être appliquées afin d'ôter des informations confidentielles de certaines données non personnelles. En effet, certaines données non personnelles sont sensibles par leurs liens avec une activité économique.

[96] L'anonymisation de données personnelles, CNIL, 19 mai 2020, disponible à cette adresse : https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles

[97] *Privacy-enhancing technologies* (PETs), *Draft anonymisation, pseudonymisation and privacy enhancing technologies guidance*, *Information Commissioner's Office, September 2022*, disponible à cette adresse : https://ico.org.uk/media/about-the-ico/consultations/4021464/chapter-5-anonymisation-pets.pdf

L'anonymisation de ces données permettraient de faciliter les échanges d'information entre plusieurs organisations concurrentes.

Il est à noter que la circulation des données non personnelles est aujourd'hui une priorité à l'échelle européenne qui y voit une réelle opportunité industrielle et économique. En ce sens, l'adoption récente du Règlement européen sur la gouvernance des données[98], du Règlement européen sur les données[99] (NB) et du Règlement européen sur les données à forte valeur[100] (NB) vise à structurer un marché unique des données, favorisant ainsi leur réutilisation et au sein duquel leurs conditions d'accès et d'utilisation seraient uniformisées entre les différents Etats membres de l'UE.

BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Analyser si l'anonymisation est bien la méthode permettant de protéger la vie privée par conception et par défaut la plus adaptée à l'usage qui doit être fait des données (il se pourrait en effet que d'autres privacy enhancing technologies soient plus adaptées).		
Identifier les informations à conserver selon leur pertinence et les contraintes techniques de conception.		
Supprimer les éléments d'identification directe ainsi que les valeurs rares qui pourraient permettre une ré-identification aisée des personnes (par exemple, la présence de l'âge des individus peut permettre de ré-identifier très facilement les personnes centenaires).		
Distinguer les informations importantes des informations secondaires ou inutiles (c'est-à-dire supprimables).		
Définir la finesse idéale et acceptable pour chaque information conservée.		

[98] Règl. (UE) n° 2022/868 du PE et du Conseil, portant sur la gouvernance européenne des données et modifiant le règlement (UE) 2018/1724 (règlement sur la gouvernance des données) (Texte présentant de l'intérêt pour l'EEE), ELI : https://eur-lex.europa.eu/legal-content/FR/TXT/?toc=OJ%3AL%3A2022%3A152%3ATOC&uri=uriserv%3AOJ.L_.2022.152.01.0001.01.FRA

[99] Règl. (UE) n°2023/2854 du PE et du Conseil, concernant des règles harmonisées portant sur l'équité de l'accès aux données et de l'utilisation des données et modifiant le règlement (UE) 2017/2394 et la directive (UE) 2020/1828 : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ%3AL_202302854

[100] Règl. d'exécution (UE) n°2023/138 de la Commission, établissant une liste d'ensembles de données de forte valeur spécifiques et les modalités de leur publication et de leur réutilisation : https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32023R0138

Effectuer une veille régulière pour préserver, dans le temps, le caractère anonyme des données produites. Cette veille doit prendre en compte les moyens techniques disponibles ainsi que les autres sources de données qui peuvent permettre de lever l'anonymat des informations.		

NB : les bonnes pratiques ci-dessus sont issues des conseils de la CNIL pour construire un processus d'anonymisation pertinent[101],[102].

CONTROLE ET CONSENTEMENT DANS L'USAGE DES DONNEES

DEFINITION

Les technologies intégrant des SIA qui traitent des données à caractère personnel doivent respecter les lois, règlements et autres normes applicables en matière de protection de ces données. Ces règles ont notamment pour objectif de renforcer l'autonomie des personnes concernées dans la gestion de leurs données ainsi que dans leurs interactions avec des SIA (cf. Principe d'autonomie). A cette fin, le consentement au traitement de leurs données et la transparence concernant le fait qu'elles interagissent avec un SIA constituent des moyens utiles. Ces règles ont par ailleurs pour objectif de garantir aux personnes concernées une effectivité des droits qu'elles doivent pouvoir exercer sur leurs données (voir ci-dessous).

Afin de garantir un niveau de protection des données conforme aux exigences du cadre réglementaire européen, il existe également des règles en matière de stockage et de transfert de données personnelles. La CNIL précise notamment que le transfert de données hors de l'Union européenne (UE) et de l'Espace Économique Européen (EEE) est possible, à condition d'assurer un niveau de protection des données suffisant et approprié. Ces transferts doivent être encadrés en utilisant différents outils juridiques[103].

Remarque : Il est à noter qu'en ce qui concerne les données à caractère non personnel, le respect de la confidentialité de certaines informations (protégées par exemple par le secret de défense nationale, les secrets industriels et commerciaux) est aussi important que le respect de la vie privée des personnes : tout service numérique rendu à des organisations impliquant une collecte de données non personnelles sensibles devrait également permettre à ces organisations une forme de contrôle et d'autonomie concernant la gestion de leurs données. Ces enjeux sont notamment pris en compte dans la stratégie numérique de la Commission Européenne[104] en matière de données et adressés juridiquement par des dispositions du

[101] Sur la base de l'avis 05/14 sur les Techniques d'anonymisation du Groupe de travail "Article 29" sur la protection des données, en date du 10 avril 2014, disponible à l'adresse : https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_fr.pdf

[102] Ces pré-requis permettent de déterminer le procédé d'anonymisation à appliquer, c'est-à-dire l'enchaînement des techniques d'anonymisation à mettre en place. Celles-ci peuvent être regroupées en deux familles : la randomisation et la généralisation.
La randomisation consiste à modifier les attributs dans un jeu de données de telle sorte qu'elles soient moins précises, tout en conservant la répartition globale. Cette technique permet de protéger le jeu de données du risque d'inférence.
La généralisation consiste à modifier l'échelle des attributs des jeux de données, ou leur ordre de grandeur, afin de s'assurer qu'ils soient communs à un ensemble de personnes. Cette technique permet d'éviter l'individualisation d'un jeu de données. Elle limite également les possibles corrélations du jeu de données avec d'autres.

[103] Les règles à respecter en la matière ainsi que les recommandations de la CNIL sont accessibles ici : https://www.cnil.fr/fr/transferer-des-donnees-hors-de-lue

[104] Comm. COM(2020) 66 final de la Commission au PE, au Conseil, au CESE, au CdR, Une stratégie européenne pour les données

Règlement européen sur la gouvernance des données[105] et du Règlement européen sur les données[106].

BONNES PRATIQUES

→ CONCERNANT LE TRAITEMENT DE DONNEES A CARACTERE PERSONNEL PAR LE SIA (LORS DE LA PHASE DE DEVELOPPEMENT ET APRES DEPLOIEMENT)

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Qualifier les traitements de données à caractère personnel opérés par le SIA (haut risque ou non) et, si nécessaire, réaliser une évaluation d'impact.		
S'assurer de la conformité légale et réglementaire des données à caractère personnel traitées par le SIA - en présence d'un délégué à la protection des données au sein de la structure, auquel cette mission doit être confiée, notamment concernant la minimisation des données personnelles.		
Documenter la conformité et maintenir un registre des activités de traitement de données personnelles.		
Faire apparaître de manière accessible, claire et intelligible aux personnes produisant les données qui sont traitées par le SIA les conditions légales et réglementaires de collecte, stockage, traitement et partage.		
Privilégier la base légale du consentement quand cela est possible dans le cas où le SIA collecte des données personnelles concernant ses utilisateurs, sinon informer l'utilisateur de l'utilisation de ses données pour qu'il puisse faire valoir ses droits		
Recueillir ce consentement de manière libre (pas de conséquences négatives en cas de refus et surtout, pas de coercition), spécifique (si plusieurs finalités, demander le consentement pour chacune), éclairée (mentionner les informations exigées par le RGPD de manière compréhensible par toutes les personnes concernées, quel que soit leur âge ou leur niveau d'études) et univoque (acte positif clair. Exemple : pas de case pré-cochées).		

[105] Règl. (UE) n° 2022/868 du PE et du Conseil, portant sur la gouvernance européenne des données et modifiant le règlement (UE) 2018/1724 (règlement sur la gouvernance des données) (Texte présentant de l'intérêt pour l'EEE), ELI : https://eur-lex.europa.eu/legal-content/FR/TXT/?toc=OJ%3AL%3A2022%3A152%3ATOC&uri=uriserv%3AOJ.L_.2022.152.01.0001.01.FRA

[106] Règl. (UE) n°2023/2854 du PE et du Conseil, concernant des règles harmonisées portant sur l'équité de l'accès aux données et de l'utilisation des données et modifiant le règlement (UE) 2017/2394 et la directive (UE) 2020/1828 : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=OJ%3AL_202302854

Définir une durée de conservation des données adéquate aux finalités de traitement et respectueuses des durées légalement définies.		
Ne pas utiliser les données personnelles traitées par le SIA de manière contraire aux intérêts des personnes concernées.		
Ne pas traiter de données sensibles (par exemple données de santé, données biométriques, données génétiques...) sauf cas exceptionnel autorisé par le RGPD.		
Mettre en place des protocoles d'accès aux données personnelles traitées par le SIA, précisant qui peut avoir accès aux données et dans quelles circonstances.		
Permettre aux personnes concernées de retirer leur consentement d'une manière aussi simple que celle par laquelle elles y ont consenti, à tout moment.		
Garantir le droit à l'effacement permettant aux personnes concernées de supprimer les données les concernant.		
Favoriser l'exercice du droit à la portabilité des données personnelles par les personnes concernées.		
Se référer aux lignes directrices du Comité Européen pour la Protection des Données sur le consentement[107],au Guide d'auto-évaluation pour les systèmes d'IA publié par la CNIL[108] et à ses premières recommandations sur le développement des SIA[109] .		

→ CONCERNANT IAG ET TRANSFER LEARNING

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Dans les phases de déploiement, prévenir l'utilisateur lorsque de nouvelles données sont utilisées pour ré-entraîner le modèle (mises à jour des corpus), en notant de la manière la plus précise leurs similitudes / différences avec les inputs de départ (mêmes types de phrases ; mêmes individus... Présence / absence de données personnelles...)		

[107] Lignes directrices 5/2020 sur le consentement au sens du Règlement (UE) 2016/679, Version 1.1, 4 mai 2020, *European Data Protection Board*, disponibles à cette adresse : https://www.cnil.fr/sites/default/files/atoms/files/lignes_directrices_du_cepd_sur_le_consentement.pdf

[108] *Guide d'auto-évaluation pour les systèmes d'intelligence artificielle*, CNIL, Avril 2022, disponible à cette adresse : https://www.cnil.fr/fr/intelligence-artificielle/guide

[109] A l'issue d'une consultation publique, la CNIL a publié ses premières recommandations sur le développement des SIA afin d'aider les professionnels à concilier innovation et respect des droits des personnes pour le développement innovant et responsable de leurs SIA, accessible ici : https://www.cnil.fr/fr/ia-la-cnil-publie-ses-premieres-recommandations-sur-le-developpement-des-systemes-dintelligence#46446

Se référer aux lignes directrices du Comité Européen pour la Protection des Données sur le consentement[110],au Guide d'auto-évaluation pour les systèmes d'IA publié par la CNIL[111] et à ses premières recommandations sur le développement des SIA[112] .		

Enquêter et communiquer avec le plus de précision possible sur la nature et les tendances des données d'entraînement et réentraînement. Pour chaque critère, si la source des données est extérieure au concepteur **COCHER** « Demande faite [date] / Informations collectées & éventuelle réponse [date] » :

- Sur la nature des données collectées en input (géographie, sites, % de données personnelles)... Faire de même avec les inputs qui ont servi à *fine-tuner* le modèle (idem si le concepteur hérite d'un modèle de fondation *déjà fine-tuned*).
- Renseigner périodiquement sur les % de contenus erronés et potentiellement indésirables présents dans les corpus d'entraînement, et ceux générés par le modèle.

→ CONCERNANT L'INTERACTION HOMME-MACHINE

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Anticiper le droit pour les destinataires, en présence d'un SIA utilisé dans le cadre d'un processus de décision, d'obtenir à la place l'intervention d'un humain[113].		
Faire comprendre à l'utilisateur ou au bénéficiaire du SIA qu'il interagit avec un SIA, notamment dans les cas suivants : interaction directe avec un SIA, utilisation d'un SIA pour reconnaître des émotions, utilisation d'un SIA ayant pour finalité de catégoriser des individus selon des critères biométriques, utilisation d'un SIA générant ou manipulant du contenu audio ou visuel[114].		

[110] Lignes directrices 5/2020 sur le consentement au sens du Règlement (UE) 2016/679, Version 1.1, 4 mai 2020, *European Data Protection Board*, disponibles à cette adresse : https://www.cnil.fr/sites/default/files/atoms/files/lignes_directrices_du_cepd_sur_le_consentement.pdf

[111] Guide d'auto-évaluation pour les systèmes d'intelligence artificielle, CNIL, Avril 2022, disponible à cette adresse : https://www.cnil.fr/fr/intelligence-artificielle/guide

[112] A l'issue d'une consultation publique, la CNIL a publié ses premières recommandations sur le développement des SIA afin d'aider les professionnels à concilier innovation et respect des droits des personnes pour le développement innovant et responsable de leurs SIA, accessible ici : https://www.cnil.fr/fr/ia-la-cnil-publie-ses-premieres-recommandations-sur-le-developpement-des-systemes-dintelligence#46446

[113] Article 22 du RGPD.

[114] Comm. COM(2021)206 Proposition de Règlement du PE et du Conseil établissant des règles harmonisées pour l'intelligence artificielle, Article 52

CONTRIBUTEURS

Coordinateurs :

- **Benoît Duchemann**, Althics
- **Thomas Souverain**, France Travail / ENS Ulm
- **Pierre Monget,** Hub France IA

Contributeurs :

- **Guilherme Alves da Silva**, INRIA
- **Charlotte Caseris,** Hormê
- **Bertrand Cassar**, La Poste
- **Miguel Couceiro**, LORIA
- **Ysens de France**, Gendarmerie Nationale
- **Karine Degorre**, XXII
- **Renaud de Mazières,** Groupe Carrousel
- **Etienne de Rocquigny**, Blaise Pascal Advisors
- **Jean-Patrice Glafkidès**, Datavaloris
- **Philippe Hermant**, Entropisme
- **Laetitia Piet**, Centrale Méditerranée
- **Ana Semedo**, IL Expansions
- **Anthéa Serafin**, Ekitia
- **Arnaud Tessalonikos**, Fidal Avocats
- **Anthony Tiangaye**, Fidal Avocats
- **Alya Yacoubi**, Zaion

Comité d'experts :

- **Rodolphe Gelin**, Ampere
- **Diane Galbois-Lehalle,** ICP

L'IA ETHIQUE EN PRATIQUE

OPERATIONNALISER VOTRE SYSTEME D'IA AVEC UNE DEMARCHE ETHIQUE

2ème version - Décembre 2024

HUB FRANCE IA