



HUB
FRANCE
IA

L'IA ETHIQUE EN PRATIQUE

OPERATIONNALISER VOTRE SYSTEME D'IA

AVEC UNE DEMARCHE ETHIQUE

Mai 2023



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

AVANT PROPOS.....	5
ABREVIATIONS.....	7
INTRODUCTION.....	8
COMMENT LIRE CE LIVRE BLANC	10
GRILLE DE MATURITE.....	11
REMERCIEMENTS.....	13
SURETE	15
• ROBUSTESSE.....	15
• FIABILITE DES RESULTATS	16
• PROTECTION PHYSIQUE CONTRE LES IMPREVUS	17
IMPACT DURABLE	20
• EQUILIBRE ENVIRONNEMENTAL	20
• APPORT SOCIO-ECONOMIQUE	22
• IMPACT SUR LA SANTE.....	23
AUTONOMIE	4
• HUMAN IN COMMAND : L'HUMAIN CONCOIT ET DECIDE	5
• HUMAN IN THE LOOP : L'HUMAIN INTERVIENT À DES ETAPES CHOISIES	5
• HUMAN ON THE LOOP : L'HUMAIN SUPERVISE & N'INTERVIENT PAS	6
• L'HUMAIN CONSENT	7
RESPONSABILITE HUMAINE	10
• ASSUMER LES CONSEQUENCES	10



•	CONTROLE ET AUDITABILITE	12
EXPLICABILITE		14
•	TRANSPARENCE	14
•	JUSTIFIABILITÉ	18
•	INTELLIGIBILITE	19
EQUITE		22
•	MESURE ET GESTION DES BIAIS.....	22
•	DIVERSITE DES CONCEPTEURS	24
•	ACCÈS OUVERT	26
RESPECT DE LA VIE PRIVEE		29
•	ANONYMAT.....	30
•	CONTROLE ET CONSENTEMENT DANS L'USAGE DES DONNEES	32
CONTRIBUTEURS.....		35



AVANT PROPOS

Guillaume Avrin, *Coordonnateur national pour l'intelligence artificielle*



Le Hub France IA, acteur fédérateur de l'écosystème national de l'intelligence artificielle (IA), a mis en place dès 2021 un groupe de travail dédié à l'éthique dont ce livre blanc couronne le travail. Il vise à promouvoir un cadre de développement et d'utilisation éthique de l'IA reposant sur des principes simples, formalisés et opérationnels. Nous devons nous en féliciter.

La mission qui m'a été confiée par le Gouvernement a pour objectif premier de diffuser l'IA dans l'économie mais aussi de l'infuser dans le milieu social. Notre stratégie est ainsi fondée, depuis 2018, sur l'impératif de « donner un sens à l'IA », c'est-à-dire canaliser les progrès pour construire un avenir meilleur pour notre société. Cela inclut non seulement de nouvelles perspectives commerciales et de compétitivité industrielle, dont notre

balance d'échanges avec l'extérieur a besoin, mais aussi un développement global, un bien-être individuel et collectif, comme par exemple la recherche d'une émancipation des contraintes du travail et de la production de richesses, une démocratie plus interactive et plus riche, l'accès à plus de savoir et d'expérience, etc. Un mouvement finalement assez banal et qui s'inscrit dans la continuité des progrès de l'humanité depuis cent siècles.

Deux des trois volets de notre programme, qui émerge au plan France 2030, ciblent le contrôle de l'impact de l'IA, l'un au bénéfice de la confiance, l'autre de la frugalité, c'est-à-dire l'économie de moyens dans la perspective de ressources planétaires de plus en plus limitées.

Nous attacherons aussi une attention particulière à la demande, jusqu'ici relativement délaissée, car on ne peut bien sûr espérer de succès commercial ou de déploiement en milieu domestique ou professionnel sans un rapprochement suffisant entre l'offre et la demande. Le Hub France IA a un rôle particulier à jouer sur ce plan en raison de sa position qui le met en mesure d'instaurer un espace de co-construction des exigences de la confiance. Nous comptons sur sa capacité à mobiliser les intermédiaires indispensables à l'établissement de relations clients-fournisseurs sécurisées, offrant suffisamment de garanties aux parties pour libérer les transactions : juristes, assureurs, banquiers, consultants, agences de certification, centres d'essais, etc.

Cela passera aussi sans doute, entre autres dispositifs, par un outil d'évaluation des intelligences renforcé (certainement multisite) et par un regroupement de données capable de couvrir des spectres de fonctionnement étendus, en largeur, en profondeur et en nature.

Nous ferons aussi davantage appel aux sciences humaines et sociales pour maîtriser, tout au long des développements et des innovations, toutes les conséquences humaines et communautaires, comme par exemple la transformation des métiers et le rapport au travail, la question du partage des richesses produites et du rôle du capital, la formation à un environnement automatisé (et peut-être aussi l'automatisation de la formation), l'équilibre géopolitique et l'endiguement de la violence, la désinformation, etc.

Cette feuille de route est donc ambitieuse et mobilisera *a minima* un milliard et demi d'euros de crédits budgétaires sur les cinq prochaines années.



Nous savons que l'IA repose sur des approches et algorithmes d'apprentissage et/ou statistiques qui présentent par construction un risque d'erreur et des fragilités intrinsèques conduisant à des comportements non-convexes et donc difficilement prédictibles. Nous savons aussi que la technologie peut être l'objet de détournements et d'usages malveillants dont les motivations sont tout aussi nombreuses que les convoitises humaines.

Mais, compte tenu des enjeux, que je n'hésiterais pas à qualifier de vitaux pour l'avenir d'une Nation qui n'échappe généralement pas à ses responsabilités, il me paraît tout aussi important de ne pas céder à la peur ou la déraison, les calculatrices n'ont pas encore pris le pouvoir. Nous devons relever le défi de la transformation avant que celle-ci ne s'impose à nous et sous une forme inadaptée. Il ne faut pas perdre de vue l'ampleur de la tâche, dans un contexte extrêmement compétitif, qui nous interdit de nous laisser distraire mais nous oblige au contraire à concentrer nos forces.



ABREVIATIONS

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



INTRODUCTION

Le champ d'étude de l'IA tend à prendre de plus en plus d'importance dans notre vie quotidienne. On le retrouve tant dans nos usages personnels que professionnels. L'arrivée fin 2022 sur la scène publique de technologies telles que l'IA générative a fait prendre conscience de ces enjeux de manière large au grand public. Se posent alors des questions tant techniques et juridiques qu'éthiques pour savoir utiliser de telles technologies dans notre quotidien.

Le groupe de travail éthique du Hub France IA est un groupe actif depuis la création de l'association en 2017. Stratégique et toujours d'actualité au sein des entreprises, 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 ont pris conscience de la nécessité, de plus en plus, d'opérationnaliser ce concept.

Courant 2021, au moment où le RIA est 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 d'utilisation éthique des systèmes d'intelligence artificielle ?

Les experts du groupe de travail ont développé un livre blanc qui se veut être un guide pour aider dans le développement, l'implémentation et le déploiement de SIA au sein d'entreprises, d'administrations ou toutes entités, afin de prendre en compte ces enjeux dès le départ d'un projet pour garantir la protection des droits et faire de notre écosystème un écosystème de confiance.

Pour concrètement associer les SIA à la notion de progrès, c'est à dire à l'épanouissement et au développement de nos sociétés, nous pensons cette technologie à travers ce qu'elle doit être (l'éthique) et ce qu'elle peut être (le droit). Concrètement, les systèmes d'intelligence artificielle devront protéger et garantir la liberté et les droits de chacun tout en cherchant constamment la sécurité de tous et la croissance du marché qu'ils occupent.

Les systèmes d'intelligence artificielle s'inscrivent ainsi dans cette promesse d'équilibre, une promesse formulée par l'UE dans sa « boussole numérique¹ » et par la France à travers « France 2030² ».

L'opérabilité de cette promesse passe ainsi par la compréhension et la mise en application de l'angle à travers lequel elle a été rédigée, l'angle éthique :

Comment garantir que mon IA correspond aux valeurs des citoyens ?

Ce livre blanc a été pensé comme un **guide pratique** qui tente précisément de répondre à cette question. **Unissant les expertises d'académiques et d'industriels de l'IA, il offre au concepteur d'IA 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 déviations 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

¹ https://ec.europa.eu/commission/presscorner/detail/fr/QANDA_21_984

² <https://www.gouvernement.fr/france-2030>



(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 répond 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 d'IA, ce livre blanc donne enfin des pistes à l'entreprise pour s'assurer que son IA promeut un avenir compatible avec nos sociétés et leurs économies, notre santé, notre environnement (impact durable).

Ce livre blanc s'appuie sur des travaux précédemment engagés par les experts du groupe, en particulier le document publié en 2021 par Dreamquark³ avec des ambitions similaires, à savoir, passer du stade de réflexions de l'éthique de l'IA au concret de l'opérationnalisation pour permettre aux entreprises de s'approprier le concept.

Le livre blanc que nous vous proposons présente un double avantage :

- Donner des **définitions simples et claires** de sept premiers principes éthiques de l'IA et de chacune de leurs dimensions ou sous-principes – l'environnement 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).

Les définitions proposées, comme les conseils pratiques, ne sont qu'un début que nous enrichirons de vos retours, et que les exigences légales de plus en plus présentes viendront bientôt préciser - version opérationnelle de l'AI Act à venir, qui imposera des mesures de fiabilité selon le « niveau de risque » ⁴ de l'application d'IA. Elles sont relatives, aussi bien aux expertises des contributeurs qu'à l'état actuel des technologies.

Nous espérons que, partant de ces définitions et recommandations simples, les acteurs de l'IA seront munis d'une balise pour **comprendre et appliquer "l'éthique de l'IA"**. Et de là, remontant au sens originel de l'éthique, que l'IA serve *un monde dans lequel nous voudrions habiter...*

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

³ SOUVERAIN T., MERIC N., *State of Ethical AI in 2021: Challenges posed by Ethics to companies developing AI*, DreamQuark, 2021, 110p. <https://www.dreamquark.com/home-dreamquark-temp/state-of-ethical-ai/>

⁴ European Commission, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*, 21/04/2021.



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 projets. 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 envisager dans votre démarche.

Bonne lecture !



GRILLE DE MATURITE

Adopter une démarche IA éthique est un enjeu tant pour les citoyens que les entreprises et les administrations. Le déploiement d'un tel processus est la garantie du respect des utilisateurs et des bénéficiaires, ainsi que l'adoption d'un rôle de tiers de confiance au sein du marché européen.

À partir de l'ensemble des principes éthiques identifiés et qualifiés par le groupe de travail dédié du Hub France IA, un outil a été réalisé afin de permettre d'identifier les étapes d'avancement dans le processus de déploiement de l'IA éthique ainsi que les points d'amélioration. La grille de maturité ainsi proposée pose un cadre structuré et indicatif, afin d'identifier les principaux jalons pour parvenir à une IA de confiance.

Cette grille est également un levier d'attractivité auprès de partenaires externes, en permettant de fournir une référence commune pour identifier votre niveau de maturité et faire la différence entre un ou plusieurs concurrents. Elle peut également être un moyen de valoriser votre démarche éthique auprès d'acteurs publics, notamment en cas d'appels d'offre ou de procédure de passation d'une commande publique.

Enfin, cette grille de maturité est un moyen de sensibiliser en interne les différents métiers concernés par le déploiement de l'IA éthique. En outre, cette démarche permet d'acculturer l'ensemble des collaborateurs à l'arrivée imminente d'un cadre normatif européen sur l'intelligence artificielle.

Grille de maturité IA et Éthique - mai 2023

	Niveau 1 - Non maîtrisé	Niveau 2 - Initial	Niveau 3 - Défini	Niveau 4 - Managé et largement diffusé	Niveau 5 - Optimisé & Intégré
Sûreté	Vérification des vulnérabilités inexistante ou sous-évaluée	Identification des vulnérabilités limitées aux cas d'application	Identification des vulnérabilités indépendantes des cas d'application et mise en place de standard sur la robustesse, fiabilités des résultats et protection physique contre les imprévus	Approche structurée de la sécurité IA déployée à l'échelle de l'entreprise pour gérer l'ensemble des risques selon le cycle de vie	Au-delà du management des risques de sécurité, l'anticipation des modifications substantielles de l'environnement est une composante essentielle de la stratégie de sécurité et est intégrée à tous les niveaux et processus IA et data de l'entreprise
Impact durable	Les enjeux de l'impact durable ne sont pas pris en considération par l'entreprise	Identification des métriques d'évaluation et sensibilisation par rapport à l'impact durable	Application au niveau des algorithmes des méthodes de prise en compte des impacts pour les optimiser	Prendre en compte les impacts dès la première étape de conception du SIA sur les 3 impacts environnemental, social et santé	Avoir ce pipeline généralisable sur tous les cas d'application
Autonomie	L'autonomie de l'humain n'est pas pensée dans le SIA	Le niveau de délégation n'est regardé que dans certains cas d'application, au regard du contexte de déploiement et des tâches remplies	Il existe des protocoles pour qualifier le contexte de l'IA, le niveau de délégation et le rôle de l'humain. Mais ils ne sont pas appliqués systématiquement à toutes les phases du cycle de vie	Les protocoles pour qualifier le contexte de l'IA, le niveau de délégation, le recueil du consentement de l'utilisateur et le rôle de l'humain sont déployés et intégrés dans tous les projets IA. Ces protocoles sont effectifs et prévoient que l'humain reprenne le contrôle de l'IA en cas de nécessité.	L'humain est au cœur de la stratégie de l'entreprise : au-delà des contrôles nécessaires au contexte, la place de l'homme est intégrée dans le développement de la stratégie IA de l'entreprise et est en ligne avec ses valeurs et propres missions
Responsabilité humaine	La responsabilité humaine n'est pas évaluée et aucune traçabilité n'a été mise en œuvre	Il existe une procédure d'enregistrement et de traitement des événements indésirables	Au-delà des responsabilités légales, il existe une procédure qui clarifie les rôles et responsabilités éthiques, pour chaque acteur de la chaîne de valeur	Il existe un dispositif de management des risques potentiels et avérés qui couvre l'ensemble des acteurs de la chaîne de valeur, le cycle de vie de l'IA ainsi que les données utilisées. Ce processus est maintenu et documenté pour réaliser des audits réguliers.	L'anticipation des responsabilités humaines est intégrée dans l'ADN de l'entreprise et se fait au-delà de l'entreprise dans l'ensemble de l'écosystème. Au-delà des procédures d'audit, de la gouvernance et de la gestion proactive des risques potentiels, il y a une culture des responsabilités humaines (e.g programmes d'éducation permanent, forum de communication, etc.)
Explicabilité	Les principes de transparence, d'explicabilité et d'interprétation sont sous-évalués et peu pris en compte	Certains cas d'application existent où des modèles sont plus transparents et des méthodes d'interprétabilité sont appliquées	Des méthodes sont développées et existent pour expliquer la conception du modèle, sa performance, les résultats obtenus ainsi que les limites techniques et choix techniques. Mais la documentation et l'utilisation de ces méthodes restent hétérogènes.	De solides standards sont systématiquement appliqués pour expliquer la conception du modèle, sa performance, les résultats obtenus ainsi que les limites techniques et choix techniques.	Les principes de transparence, d'explicabilité et d'interprétation font partie intégrante des cycles de conception de l'IA et ils sont revus régulièrement pour anticiper tout risque notamment éthique.
Équité	L'interaction humain machine n'est pas appréhendée, ni aucune appréhension des biais n'est appliquée	Des approches de détection et correction des biais sont déployées pour certains modèles	Lorsque l'IA aboutit à des conséquences négatives pour l'utilisateur, un humain a la possibilité de vérifier les biais et doit valider la décision	Les méthodes de détection de biais sont disponibles et systématiquement utilisées pour tout modèle IA	L'utilisation de méthodes de détection de biais, l'implication d'équipes diverses et la révision régulière des modèles sont systématiques
Respect de la vie privée	Peu d'attention aux aspects de vie privée et/ou confidentialité des données	Seulement des contrôles légalement obligatoires sont appliqués pour chaque cas traité	Toutes les étapes du cycle de vie de la donnée et du SIA respectent les obligations réglementaires dont celles autres que le respect de la vie privée	Mise en place d'une gouvernance afin d'intégrer dans la méthodologie projet des vérifications, des procédures et des comités opérationnels	Mettre en place un écosystème actif lié à la protection des utilisateurs par les données et le SIA. L'entreprise va au-delà de ses obligations réglementaires, de gouvernance et de révision. Et participation externe à la reconnaissance de leur ligne de conduite



REMERCIEMENTS

- **Thierry Berthier**, Enseignant-chercheur expert en cybersécurité – Cyberdéfense
- **Bertrand Cassar**, Responsable gouvernance des données et Docteur en droit du numérique – La Poste Groupe
- **Gauthier Chassang**, Juriste en droit européen et international – Inserm
- **Christophe Denis**, Maître de conférences en Intelligence Artificielle – Sorbonne Université, Laboratoire d'Informatique LIP6
- **Sonia Desmoulin-Canselier**, Chargée de recherche – CNRS – Université Paris 1 Panthéon-Sorbonne
- **Yann Ferguson**, Enseignant-chercheur – Icam, Expert – Partenariat Mondial pour l'Intelligence Artificielle (PMIA), Responsable scientifique – LaborIA
- **Karën Fort**, Maîtresse de conférences – Sorbonne – LORIA
- **Diane Galbois-Lehalle**, Maître de conférences et Directrice du Master Droit de l'IA – ICP
- **Rodolphe Gelin**, Expert *Deep Learning and Robotics* – Renault



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*.

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.⁵

⁵ 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 <i>checklist</i> 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.		

• 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. 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.

Reproduire les résultats nécessiterait donc une grande capacité de calcul due aux nombres de paramètres, et au nombre des combinaisons possibles⁶.

⁶ 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



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.		

• 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. 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.

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



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...).		
Permettre le passage en mode " <i>alternate</i> ", dans lequel le système donne la main à l'humain.		
Identifier des exemples d'impact direct physique lorsque le SIA est intégré dans le <i>hardware</i> (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.		
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 techniques par ses automatismes, porteurs d'applications rapides et impactant à large échelle (des *chats* informatifs aux opérations médicales, en passant par les réseaux sociaux).

Il est donc du **devoir des entreprises d'IA** de voir plus loin, en imaginant **quels futurs son SIA ouvre**. Les applications d'IA doivent aider à construire un **futur désirable** par l'entreprise, ses consommateurs et les citoyens.

Un SIA à impact durable s'inscrit dans un projet prenant en considération les enjeux environnementaux et participant à la transition écologique, par exemple en contribuant aux objectifs de développement durable¹⁰. La définition de référence de cette notion est donnée par le Rapport Brundtland, qui la définit comme un mode de développement qui répond aux besoins du présent sans compromettre la capacité des générations futures de répondre aux leurs¹¹. Il s'agit donc de développer chaque SIA en mettant en balance les avantages et les inconvénients qu'il pourrait engendrer pour l'environnement, la société, l'économie et la santé des générations actuelles et futures, et de ne développer que ceux dont les avantages seraient supérieurs aux inconvénients.

• 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.

Tout d'abord, il convient donc de minimiser le coût environnemental de la chaîne de production des matériaux nécessaires à la création de SIA, en favorisant par exemple les circuits courts. Vient ensuite la nécessité de minimiser le coût environnemental du développement et du fonctionnement des SIA, en réduisant l'énergie nécessaire pour cela – vers une IA frugale. Enfin, afin de compenser le coût environnemental de la chaîne de production, du développement et du fonctionnement, il serait intéressant de favoriser des usages à impact positif sur l'environnement, par exemple en favorisant des usages qui contribuent à la réalisation des Objectifs de Développement Durable définis par les Nations Unies à l'horizon 2030 précédemment évoqués.

¹⁰ 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/>

¹¹ Commission mondiale sur l'environnement et le développement de l'ONU, *Rapport Brundtland : Notre avenir à tous*, 1987, disponible à cette adresse : <https://www.are.admin.ch/are/fr/home/media-et-publications/publications/developpement-durable/brundtland-report.html>

**BONNES PRATIQUES¹²**

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Vérifier que le besoin auquel le SIA répond soit une action positive pour l'environnement.		
Prendre en compte les enjeux environnementaux dans le choix des fournisseurs, prestataires et sous-traitants impliqués dans le cycle de vie du SIA.		
Mesurer la consommation carbone, en particulier lors de l'entraînement de modèles (des outils comme Carbon tracker, bien qu'imparfaits, permettent d'aider). Cf. GT Environnement du Hub France IA ¹³		
Minimiser la consommation de ressources énergétiques, matérielles et logicielles par le SIA au cours de son cycle de vie et minimiser la complexité de l'architecture nécessaire à son fonctionnement.		
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).		
Se concentrer sur les fonctionnalités essentielles et sur le besoin initial auquel est censé répondre le SIA et, au cours de son cycle de vie, supprimer celles qui ne sont pas utilisées.		
Minimiser le volume de données récupérées au cours du cycle de vie du SIA (notamment après avoir réalisé l'apprentissage) 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.		

¹² AFNOR SPEC 2201. *Ecoconception des services numériques*. Z77-102-0. Avril 2022, 58 p.

¹³ Les travaux du groupe de travail Environnement sont accessibles à l'URL : <https://www.hub-franceia.fr/groupe-de-travail/>



• 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 permettent d'évaluer l'apport socio-économique d'un SIA. Tout d'abord, les usages d'un SIA devrait contribuer à un bien-être collectif, à travers la prospérité des individus et la réalisation d'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.

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.		
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 les (futurs) utilisateurs du SIA dans son développement et/ou son suivi.		
Évaluer l'accessibilité de l'interface utilisateur du SIA ¹⁴ et l'améliorer en continu.		
Mettre en place une stratégie de compatibilité du SIA avec les terminaux et versions logicielles obsolètes.		

¹⁴ DINUM, *Référentiel Général d'Amélioration de l'Accessibilité*, version 4.1, Décembre 2020, 128 p.



• IMPACT SUR LA SANTE

DEFINITION

Les SIA peuvent avoir un impact sur la santé des individus. Cet impact sur la santé 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 normes en vigueur, notamment afin de permettre une protection effective des données personnelles.

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.

BONNES PRATIQUES

→ 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, il n'y a pas à notre connaissance d'outils ou de méthodes permettant d'anticiper 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**¹⁵

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 ¹⁶ .		
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.		
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) ¹⁷ .		
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) ¹⁸ .		
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.		
Héberger les données de santé traitées par le SIA dans des centres de données certifiés "Hébergeurs de données de santé" ¹⁹ .		
Assurer une supervision humaine du fonctionnement du SIA et de ses résultats.		

¹⁵ 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>

¹⁶ 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>.

¹⁷ Cf. 7. Respect de la vie privée.

¹⁸ Cf. 6. Équité

¹⁹ 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 Lumières²⁰ 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...).

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 fondamentaux²¹. Ici l'autonomie se définit par la négative, par exemple comme absence de risque pour la liberté d'expression, contre les particularités religieuses, ou même l'intégrité physique des individus (Cf. Principe 1. Sécurité).

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.

Tout projet de SIA se construit autour de la notion de "garantie humaine".

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 ?

²⁰ KANT E., *Qu'est-ce que les Lumières*, éd. Flammarion, trad. POIRIER J.F & PROUST F., 2020 (1784), 128 p.

²¹ Union Européenne, Charte des droits fondamentaux, 18 décembre 2000, ELI : https://www.europarl.europa.eu/charter/pdf/text_fr.pdf



• HUMAN IN COMMAND : L'HUMAIN CONÇOIT ET DÉCIDE

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 ?
Se former à la gestion de projet 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 consommateurs.		
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 À DES ÉTAPES 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 système d'IA 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 <i>feedback</i> 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 & 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 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. Sécurité).		

• 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 de l'*AI Act* dans le cas spécifique des technologies biométriques s'étend en réalité à tous les systèmes, en ce sens qu'elle sous-entend 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.

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 *data scientists* sur ce sujet central du consentement ;
- Mieux saisir la finalité du SIA au consommateur 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 système d'IA, 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).		
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.		



LIVRE BLANC :

L'IA Éthique en pratique – Opérationnaliser votre système d'IA avec une démarche éthique

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 obtenu²².

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 projet européen de réglementation de 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.

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.

D'un point de vue légal, le « fournisseur » légal d'un SIA est responsable de sa mise sur le marché²³, il est en particulier responsable de la conformité 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 de l'utilisateur pourra être engagée et le poste de travail devra lui aussi être conforme.

²² FERGUSON Y., PECTOSE C., LEBLANC A., CRESPIEN 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

²³ European Commission, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*, 21/04/2021, Titre 53, p.32



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.		
Organiser et mettre en place une procédure d'enregistrement et de traitement 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 <i>model owner</i> qui sera « responsable » du modèle en production et sera le représentant unique des différents acteurs humains impliqués dans le cycle de développement.		
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.		



• 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.

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.

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 ²⁴) : - 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.		

²⁴ 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 fonctionnement²⁵, 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 système d'IA 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 »²⁶.

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 ²⁷. 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 programme²⁸.

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...).

L'AI Act²⁹ ajoute à cette transparence technique une transparence sociale renvoyant à la notification à l'utilisateur humain de son interaction avec une IA (encore au stade de

²⁵ 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

²⁶ JEBEILE J., *Épistémologie des Modèles et des Simulations numériques – De la Représentation à la Compréhension scientifique*, CNRS Éditions, 2019, p.127

²⁷ FELZMANN H., FOSCH-VILLARONGA E., LUTZ C., TAMÖ-LARRIEUX A., *Towards transparency by design for artificial intelligence*, *Science and Engineering Ethics*, 2020, p. 3333-3361

²⁸ WELLER A., *Challenges for Transparency*, Arxiv :1708.01870v1 [cs.CY] 29 Jul 2017, consulté le 29/07/2022

²⁹ European Commission, *Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts*, 21/04/2021, Titre 70, p.34



proposition de règlement, l'exigence est en cours de définition) – faire connaître à l'utilisateur que l'IA est utilisée, condition nécessaire pour qu'il demande à l'expliquer.

Cette prise en charge de la transparence peut être réalisée en considérant le SIA lors de sa conception technique, en interaction avec un utilisateur ou encore en évolution au sein de son environnement.

→ 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 (excepté quand les données sont confidentielles, mais dans ce cas leur absence doit être exprimée)

- Au moment de la conception, choisir l'algorithme le plus simple (si une régression logistique peut être employée sans trop de perte de performance pour un modèle de score de crédit, ne pas employer d'apprentissage-machine pour le seul défi technique ou le seul argument « IA »), reposant si possible sur des règles (exemple : penser à concevoir des systèmes hybrides³⁰, où les règles métier introduisent des contraintes intelligibles en 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 »³¹). 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 :

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 ³², 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.

→ 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

³⁰ BARREDO ARRIETA A. et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion 58, 2020, p. 100-103

³¹ Numeum, *Behind the Codes and the Data, Guide Pratique pour des IA éthiques*, 2021, p.44

³² 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



d'outils de l'«XAI»³³. Si l'IA est un modèle d'apprentissage-machine complexe³⁴, dont le cheminement complet et détaillé est inaccessible, des outils tels qu'Anchor³⁵ ou SHAP³⁶ ou global (SHAP) peuvent ainsi être utilisés pour en reconstituer les traits majeurs, mais *en admettant leurs limites*³⁷.

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)³⁸.

→ SI LE MODELE EVOLUE AVEC SON ENVIRONNEMENT

Informez 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.

Exemple : «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 (*continuous 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.

³³ Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022, April). Explainable AI methods-a brief overview. In *xxAI-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.

³⁴ 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.

³⁵ 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

³⁶ 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>.

³⁷ DUCHEMANN B., *Intelligence artificielle et santé publique : normes, savoirs, appropriations et sécurité*, Université de Paris, 2021, p.294-297

³⁸ 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



BONNES PRATIQUES

→ CONCERNANT LA CONCEPTION D'UN SIA

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
Choisir l'algorithme le plus simple, reposant si possible sur des règles.		
Documenter les méthodes employées, l'usage des données, l'historique des changements majeurs : - 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. En faire une notice synthétique .		

→ 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.		



• JUSTIFIABILITÉ

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 raison³⁹.

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 ferait⁴⁰. 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 humaines⁴¹.

Il s'agit aussi d'être en mesure d'explicitier 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 causaux⁴² ; 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 utilisateur⁴³. L'utilisateur sera ainsi en mesure de choisir d'adhérer à l'offre d'IA, et y adhèrera 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 ».

³⁹ HÉNIN C., LE MÉTAYER D., *Beyond explainability: justifiability and contestability of algorithmic decision systems*, AI & SOCIETY, 2021, p. 1-14.

⁴⁰ MITCHELL M., *Abstraction and analogy-making in artificial intelligence*, Annals of the New York Academy of Sciences, 1505(1), 2021, p. 79-101.

⁴¹ BARREDO ARRIETA A. et al., *Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI*, Information Fusion 58, 2020.

⁴² 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.

⁴³ 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.



BONNES PRATIQUES

Description	Cela concerne-t-il mon projet ?	Si oui, ai-je pris en compte ce point ?
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 : <i>machine-learning</i> 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).		

• 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 considéré⁴⁴.

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 sélectionnera⁴⁵ 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 SHAP⁴⁶ pour expliquer des résultats individuels. S'il s'agit pour lui de convaincre un régulateur, l'explicabilité globale peut être présentée comme influence de

⁴⁴ SOUVERAIN T., *Les Enjeux de l'explication en Intelligence Artificielle, appliquée à l'octroi de prêt financier*, Paris, ENS Ulm, 2020.

⁴⁵ MILLER T., *Explanation in artificial intelligence: Insights from the social sciences*, Artificial Intelligence n°267, 2018.

⁴⁶ Voir le site de l'outil, détectant les variables qui influent sur le résultat : <https://shap.readthedocs.io/en/latest/>



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 Europe⁴⁷ : 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 (<i>scenarii-types</i>) ⁴⁸ .		
Avoir en stock différents supports possibles : interface ⁴⁹ , démo, vocabulaire pour les différentes équipes (formation au métier pour les équipes techniques, et aux techniques pour les équipes métier).		
Tester si mon support est perçu comme explicatif chez les utilisateurs visés ⁵⁰ .		

⁴⁷ 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>

⁴⁸ Cf. Exemples dans la partie 5.3.1. Définition

⁴⁹ TRIBE D. U. C., *Model reports, a supervision tool for Machine Learning engineers and users*, 2021. Cf. <https://pypi.org/project/shapash/>

⁵⁰ 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.



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 ligne⁵¹ - exigence que l'*AI Act* 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. Penser les 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 » est centrale et structure le futur *AI Act*.

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 :

⁵¹ 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 Règlement).



- (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^{52,53} ;
- (ii) Par groupes (ou sous-groupes⁵⁴) : lorsque la loi ou la sensibilité des utilisateurs définit une variable (« sexe ») sur laquelle les groupes doivent être traités d'une manière semblable.

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 statistique⁵⁵) ; ou bien l'entreprise peut se focaliser sur un type d'erreurs à éviter comme les faux positifs et les faux négatifs (égalité des chances⁵⁶), 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 contrefactuels⁵⁷.

Dans l'équité par groupes fréquemment utilisée, les métriques sont souvent incompatibles entre elles⁵⁸ ce qui implique pour le praticien d'arbitrer pour en cibler certaines en fonction du cas d'usage.

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. *Reweighting*⁵⁹ et *FairBatch*⁶⁰) - 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*⁶¹ et *AdaFair*⁶²) : 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*⁶³ et SBD (*Shifted Decision Boundary*)⁶⁴)- 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).

⁵² 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.

⁵³ BIEGA A.J., GUMMADI K.P., WEIKUM G., *Equity of attention: amortizing individual fairness in rankings*, SIGIR, 405–414. ACM, 2018.

⁵⁴ 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

⁵⁵ 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

⁵⁶ HARDT M., PRICE E., SREBRO N., *Equality of opportunity in supervised learning*, NIPS, 2016, p. 3315–3323

⁵⁷ PEARL J., *Causality*, Cambridge university press, 2009

⁵⁸ KLEINBERG J.M., MULLAINATHAN S., RAGHAVAN M., *Inherent Trade-Offs in the Fair Determination of Risk Scores*, IJCAI 2017: 43:1–43:23

⁵⁹ CALDERS T., KAMIRAN F., PECHENIZKIY M., *Building classifiers with independency constraints*, IEEE International Conference on Data Mining Workshops, IEEE, 2009, p. 13–18

⁶⁰ ROH Y., LEE K., WHANG S.E., SUH C., *Fairbatch: Batch selection for model fairness*, International Conference on Learning Representations, 2020

⁶¹ 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

⁶² 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

⁶³ HARDT M., PRICE E., SREBRO N., *Equality of opportunity in supervised learning*, Advances in neural information processing systems 29, 2016, p. 3315–3323

⁶⁴ 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



Les approches de traitement hybride combinent ces familles de méthodes pour corriger les défauts de chacun (e.g. LRF (*Learning Fair Representations*)⁶⁵ et FixOut (*Fairness through eXplanations and feature dropOut*)⁶⁶). 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.		
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 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.		

• 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 société⁶⁷. 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.

⁶⁵ ZEMEL R., WU Y., SWERSKY K., PITASSI T., DWORK C., *Learning fair representations*, International Conference on Machine Learning, 2013, p. 325–333

⁶⁶ 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

⁶⁷ CHOU J., IBARS R., MURILLO O., *In pursuit of inclusive AI*, Inclusive Design, 2018



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 future obligation présente dans le projet européen de réglementation de l'IA. 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 SIA⁶⁸. 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 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, AI ACT à venir, 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'entreprise 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-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 ?

⁶⁸ 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)



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) ⁶⁹ .		
Associer des intervenants internes ou externes spécialistes d'éthique pour former les <i>data-scientists</i> aux aspects éthiques et juridiques du projet ⁷⁰ .		
Pratiquer et publier des auto-évaluations régulières (par exemple semestrielles) de la conformité des travaux des <i>data-scientists</i> à la régulation et aux valeurs de l'entreprise (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).		

• ACCÈS 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.

⁶⁹ 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

⁷⁰ 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.		



PRINCIPE DU RESPECT DE LA VIE PRIVEE



RESPECT DE LA VIE PRIVÉE

DEFINITION DU PRINCIPE

Le droit au respect de la vie privée et le droit à la protection des données à caractère personnel sont deux droits fondamentaux consacrés au niveau européen⁷¹ susceptibles d'être mis en cause par les SIA dès lors que ceux-ci traitent des données à caractère personnel. Les modalités d'application de ces droits fondamentaux sont définies au niveau européen par le RGPD⁷², au niveau français par la Loi Informatique et Libertés⁷³ 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 minimisées⁷⁴, 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 loi⁷⁵, et que toute personne a le droit d'accéder aux données collectées la concernant⁷⁶ et d'en obtenir la rectification⁷⁷ ou la suppression⁷⁸. Il est par ailleurs à noter la distinction entre « données à caractère personnel »⁷⁹ et « catégories particulières de données à caractère personnel » (ou « données sensibles »)⁸⁰. 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 données⁸¹ - exigences qui seront renforcées par le futur *AI Act*.

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.

⁷¹ Union Européenne, *Charte des droits fondamentaux*, Articles 7 et 8, 18 décembre 2000, ELI : <https://fra.europa.eu/fr/eu-charter>

⁷² 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>

⁷³ 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>

⁷⁴ 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.

⁷⁵ Article 5 du RGPD.

⁷⁶ Article 15 du RGPD.

⁷⁷ Article 16 du RGPD.

⁷⁸ Article 17 du RGPD.

⁷⁹ 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 "

⁸⁰ 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 "

⁸¹ 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 »⁸². 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'émerger⁸³.

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.

⁸² L'anonymisation de données personnelles, CNIL, 19 mai 2020, disponible à cette adresse : <https://www.cnil.fr/fr/lanonymisation-de-donnees-personnelles>

⁸³ *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. La réglementation européenne est en train d'évoluer avec la mise en place du *Data Governance Act*⁸⁴ visant à encadrer les échanges de données non personnelles non anonymisées, notamment grâce à l'apparition de nouveaux acteurs comme les intermédiaires de données et les organismes altruistes en matière de données.

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 <i>privacy enhancing technologies</i> soient plus adaptées).		
Identifier les informations à conserver selon leur pertinence.		
Supprimer les éléments d'identification directe ainsi que les valeurs rares qui pourraient permettre un 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.		
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.		
Instaurer un cadre juridique visant à encadrer l'usage des données traitées par le SIA, même si celles-ci sont anonymisées.		

NB : les bonnes pratiques ci-dessus sont issues des conseils de la CNIL pour construire un processus d'anonymisation pertinent^{85, 86}.

⁸⁴ 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

⁸⁵ 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

⁸⁶ 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.



• 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. 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 juridiques⁸⁷.

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 Européenne⁸⁸, à travers la publication de la proposition *Data Act*⁸⁹ et du *Data Governance Act*⁹⁰.

⁸⁷ Les règles à respecter en la matière ainsi que les recommandations de la CNIL sont accessibles ici : <https://www.cnil.fr/fr/transférer-des-donnees-hors-de-lue>

⁸⁸ Comm. COM(2020) 66 final de la Commission au PE, au Conseil, au CESE, au CdR, Une stratégie européenne pour les données

⁸⁹ Comm. COM(2022) 68 final de la Commission au PE, au Conseil, au CESE, au CdR, Proposition de règlement sur les règles harmonisées concernant l'accès équitable aux données et leur utilisation (loi sur les données)

⁹⁰ 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%3AOJL_2022.152.01.0001.01.FRA

**BONNES PRATIQUES****→ CONCERNANT LE TRAITEMENT DE DONNEES A CARACTERE PERSONNEL PAR LE SIA**

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.		
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 dans le cas où le SIA collecte des données personnelles concernant ses utilisateurs.		
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).		
Définir une durée de conservation des données adéquate aux finalités de traitement.		
Ne pas utiliser les données personnelles traitées par le SIA de manière contraire aux intérêts des personnes concernées.		
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 ⁹¹ et au Guide d'auto-évaluation pour les systèmes d'IA publié par la CNIL ⁹² .		

→ 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 ⁹³ .		
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 ⁹⁴ .		

⁹¹ 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

⁹² 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>

⁹³ Article 22 du RGPD.

⁹⁴ Comm. COM(2021)206 Proposition de Règlement du PE et du Conseil établissant des règles harmonisées pour l'intelligence artificielle, Article 52



CONTRIBUTEURS

Coordinateurs :

- **Pierre Monget**, Hub France IA
- **Camille Souillart**, Hub France IA

Contributeurs :

- **Bertrand Cassar**, La Poste
- **Miguel Couceiro**, LORIA
- **Ysens de France**, Gendarmerie Nationale
- **Benoît Duchemann**, Althics
- **Jean-Patrice Glafkidès**, Datavaloris
- **Thomas Gomez**, Agdatahub
- **Philippe Hermant**, Entropisme
- **Christine Laüt**, Safe AI Now
- **Anthéa Serafin**, Ekitia
- **Thomas Souverain**, DreamQuark & ENS Ulm
- **Alya Yacoubi**, Zaion



L'IA ETHIQUE EN PRATIQUE
OPERATIONNALISER VOTRE SYSTEME D'IA
AVEC UNE DEMARCHE ETHIQUE

Mai 2023



HUB
FRANCE
IA