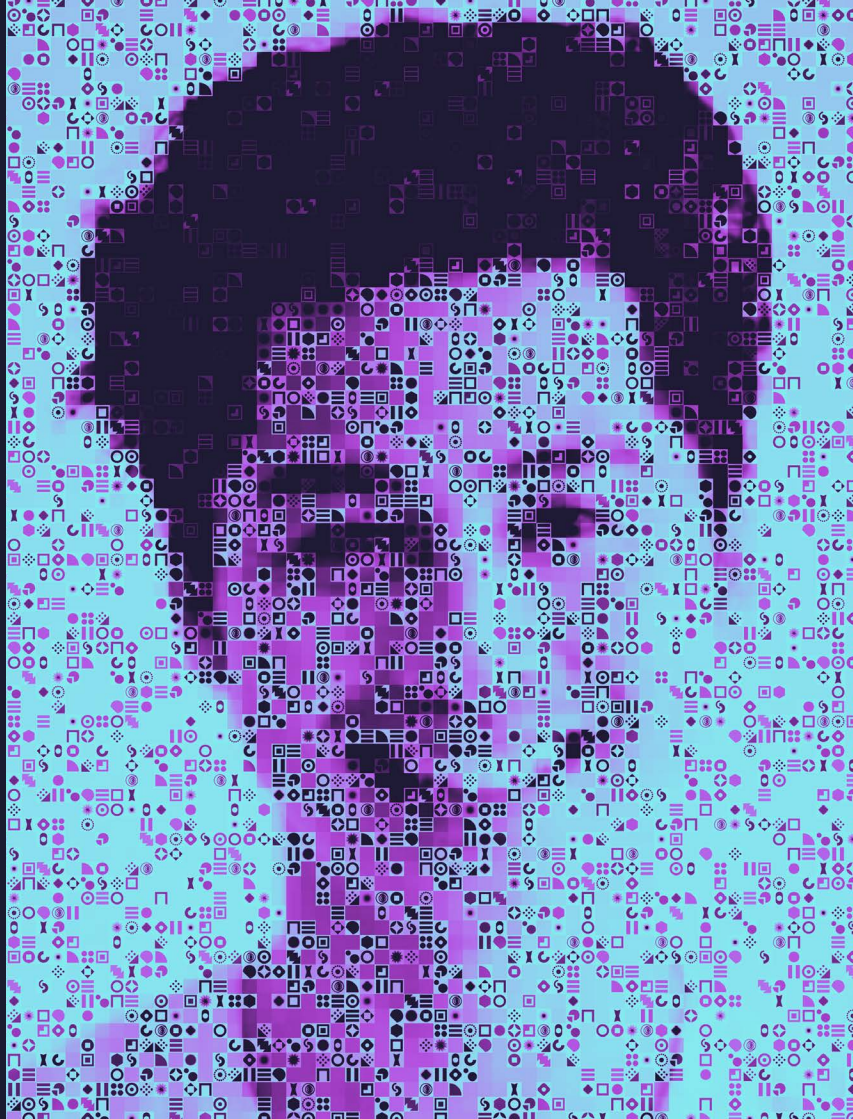
numeum

BEHIND THE CODES AND THE DATA

COMPRENDRE L'IA ÉTHIQUE

[Version 1 - Octobre 2024]



LA DÉMARCHE

- L'ambition de Numeum et de ses partenaires est de traduire les principes généraux dictant le développement d'IA éthiques **en méthodes pratiques.**
- En 2021, l'organisation de 7 ateliers d'intelligence collective rassemblant plus de **350 participants** ont permis de faire émerger un ensemble de recommandations et de les consolider dans une **première version** du Guide Ethical AI ; la publication du document s'est accompagnée d'un **manifeste pour des IA éthiques** que **60** entreprises du secteur du numérique désireuses d'œuvrer pour le développement d'IA éthiques by design ont signé.
- Conformément à son engagement, Numeum propose aujourd'hui une **mise à jour du Guide Ethical AI** qui se présente désormais en deux parties :
  - **un livret pédagogique**, qui reprend la structure du guide initial et qui est destiné à poser les enjeux, décrire les différents aspects d'une IA éthique et fournir des exemples concrets,
  - **un outil en ligne**, complément pratique du livret, pour guider pas à pas la réalisation d'une IA éthique sur tout son cycle de vie.
- Cette nouvelle version du Guide a été officiellement présentée aux pouvoirs publics.
- Tous les éléments – le livret pédagogique, l'outil en ligne et le manifeste – sont à retrouver sur le site Ethical AI dédié à la démarche.

ÉDITO

Guillaume AVRIN,
Coordinateur National pour l'IA

Katya LAINÉ, *Présidente de la Commission IA de Numeum*

L'intelligence artificielle transforme notre société à un rythme sans précédent. Son développement doit être accompagné d'une réflexion éthique pour garantir un numérique responsable, en accord avec nos valeurs et respectueux de nos droits fondamentaux.
L'essor des IA génératives ouvre des perspectives inédites, mais soulève également des défis spécifiques. Capables de générer des contenus et d'influencer les idées ainsi que les décisions, ces technologies imposent de nouvelles responsabilités à ceux qui les conçoivent, les intègrent et les utilisent. Il ne s'agit plus simplement de déterminer ce que l'IA peut accomplir, mais de définir ce qu'elle doit faire, et dans quelles conditions elle doit être employée.
Ce constat nous confronte à des enjeux majeurs, tels que la transparence, la gestion des biais algorithmiques, et à la protection des données personnelles.

Pour que l'IA s'inscrive pleinement dans un numérique responsable, il est prioritaire d'établir une gouvernance éthique robuste et durable. Cela signifie définir des cadres clairs, engager un dialogue constant avec l'ensemble des parties prenantes et assurer une formation adéquate des professionnels de l'IA. Le respect des valeurs humaines, la protection des droits fondamentaux et la promotion d'une innovation inclusive et équitable doivent constituer les fondements de nos actions dans le domaine.

Ce guide pratique se veut être bien plus qu'un simple code de conduite volontaire. Il s'agit d'une boussole éthique destinée à orienter le développement d'IA de confiance. Il permet de s'approprier certains principes clés de la réglementation européenne sur l'IA (AI Act), tout en offrant un cadre de référence pour accompagner les acteurs dans une démarche proactive de mise en conformité aux exigences réglementaires. Avec cette mise à jour, nous invitons chaque acteur à prendre connaissance des 117 recommandations articulées autour de cinq caractéristiques fondamentales — transparence, maîtrise, sûreté, respect de la vie privée et équité — qui couvrent l'ensemble du cycle de vie d'une solution d'IA, depuis la gouvernance du projet jusqu'à sa conception, son développement et sa mise en production.

Nous lançons un appel aux entreprises et aux chercheurs : engageons-nous collectivement pour que l'éthique ne soit pas perçue comme une contrainte, mais comme un levier essentiel de notre succès technologique. L'éthique de l'IA doit être l'occasion de construire des technologies de confiance, au service des valeurs de notre société et du respect des droits de chacun.

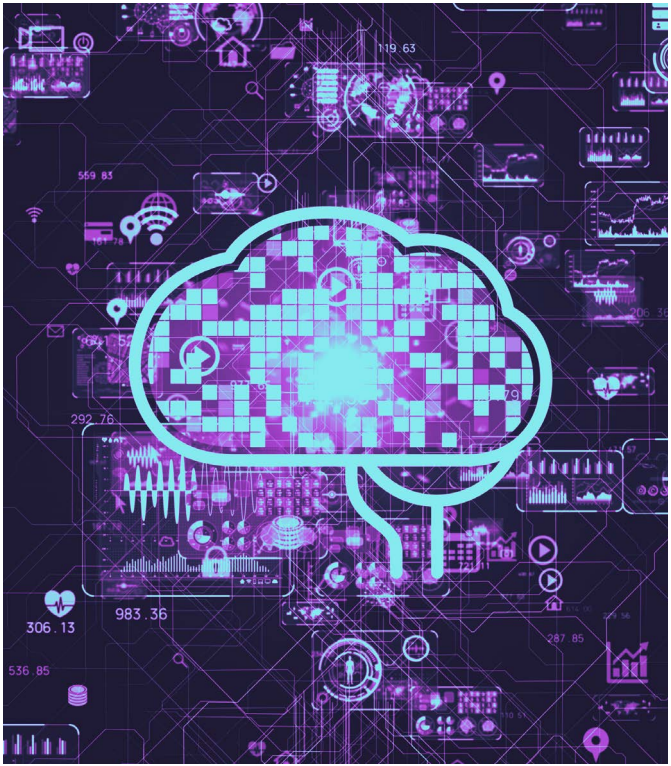
Ensemble, nous avons le pouvoir de façonner des intelligences artificielles éthiques et responsables, garantes d'un avenir numérique digne de confiance.

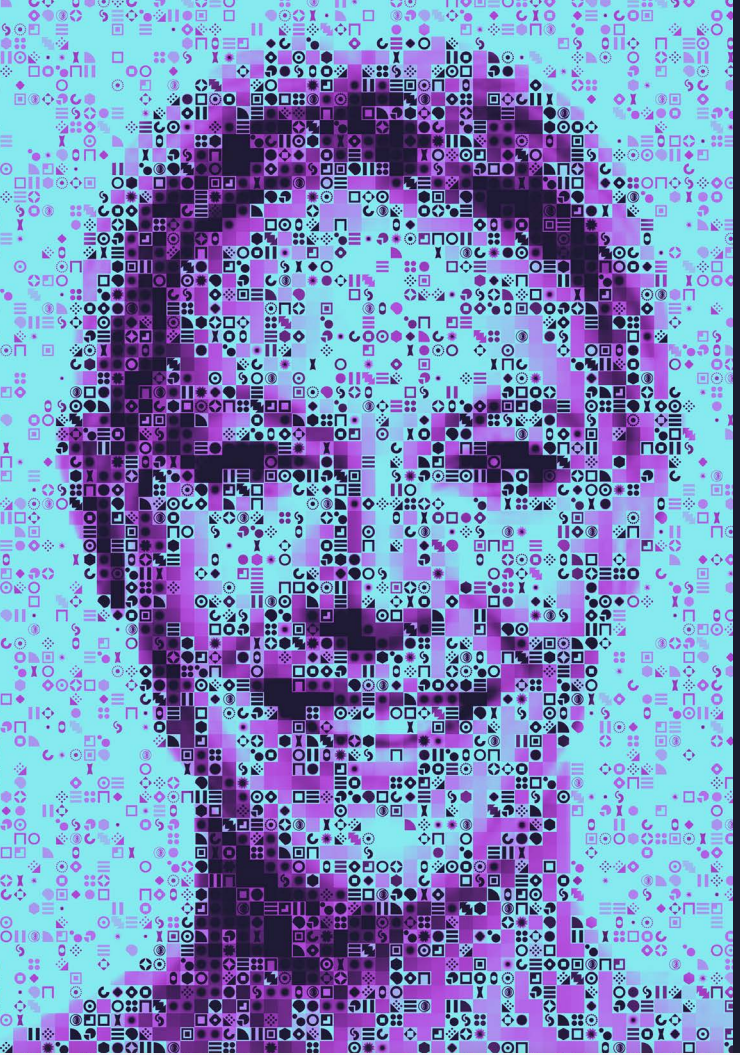
LES PARTENAIRES

SOMMAIRE

MODE D'EMPLOI DU GUIDE

- **PARTIE 1 :** elle pose les définitions et le cadre dans lequel s'inscrit la démarche.
- **PARTIE 2 :** elle souligne les enjeux de l'éthique des IA, présente les principes essentiels (mise en place d'une gouvernance appropriée, évaluation des risques éthiques) et décrit les 5 qualités qui, selon Numeum, caractérisent les systèmes d'IA éthiques.
- **PARTIE 3 :** elle rassemble, dans un outil pratique et accessible en ligne, les 117 recommandations et pistes de mise en œuvre identifiées comme essentielles pour développer des systèmes d'IA éthiques by design.





PARTIE 1

CADRE ET REPÈRES

NOTES :

L'ENJEU

L'intelligence artificielle (IA) porte en elle de nombreuses promesses : elle peut nous aider à relever les défis sociétaux et environnementaux, rendre nos entreprises plus efficaces, réduire la pénibilité de certaines tâches, mieux répondre aux clients, améliorer les soins ou encore nous assister dans notre quotidien.

Mais elle suscite des craintes. Les vagues technologiques se succèdent à un rythme effréné, tendant chacune à amplifier les dangers existants quand elles n'en charrient pas de nouveaux.

Dans ce contexte, prévoir la direction à prendre pour anticiper et maîtriser les dérives de l'IA relève de la gageure… Il y a dix ans on s'inquiétait de l'opacité des systèmes à base de réseaux de neurones profonds (deep learning). Aujourd'hui, ce sont les modèles génératifs qui mettent mal à l'aise. En cause, notamment : leur capacité à rendre la conception de virus ou de deepfakes accessible à n'importe quel individu malveillant.
Et demain, que redouter ? La disparition du droit d'auteur ?
Des détournements de finalité à grande échelle ?
La perte de l'autonomie individuelle de décision ?

L'AMBITION

Les professionnels du numérique, dont les métiers consistent à concevoir, développer, exploiter et décommissionner des solutions d'IA, se sont, très tôt, convaincus de la nécessité d'encadrer l'exploitation de ces technologies pour éviter le risque d'un rejet massif et durable de la technologie, malgré ses potentialités positives.

Pour cette raison, Numeum, qui les représentent, et ses partenaires (Hub France IA, Institut Data IA, Institut 3IA Côte d'Azur, Impact AI, l'Ecole 42, Aivancity, Telecom Valley, La Région Grand Est, ANITI et Grand E-nov+), ont élaboré, dès 2021, un premier guide méthodologique destiné à aider les professionnels à créer et à utiliser des systèmes d'IA respectueux des valeurs et principes fondamentaux de la société. La publication du Guide Ethical AI s'est accompagnée de la diffusion d'un manifeste signé par 60 entreprises du secteur du numérique souhaitant valoriser leur engagement pour la conception et l'usage d'IA éthiques by design.

Trois ans après cette première initiative, Numeum propose une mise à jour du Guide Ethical AI qui se présente désormais sous deux formes :

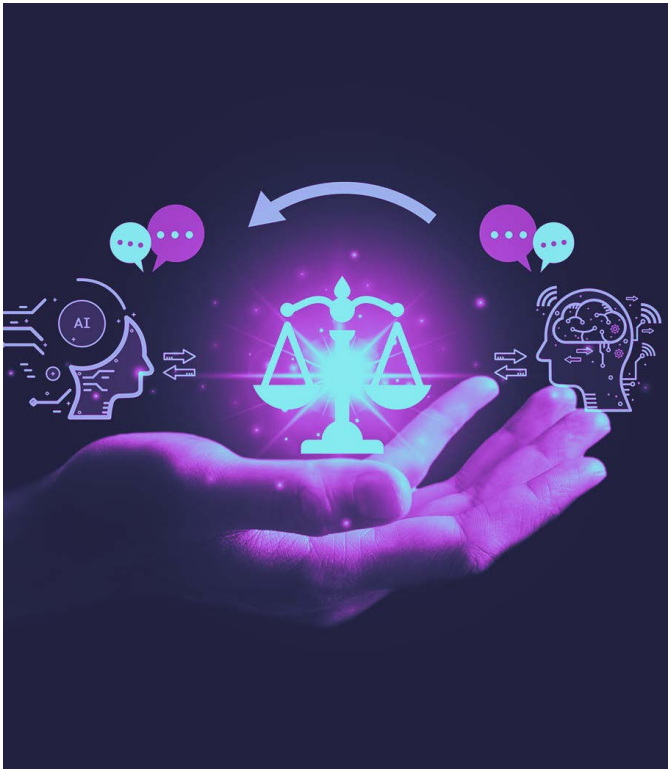
- un livret pédagogique, qui reprend la structure du guide initial et qui est destiné à poser les enjeux, décrire les différents aspects d'une IA éthique et fournir des exemples concrets,
- un outil, complément pratique du livret, pour guider pas à pas la réalisation d'une IA éthique sur tout son cycle de vie.

LE LIEN AVEC L'AI ACT

Le développement et l'utilisation de systèmes d'IA dans l'UE sont aujourd'hui encadrés par deux règlements : l'AI Act, publié au journal officiel en juillet 2024 et qui traite des IA en général, et le RGPD (entré en application en 2018) pour ce qui concerne le sujet spécifique de la protection des données à caractère personnel.
Tout système d'IA doit se soumettre à ces deux législations sous peine d'être illicite et de ne pouvoir accéder au marché européen.

Cette nouvelle version du Guide Ethical AI s'inscrit dans la trace de l'AI Act – les deux documents se recoupent sur la plupart des sujets et ne présentent pas de désaccords fondamentaux. Mais les réglementations se différenciant d'une région à l'autre du monde et la technologie progressant plus rapidement que le droit, Numeum a jugé préférable d'aborder le sujet indépendamment des considérations réglementaires, se donnant la liberté de se détacher du seul objectif de conformité.

Ainsi, si l'outil pratique du Guide Ethical AI peut s'utiliser comme une aide à la mise en conformité avec l'AI Act, il se présente avant tout comme un code de bonne conduite et ne peut en aucun cas constituer une garantie absolue de conformité.



L'ESPRIT DE L'AI ACT

La philosophie de la régulation de l'IA en Europe est d'adapter les obligations au niveau de risque engendré par le système d'IA considéré. D'où une classification des systèmes selon quatre niveaux de risque, avec des obligations spécifiques pour chacun :

- **Risque inacceptable :** cas d'usage interdits dans l'UE, comme la notation sociale, l'identification biométrique généralisée et la manipulation de contenus (voir la liste complète dans l'article 5.1 du règlement).
- **Risque élevé :** systèmes embarqués dans des produits à haut risque (voir la liste dans l'annexe II du règlement) ou déployés dans des secteurs jugés à haut risque, comme l'éducation, les ressources humaines, la santé, la justice, certains services bancaires, les transports autonomes… (voir la liste dans l'annexe III) ; les systèmes appartenant à cette classe doivent être enregistrés dans la base de données de l'UE et mettre en œuvre un certain nombre d'obligations pour être certifiés (système de gestion des risques, traçage des données, documentation technique complète, sécurisation, supervision humaine, etc.).
- **Risque faible :** systèmes dont la fonction est d'interagir avec les humains (chatbot, par exemple) et qui ne figurent pas dans les deux précédentes catégories ; leur seule obligation est la transparence vis-à-vis de l'utilisateur.
- **Risque minime :** tous les autres systèmes ; aucune obligation.

Les modèles à usage général – ou de fondation – tels que les grands modèles linguistiques sont l'objet d'une réglementation propre. Ils doivent, en particulier, fournir des informations sur les données d'entraînement et se soumettre aux réglementations relatives au droit d'auteur. Des dispositions spécifiques sont, en outre, prévues selon la taille des modèles et selon qu'ils sont en open source ou non. Les très grands modèles, qui présentent un risque systémique, sont, par exemple, tenus d'effectuer des tests d'adversité et de déclarer les incidents.
Les délais de mise en conformité des systèmes dépendent de la classe de risques à laquelle ils appartiennent et s'étalent approximativement entre janvier 2025 (bannissement du marché européen des systèmes à risque inacceptable) et juillet 2027.

ÉTHIQUE ET IA : DE QUOI PARLE-T-ON ?

Un système d'IA peut être considéré comme conforme aux valeurs éthiques s'il est en mesure, tout au long de son cycle de vie, de préserver les droits humains fondamentaux : droits à la dignité, à l'intégrité mentale et physique, à la liberté, à l'autonomie, à l'équité de traitement, à l'intimité et à la vie privée.

Or, les systèmes d'IA actuels, par leur nature – notamment ceux faisant appel aux technologies de deep learning et d'IA générative –, leur place dans nos vies et leur capacité de diffusion à l'échelle planétaire concentrent un nombre élevé de risques d'ordre éthique : discrimination, déshumanisation des relations sociales, opacité des processus de déduction, utilisation abusive des données personnelles, malfaisance (manipulation, détournement d'usage, prise de contrôle indue…), etc.

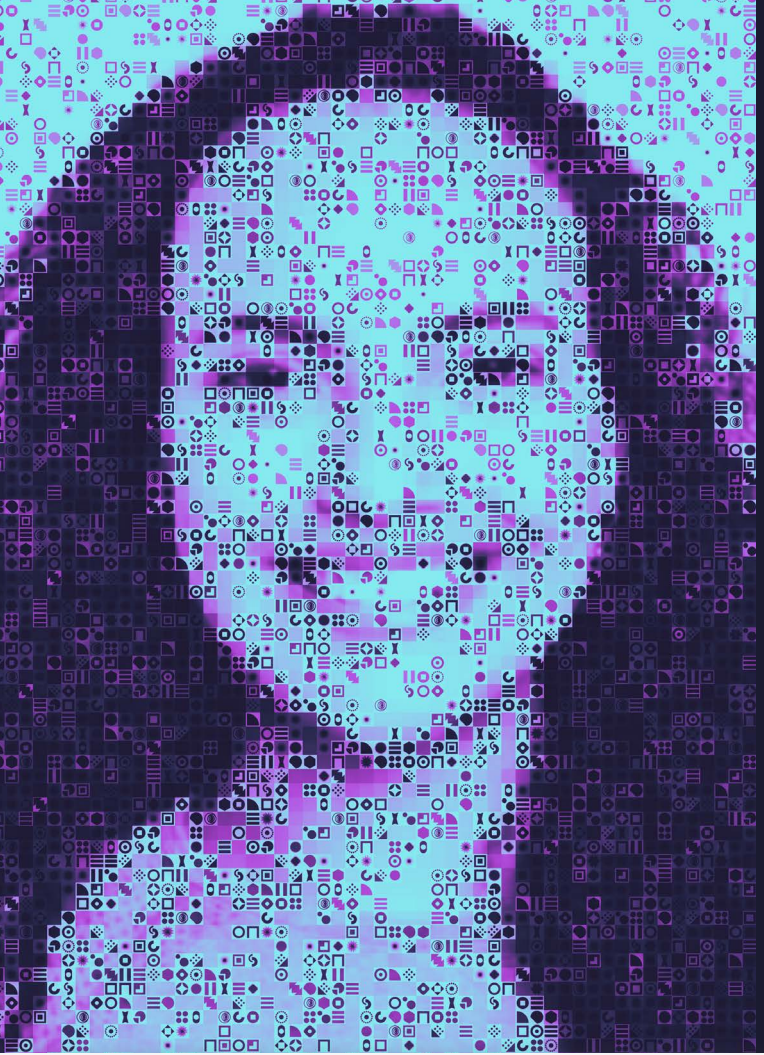
Des dispositions et précautions spécifiques s'imposent donc en amont de la conception des systèmes d'IA pour éviter toute dérive volontaire ou involontaire et faciliter les arbitrages entre objectifs de performance et principes éthiques. Car – faut-il le rappeler ? – l'humain reste finalement le seul responsable des systèmes qu'il crée.

5 QUALITÉS MAJEURES

Numeum considère **5 grandes thématiques**, ou « qualités », à aborder pour rendre un système d'IA conforme aux valeurs éthiques :

1. **Respect** de la vie privée
2. **Équité** des résultats et inclusion
3. **Transparence** du modèle et de son artificialité
4. **Maîtrise** et contrôle par l'humain
5. **Sûreté** et robustesse face aux cyberattaques

Ces 5 qualités et leurs sous-jacents concrets sont décrits dans les pages suivantes du livret. Les dispositions techniques à mettre en œuvre pour les remplir figurent, quant à elles, dans l'outil.



PARTIE 2

LES PRINCIPES DE L'IA ÉTHIQUE

NOTES :

UNE GOUVERNANCE DE L'IA AU PLUS HAUT NIVEAU DE L'ENTREPRISE

L'enjeu lié au risque éthique d'un système d'IA est trop élevé pour reposer sur les seules têtes des data scientists et autres spécialistes techniques de l'entreprise. Le sujet est souvent d'une grande complexité. Il nécessite parfois des décisions de haut niveau qui réclament de réunir des compétences autres que celles présentes dans une équipe projet. Il peut, en outre, faire porter un risque réputationnel ou juridique sur l'entreprise.

C'est pourquoi la première recommandation de ce guide est la mise en place, au plus haut niveau de l'organisation, d'une gouvernance visant à définir et faire appliquer la politique en matière d'IA de l'entreprise. Reflétant les valeurs éthiques de cette dernière, cette politique posera le cadre dans lequel les projets d'IA seront développés, à savoir un ensemble de processus et de méthodes standardisés visant à sécuriser le traitement des questions éthiques et à se conformer à la réglementation.

Le cadre définira notamment :

- la méthode d'évaluation des risques éthiques,
- les modalités d'arbitrage et d'instruction des cas critiques ainsi que les seuils et critères d'équité et d'explicabilité auxquels se référer,
- le processus d'alerte et d'escalade à suivre si survient un risque ou un incident,
- les moyens à mettre en œuvre pour sensibiliser les équipes projet aux enjeux éthiques.

Il précisera également les attributions des gouvernances de projet d'IA concernant les sujets éthiques.

L'organisation et le périmètre de cette gouvernance d'entreprise varieront d'une organisation à l'autre. Dans certains cas, les questions de conformité RGPD et NIS (Network and Information System Security) pourraient aussi lui incomber.

ÉVALUER LES RISQUES ÉTHIQUES DU PROJET

La première étape de la création d'un système d'IA éthique by design, consiste à évaluer les risques éthiques qu'il porte : ceux propres au cas d'usage et au secteur dans lequel le système sera déployé, mais aussi ceux induits par les méthodes et algorithmes employés ainsi que ceux liés à la gouvernance du projet.

La réalisation d'une cartographie soignée mettant en évidence la sensibilité au risque éthique du futur système selon ces trois axes permettra d'orienter l'effort dans la bonne direction et selon la bonne mesure.

Pour se conformer à l'AI Act, l'étude devra par ailleurs définir dans quelle classe de risques se situe le système d'IA et appliquer les exigences correspondant à la classe.

EXEMPLES DE RISQUES LIÉS AU CAS D'USAGE

- résultat de l'IA conduisant à refuser ou limiter l'accès d'une personne à un droit fondamental (refus d'accès à un service essentiel, par exemple),
- génération d'addiction,
- faux contenus destinés à duper l'utilisateur,
- génération de contenus protégés,
- production de résultats erronés qui faussent la décision,
- survenance de cas non prévus et potentiellement non maîtrisés pouvant porter atteinte à l'intégrité ou à la dignité humaine ou nuire à l'environnement (ce risque peut apparaître sur les systèmes autonomes), etc.

EXEMPLES DE RISQUES PROPRES À LA TECHNOLOGIE

- non-conformité aux règlements en vigueur (l'AI Act et le RGPD, en Europe, par exemple),
- cyberattaque entraînant la divulgation d'informations à caractère personnel ou un détournement de finalité,
- opacité du chemin de décision du système d'IA,
- non-traçabilité des données et des méthodes qui empêcherait la correction du système ou la recherche de responsabilités,
- utilisation de contenus protégés pour l'apprentissage, etc.

EXEMPLES DE RISQUES LIÉS AU CONTEXTE DU PROJET

- absence de prise en compte des sujets éthiques par la gouvernance projet,
- absence d'analyse des risques éthiques,
- méconnaissance des enjeux par les équipes et/ou l'entreprise,
- absence de processus de correction et de réparation,
- absence de référents, etc.

→ La grille d'auto-évaluation ci-dessous vise à faciliter l'évaluation du projet sur le plan de sa sensibilité au risque éthique.

SENSIBILITÉ D'UN SYSTÈME D'IA AUX DIFFÉRENTS ENJEUX ÉTHIQUES

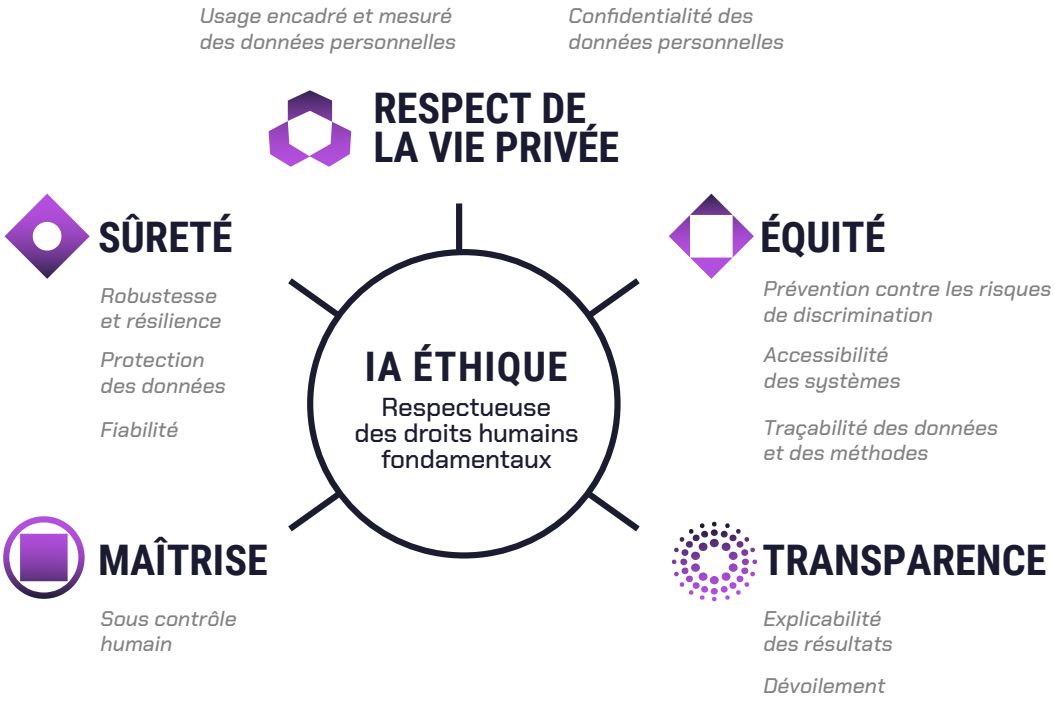
Sujets éthiques à considérer spécifiquement si la réponse est Vrai →			Respect vie privée		Équité		Transparence			Maîtrise	Sûreté	
	↓ Finalité et cadre de mise en œuvre du système	S'applique au projet	Usage encadré et mesuré des données à caractère personnel	Confidentialité des données à caractère personnel	Prévention contre les risques de discrimination	Accessibilité de la solution	Traçabilité des données et des méthodes	Explicabilité des résultats	Dévoilement	Fonctionnement sous contrôle humain	Robustesse et réslience de la solution	Fiabilité des résultats
Le cas d'usage	Le système automatise ou aide une prise décision qui concerne des personnes physiques	Vrai/Faux	⊗	⊗	⊗			⊗		⊗	⊗	⊗
	Le système automatise l'exécution de tâches pour l'utilisateur	Vrai/Faux						⊗	⊗	⊗	⊗	⊗
	Le système est voué à être utilisé à très large échelle dans le grand public	Vrai/Faux					⊗			⊗	⊗	⊗
	Le système génère des contenus (images, textes, audio, vidéo…)	Vrai/Faux					⊗		⊗	⊗	⊗	⊗
	Le système interagit avec l'utilisateur	Vrai/Faux	⊗			⊗			⊗	⊗	⊗	⊗
	Le système automatise des actions	Vrai/Faux					⊗	⊗	⊗	⊗	⊗	⊗

Sujets éthiques à considérer spécifiquement si la réponse est Vrai →			Respect vie privée		Équité		Transparence			Maîtrise	Sûreté	
↓ Finalité et cadre de mise en œuvre du système		S'applique au projet	Usage encadré et mesuré des données à caractère personnel	Confidentialité des données à caractère personnel	Prévention contre les risques de discrimination	Accessibilité de la solution	Traçabilité des données et des méthodes	Explicabilité des résultats	Dévoilement	Fonctionnement sous contrôle humain	Robustesse et réslience de la solution	Fiabilité des résultats
La gouvernance projet	L'équipe projet ne peut ni solliciter d'instance dans l'entreprise responsable des sujets d'éthique et d'IA ni se référer à des règles de gouvernance sur le plan éthique des projets d'IA	Vrai/Faux	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	L'équipe projet n'a pas été sensibilisée aux enjeux éthiques	Vrai/Faux	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗	⊗
	L'équipe projet n'a pas été sensibilisée aux enjeux de cybersécurité	Vrai/Faux		⊗			⊗			⊗	⊗	
	L'équipe projet présente un défaut de diversité (genre, origine, culture, métier…)	Vrai/Faux			⊗							
	Certains acteurs de la chaîne sont des partenaires extérieurs	Vrai/Faux		⊗			⊗				⊗	

Sujets éthiques à considérer spécifiquement si la réponse est Vrai →			Respect vie privée		Équité		Transparence			Maîtrise	Sûreté	
↓ Finalité et cadre de mise en œuvre du système		S'applique au projet	Usage encadré et mesuré des données à caractère personnel	Confidentialité des données à caractère personnel	Prévention contre les risques de discrimination	Accessibilité de la solution	Traçabilité des données et des méthodes	Explicabilité des résultats	Dévoilement	Fonctionnement sous contrôle humain	Robustesse et réslience de la solution	Fiabilité des résultats
La solution technique d'IA	Le système est embarqué dans une solution plus large	Vrai/Faux					⊗			⊗	⊗	⊗
	Le système utilise un modèle génératif	Vrai/Faux					⊗	⊗	⊗	⊗	⊗	⊗
	L'apprentissage requiert l'utilisation de données sensibles et/ou à caractère personnel	Vrai/Faux	⊗	⊗			⊗		⊗		⊗	
	L'apprentissage utilise des jeux de données d'origine inconnue	Vrai/Faux	⊗		⊗		⊗					⊗
	Le système met en œuvre des technologies non naturellement explicables	Vrai/Faux					⊗	⊗				
	Le système traite des données sensibles et/ou à caractère personnel	Vrai/Faux	⊗	⊗			⊗		⊗	⊗	⊗	
	Le système apprend en continu	Vrai/Faux	⊗	⊗	⊗		⊗			⊗	⊗	⊗

LES 5 QUALITÉS DES IA ÉTHIQUES

Les parties suivantes décrivent les qualités des IA éthiques et ce qu'elles recouvrent en donnant quelques bonnes pratiques. Les recommandations pour chacune des étapes du cycle de vie d'un projet sont détaillées dans l'outil.



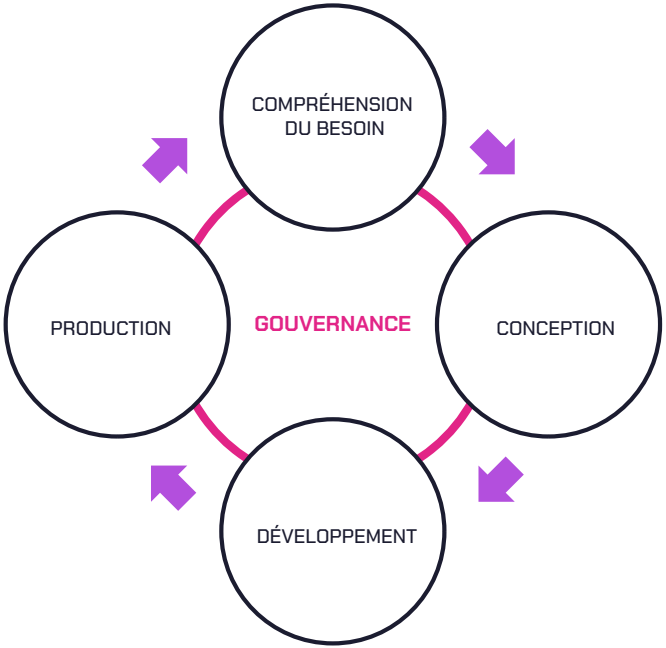
LE CYCLE DE VIE

GOUVERNANCE PROJET

- Les attributions de la gouvernance du projet concernant l'éthique seront définies par la gouvernance IA de l'entreprise. Elles pourront varier d'une organisation à l'autre, mais consisteront, en général, à suivre les questions suivantes tout au long de la vie du projet : analyse de risques et d'impact, sensibilisation et formation des parties prenantes aux risques éthiques du projet et aux exigences légales, mise en place et suivi des systèmes de documentation technique et de traçabilité, définition des habilitations nécessaires, organisation du processus de remontées des alertes et de traitement des recours en cas d'attaques du système, etc.

COMPRÉHENSION DU BESOIN

- Phase durant laquelle les membres de l'équipe projet, en particulier les data scientists, prennent connaissance du sujet et du cadre du projet auprès de leurs donneurs d'ordre. C'est à ce moment que les questions éthiques sont abordées et que les différents risques, critères et seuils sont posés.



CONCEPTION

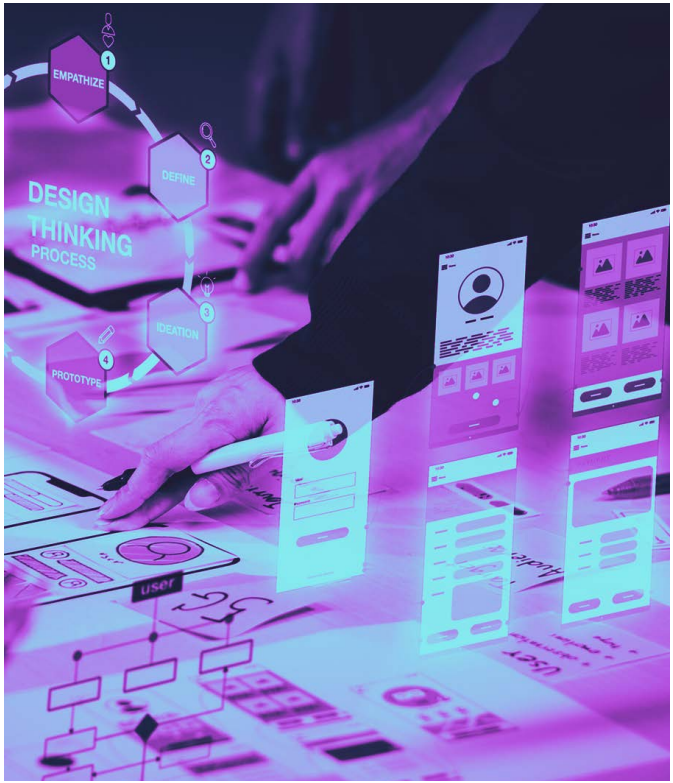
- Cycle itératif au cours duquel les data scientists construisent les différents jeux de données nécessaires à l'apprentissage et au test puis élaborent et testent le modèle. Les arbitrages entre performances et enjeux éthiques se passent souvent pendant cette phase.

DÉVELOPPEMENT/INTÉGRATION

- Étape qui consiste à intégrer le modèle dans son environnement de production (dans un système, dans une application, etc.) et à développer les modules complémentaires au modèle lui-même (les interfaces, par exemple). Cette phase requiert en général des compétences informatiques.

PRODUCTION

- Le système d'IA est opérationnel. Il est mis en ligne et monitoré pour surveiller les dérives possibles. Il pourra faire l'objet de mises à jour. En fin de vie, il sera décommissionné.



RESPECT DE LA VIE PRIVÉE

Les systèmes d'IA doivent garantir le respect de la vie privée. Cela signifie faire un usage mesuré et encadré par la loi des données à caractère personnel, et garantir leur confidentialité. Cela concerne les informations initialement fournies par l'utilisateur ainsi que celles produites par le système lui-même.

CE QUE DIT LE RGPD

Les sociétés conceptrices d'IA doivent bien évidemment se conformer aux réglementations en vigueur dans les territoires ciblés. Au sein de l'Union européenne, le règlement à appliquer en la matière est le RGPD. Il repose sur 5 principes qui doivent guider la conception et la mise en œuvre des systèmes d'IA :

- les données à caractère personnel sont utilisées dans un but précis et défini,
- le système ne collecte que les données à caractère personnel qui lui sont strictement nécessaires,
- la durée de conservation des données personnelles est raisonnable et le décommissionnement est prévu,
- des mécanismes de confidentialité et de sécurité protègent ces données,
- les personnes peuvent accéder à leurs données, les modifier, les supprimer et les transporter d'un traitement à un autre.

USAGE ENCADRÉ ET MESURÉ DES DONNÉES PERSONNELLES

En se nourrissant de quantités gigantesques de données, les systèmes d'IA par apprentissage automatique exacerbent les risques d'abus concernant les données personnelles tout en défiant les principes mêmes sur lesquels reposent les réglementations existantes de protection des données personnelles telles que le RGPD. Les concepteurs de systèmes d'IA doivent s'attacher à réduire au minimum leur utilisation des données personnelles, voire à les remplacer par des données synthétiques. Ils doivent aussi informer l'utilisateur de l'usage que le système fait de ses données personnelles et trouver le juste équilibre entre durée de conservation pour le traçage et suppression.

CONFIDENTIALITÉ DES DONNÉES PERSONNELLES

Le respect de la confidentialité des données requiert la mise en place de dispositifs de sécurité visant à empêcher l'intrusion illicite dans les bases de données (voir le chapitre Sûreté). Il passe aussi par l'emploi de techniques empêchant d'identifier l'individu à partir de ses données personnelles.

Plusieurs solutions existent telles que :

- l'anonymisation (suppression des données permettant d'identifier directement un individu),
- la pseudonymisation (remplacement des données directement identifiantes par des valeurs ou des alias),
- la collecte de données moins précises (une tranche d'âge plutôt qu'un âge),
- l'apprentissage fédéré (réduction des points de centralisation des données par l'emploi d'une base de données distribuée),
- le chiffrement des données, etc.

Aucune n'est totalement fiable individuellement. La bonne voie consiste souvent à combiner plusieurs approches, en les équilibrant en fonction de la finalité de l'application et des objectifs de performance.

ÉQUITÉ

PRÉVENTION CONTRE LES RISQUES DE DISCRIMINATION

Il est important de veiller à l'impartialité des systèmes d'IA, en particulier si leurs conclusions concernent des personnes (systèmes de notation ou de recrutement, par exemple). Un système d'IA par apprentissage machine peut, en effet, présenter des biais en particulier s'il a été entraîné avec un jeu de données lui-même biaisé (un historique biaisé, par exemple). Le risque dans ce cas est d'aboutir à des résultats faux ou discriminatoires et donc préjudiciables. L'utilisation de jeux de données d'apprentissage représentatifs et la réalisation de tests d'équité contribuent, en général, à réduire le risque.

Il ne faut, cependant, négliger ni le caractère éminemment culturel de la notion (dépendant donc du marché visé) ni sa complexité (l'impartialité prend parfois des formes incompatibles...). D'où le besoin de préciser en amont les critères d'équité que l'on souhaite obtenir pour faciliter la recherche de la solution optimale.

La diversité de l'équipe de conception (origine, genre, âge, etc.) peut s'avérer un atout pour se poser les bonnes questions et aborder le sujet avec un état d'esprit ouvert.

ACCESSIBILITÉ DE LA SOLUTION

L'IA présente un vrai potentiel pour faciliter l'accès au numérique de personnes souffrant de handicaps ou en difficulté par rapport au numérique. Par-delà les caractéristiques réglementaires dont une interface graphique doit se prévaloir, l'intégration de fonctions de reconnaissance d'images automatique, de traduction automatique, de synthèse vocale ou d'analyse de texte dans un système d'IA en interaction avec des humains joue en faveur de l'accessibilité et de l'inclusion.

TRANSPARENCE

TRAÇABILITÉ DES DONNÉES ET DES MÉTHODES

La traçabilité des données, méthodes et processus est une condition nécessaire à la possibilité de vérifier l'absence de biais, auditer le système, garantir la fiabilité et l'intégrité des systèmes et des données, etc. Cela signifie documenter pendant tout le cycle de vie tout ce qui a trait au système d'IA : méthode de collecte et origine des jeux de données, algorithmes et méthodes de conception employés, arbitrages effectués, méthodes et résultats des tests, résultats en production, incidents éventuels, etc.

Il s'agit, ni plus ni moins, que d'appliquer une démarche qualité, comme cela se pratique dans les autres domaines de l'informatique.

EXPLICABILITÉ DES RÉSULTATS

Les causes et critères qui conduisent aux conclusions d'un système d'IA doivent pouvoir être portés à la connaissance des utilisateurs ou des personnes affectées ou concernées par la conclusion. Malheureusement, certaines IA très performantes – comme celles à base de réseaux de neurones profonds (deep learning) – se comportent comme des boîtes noires : elles ne permettent pas d'identifier facilement les variables qui ont influencé le résultat et donc d'interpréter ce dernier.

Il est, de ce fait, important d'estimer en amont le besoin et le degré d'explicabilité requis, en fonction du cas d'usage et de la finalité du produit – un algorithme de notation pour l'obtention d'un prêt bancaire exigera un niveau probablement plus élevé d'explicabilité qu'un service de recommandation de livres. Cela, afin de sélectionner les algorithmes et méthodes de conception en connaissance de cause et, le cas échéant, définir les critères selon lesquels arbitrer, entre transparence et précision des résultats.

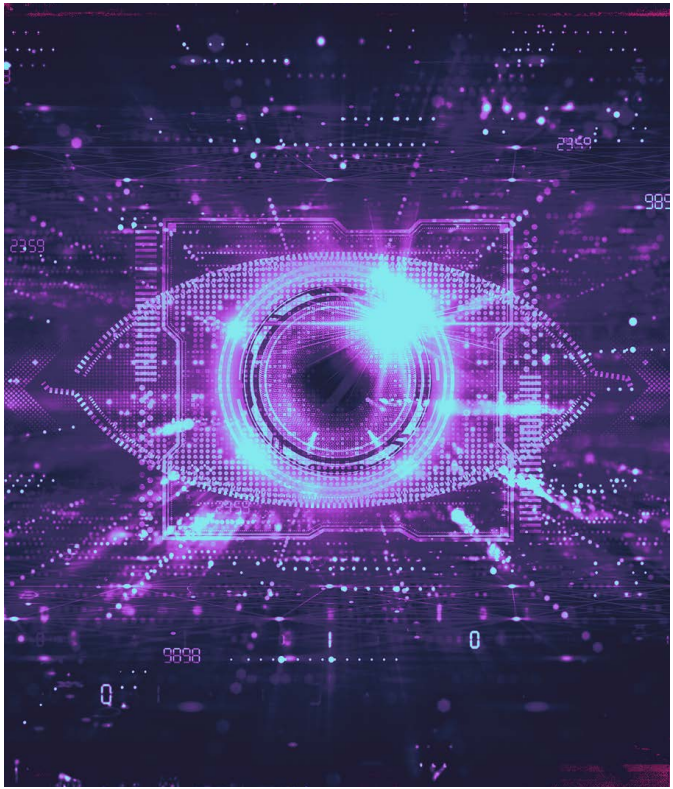
La question de l'explicabilité doit être reposée régulièrement tout au long de la conception, au fur et à mesure de l'amélioration des résultats.

DÉVOILEMENT

La possibilité pour l'utilisateur de comprendre qu'il interagit avec une IA (s'il s'adresse à un agent conversationnel, par exemple) ou qu'il est en présence d'un contenu généré artificiellement est un élément clé pour instaurer la confiance des utilisateurs, réduire les risques d'abus de faiblesse et la circulation de fausses informations (deepfakes).

Dans le cas d'un agent conversationnel, il peut suffire, d'entrée de jeu, d'informer l'humain qu'il interagit avec un chatbot. Pour une création visuelle artificielle, la tendance est d'intégrer des tatouages dans les images.

Sensibiliser les parties prenantes du projet à cet enjeu de transparence peut aider les équipes à identifier les informations à communiquer à l'utilisateur et à choisir les bons canaux.



MAÎTRISE

Maintenir les systèmes d'IA sous contrôle humain est un facteur clé pour éviter la déshumanisation de nos sociétés, protéger l'autonomie de décision des individus et rendre viable l'idée que l'humain reste seul responsable des actions et décisions des IA.

Cela signifie concrètement de respecter les points suivants :

- dans tous les domaines où une décision qui affecte un individu (sa vie, sa réputation, sa santé, etc.) doit être prise, la décision finale doit revenir à une personne,
- l'utilisateur d'un système d'IA doit pouvoir prendre connaissance des recommandations émises, mais rester décisionnaire et autonome dans ses choix individuels,
- tout système en interaction avec un humain doit prévoir la possibilité d'émettre un recours ou de signaler une anomalie (ce qui implique de mettre en place le processus interne permettant de traiter les recours et remontées d'informations des utilisateurs),
- il doit rester possible de ne pas faire appel au système d'IA s'il est jugé que les conditions éthiques ou de sûreté ne sont pas remplies.

SÛRETÉ

ROBUSTESSE ET RÉSILIENCE

Comme tout système informatique, un système d'IA est une cible potentielle du cybercrime. Il doit donc bénéficier des mêmes règles, principes et dispositifs de protection que n'importe quel autre système d'information. Mais il est aussi à la merci de menaces spécifiques :

- **Empoisonnement :** l'attaquant cherche à biaiser le comportement d'un modèle en modifiant les données d'apprentissage ; les systèmes apprenant en continu sont particulièrement exposés à ce type d'attaque.
- **Évasion :** l'attaquant modifie de manière imperceptible les données d'entrée de l'application pour leurrer le système et lui faire produire une décision différente de celle normalement attendue ; les systèmes traitant des données d'entrée complexes comme les images sont particulièrement sensibles à ce type d'attaque.
- **Inférence :** l'attaquant assaille l'IA de requêtes pour comprendre son fonctionnement et saisir les paramètres clés, dans le but d'imiter le système ; les systèmes qui diffusent beaucoup d'informations s'exposent plus facilement à ce type d'attaque.

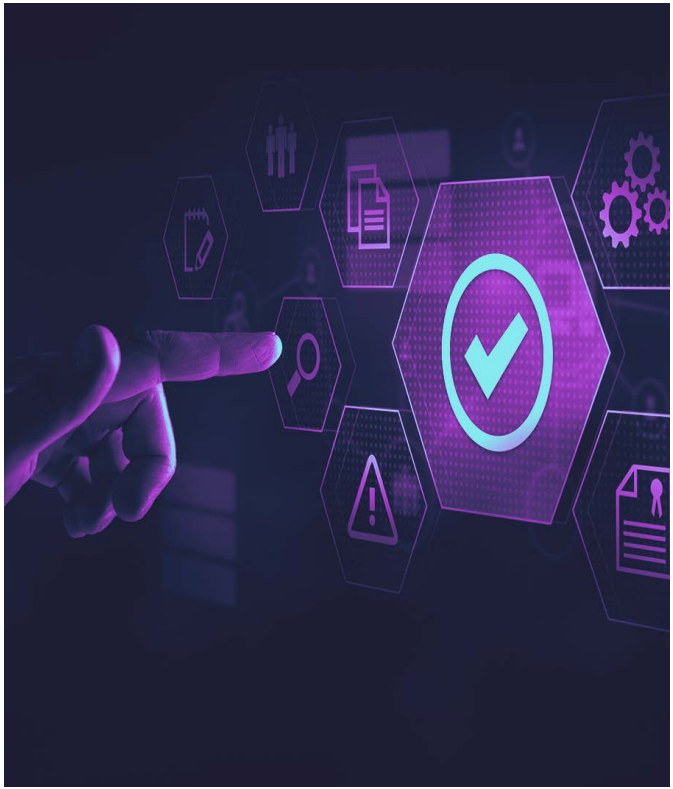
Le système doit comporter les mécanismes lui permettant de se prémunir contre ces risques spécifiques, et de bloquer et corriger les effets des éventuelles attaques.

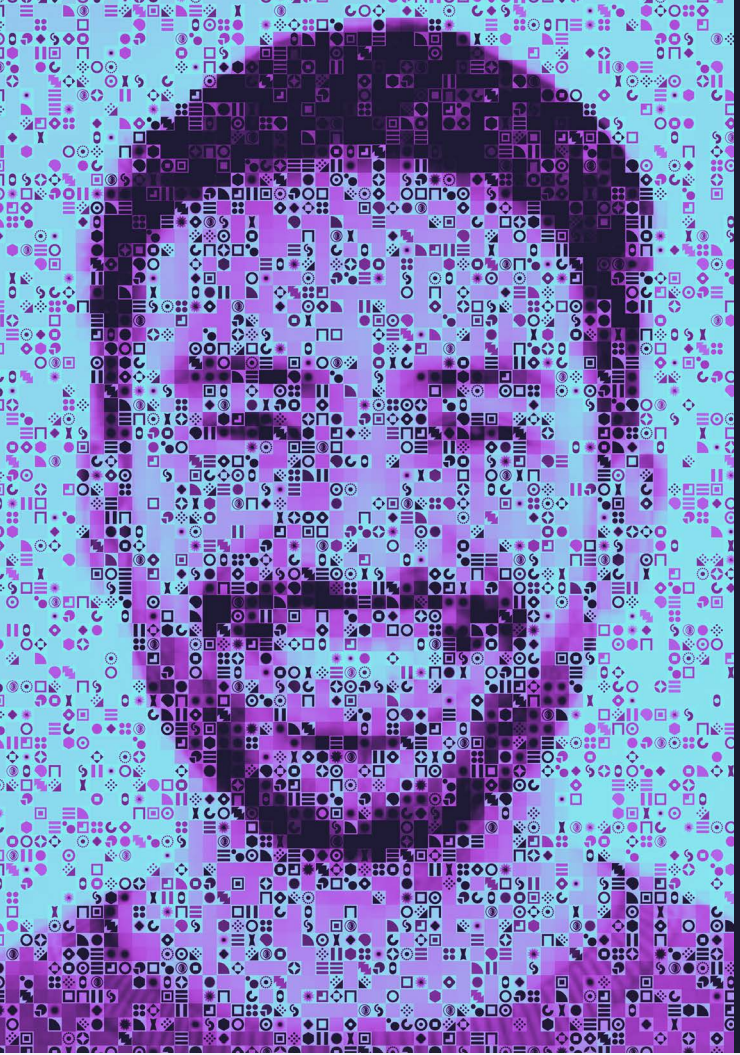
La première mesure à prendre consiste sans doute à sensibiliser les data scientists et autres métiers de la data aux enjeux de cybersécurité. Contrairement aux informaticiens, cette population, souvent issue des filières mathématiques et statistiques, s'avère naturellement moins préoccupée par ces questions. La mise en place d'un dispositif d'habilitation permettra de sécuriser les accès au modèle et aux données et de réduire les risques d'attaques par empoisonnement et de vols de données.
La mise en œuvre de méthodes comme le bruitage des données, la distillation de modèles ou l'apprentissage adversarial permettra de réduire la sensibilité des modèles aux attaques par inférence et par évasion.

FIABILITÉ DES RÉSULTATS

La problématique figure parmi les plus complexes à traiter, car par nature, les IA par apprentissage automatique ne peuvent garantir à 100 % la reproductibilité de leurs résultats, à fortiori si elles apprennent en continu. Pour les systèmes complexes qui manipulent un très grand nombre de paramètres, on se heurte à la difficulté supplémentaire de ne pas pouvoir tester et donc valider l'ensemble des cas possibles.

Des règles contraignantes et des limites à ne pas franchir devront être définies à l'aune des résultats de l'étude de risques et d'impacts réalisée en amont pour encadrer et sécuriser le fonctionnement du système. Là encore, des compromis entre simplicité et performance devront être décidés pour mieux maîtriser la reproductibilité des résultats.





PARTIE 3

RECOMMANDATIONS PRATIQUES

NOTES :

Les 117 recommandations sont accessibles sur le site Ethical AI ou en scannant le QR Code. Elles sont accompagnées de pistes de solutions pour les mettre en œuvre.

Organisées par qualité et grande problématique ainsi que par étape du cycle de vie pour vous permettre de les consulter de la manière qui vous convient le mieux grâce à un système de filtres, ces recommandations visent à vous guider pas à pas dans la conception de systèmes d'IA éthiques by design.

- **Directrice de la publication :**
  Véronique TORNER (NUMEUM)
- **Conception et coordination :**
  Katya LAINE (NUMEUM et TALRK.AI)
  Atef BEN OTHMAN (NUMEUM)
- **Rédaction :**
  Bénédicte DE LINARES (BDL CONSEIL)
- **Conseils techniques :**
  Aziz AMAL (ASTEK)
  Mathilde VERON (ASTEK)
  Sylvain AKRICHE (CEGID)
  Francois Marie LESAFFRE (SOPRA STERIA)
- **Crédits photos :** IStock

Initiative conduite par :

numeum

22-28 Rue Joubert – 75009 Paris
01 44 30 49 70 - contact@numeum.fr
@numeum

Soutenue par :

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Premier ministre

3iA Côte d'Azur
Institut interdisciplinaire d'intelligence artificielle

42

aivancity
SCHOOL FOR TECHNOLOGY, BUSINESS & SOCIETY
PARIS-CACHAN

ANITI
Université Fédérale Toulouse Midi-Pyrénées
TOULOUSE INSTITUTE

GRAND ENOV+
AGENCE D'INNOVATION & DE PROSPECTION INTERNATIONALE

HUB FRANCE IA

IMPACT AI

INSTITUT DATAIA
Science des données, Intelligence & Société

Grand Est
ALSACE CHAMPAGNE-ARDENNE LORRAINE
L'Europe s'invente chez nous

Telecom valley
Animateur Azuréen Numérique

Conception graphique et réalisation : MAP ADVERTISING