Une image contenant motif, blanc, conception, noir et blanc

Le contenu généré par l’IA peut être incorrect.

Intelligence artificielle

Bonnes pratiques

Objet

Ce document s’inscrit dans un ensemble de documents méthodologiques destinés à aider les organismes à gérer les risques liés à l’intelligence artificielle (IA). Ils peuvent être utiles et utilisés ensemble ou séparément.

**Il propose une liste de bonnes pratiques qui permettent de respecter les critères de confiance que tout systèmes d’IA devrait respecter**.

Il peut être directement utilisé dans le cadre des projets de nouveaux services qui reposent sur l’IA, et peut également être intégré à une démarche de gestion des risques.

Contributeurs

**Matthieu GRALL**, expert-conseil en management des données, sécurité de l’information, protection de la vie privée et nouvelles technologies

<autres contributeurs>

Ressources utilisées

**Page de garde** : Grid, par Magic Creative, de PIXABAY.

Versions

| Version | Action | Éditeur | État |
| --- | --- | --- | --- |
| 06/04/2025 | Création du document | Matthieu GRALL | Document de travail |
|  |  |  |  |

Documents de référence

Les références suivantes sont utilisées entre crochets dans le corps du document :

| Libellé court | Libellé long |
| --- | --- |
| [AI Act] | Proposition de Règlement du Parlement européen et du Conseil établissant des règles harmonisées concernant l’intelligence artificielle (législation sur l’intelligence artificielle) et modifiant certains actes législatifs de l’Union, COM/2021/206 final  🡪 [Lien](https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:52021PC0206) |
| [ATLAS] | *Adversarial Threat Landscape for Artificial-Intelligence SystemsTM (ATLAS)*, MITRE  🡪 [Lien](https://atlas.mitre.org/matrices/ATLAS) |
| [EBIOS Risk Manager] | Expression des besoins et identification des objectifs de sécurité – Méthode de gestion des risques, EBIOS *Risk Manager*, Agence nationale de la sécurité des systèmes d’information (ANSSI, 2022)  🡪 [Lien](https://cyber.gouv.fr/la-methode-ebios-risk-manager) |
| [EN 301 549] | Exigences d’accessibilité pour les produits et services ICT, *European Telecommunications Standards Institute* (ETSI, 2018)  🡪 [Lien](https://accessibilite.numerique.gouv.fr/doc/fr_301549v020102p.pdf) |
| [ISO/IEC 27001] | Sécurité de l’information, cybersécurité et protection de la vie privée – Systèmes de management de la sécurité de l’information – Exigences, *International Organization for Standardization* (ISO, 2022)  🡪 [Lien](https://www.iso.org/fr/standard/27001) |
| [ISO/IEC 27005] | Sécurité de l’information, cybersécurité et protection de la vie privée – Préconisations pour la gestion des risques liés à la sécurité de l’information, *International Organization for Standardization* (ISO, 2022)  🡪 [Lien](https://www.iso.org/fr/standard/80585.html) |
| [ISO/IEC 42001] | Technologies de l’information – Intelligence artificielle – Système de management, *International Organization for Standardization* (ISO, 2023)  🡪 [Lien](https://www.iso.org/fr/standard/81230.html) |
| [Loi I&L] | Loi n°78-17 du 6 janvier 1978 relative à l’informatique, aux fichiers et libertés, modifiée  🡪 [Lien](https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460/) |
| [NIST AI RMF] | *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, *National Institute of Standards and Technology* (NIST, 2024)  🡪 [Lien](https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf) |
| [Rapport Task Force IA] | *L’intelligence artificielle au service de la Défense*, Rapport de la Task Force IA, ministère des Armées (2019)  🡪 [Lien](https://www.defense.gouv.fr/sites/default/files/aid/20200108-NP-Rapport%20de%20la%20Task%20Force%20IA%20Septembre.pdf) |
| [Recommandation IA de l’OCDE] | Recommandation du Conseil sur l’intelligence artificielle, OECD/LEGAL/0449, Organisation de coopération et de développement économiques (OCDE, 2023)  🡪 [Lien](https://legalinstruments.oecd.org/fr/instruments/oecd-legal-0449) |
| [Recommandations IA de l’ANSSI - 2024] | Recommandations de sécurité pour un système d’IA générative, Agence nationale de la sécurité des systèmes d’information (ANSSI, 2024)  🡪 [Lien](https://cyber.gouv.fr/sites/default/files/document/Recommandations_de_s%C3%A9curit%C3%A9_pour_un_syst%C3%A8me_d_IA_g%C3%A9n%C3%A9rative.pdf) |
| [Recommandations IA de l’ANSSI - 2025] | Développer la confiance dans l’IA par une approche par les risques cyber (ANSSI, 2025)  🡪 [Lien](https://cyber.gouv.fr/sites/default/files/document/analyse_commune_haut_niveau_des_risques_cyber_ia.pdf) |
| [Recommandations IA de la CNIL] | Recommandations sur le développement des systèmes d’intelligence artificielle, Commission nationale de l’informatique et des libertés (CNIL, 2024)  🡪 [Lien](https://www.cnil.fr/fr/les-fiches-pratiques-ia) |
| [RGAA] | Référentiel général d’amélioration de l’accessibilité (RGAA), Direction interministérielle du numérique (DINUM, 2023)  🡪 [Lien](https://accessibilite.numerique.gouv.fr/doc/RGAA-v4.1.2.pdf) |
| [RGESN] | Référentiel général d’écoconception de services numériques (RGESN), Autorité de régulation de la communication audiovisuelle et numérique (ARCOM, 2024)  🡪 [Lien](https://ecoresponsable.numerique.gouv.fr/docs/2024/rgesn-mai2024/referentiel_general_ecoconception_des_services_numeriques_version_2024.pdf) |
| [RGI] | Référentiel général d’interopérabilité (RGI), Direction interministérielle du numérique (DINUM, 2020)  🡪 [Lien](https://www.numerique.gouv.fr/uploads/Referentiel_General_Interoperabilite_V2.pdf) |
| [RGIAF] | Référentiel général pour l’IA frugale (RGIAF), ministère de la transition écologique et de la cohésion des territoires (2024)  🡪 [Lien](https://greentechinnovation.fr/storage/2024/06/Referentiel-general-pour-lIA-frugale.pdf) |
| [RGPD] | Règlement (UE) 2016/679 du Parlement européen 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  🡪 [Lien](https://eur-lex.europa.eu/eli/reg/2016/679/oj?locale=fr) |
| [US EO 13960] | *Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government*, *Executive Order* 13960 (2020)  🡪 [Lien](https://trumpwhitehouse.archives.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/) |

Sommaire

[1 Introduction 6](#_Toc194868494)

[2 Bonnes pratiques 6](#_Toc194868495)

[2.1.1 Gouvernance de l’IA 7](#_Toc194868496)

[2.1.2 Robustesse 7](#_Toc194868497)

[2.1.3 Équité (fairness) 8](#_Toc194868498)

[2.1.4 Transparence et explicabilité 9](#_Toc194868499)

[2.1.5 Sécurité du système d’IA 10](#_Toc194868500)

[2.1.6 Protection des droits et libertés des personnes concernées 10](#_Toc194868501)

[2.1.7 Maintenabilité et évolutivité 10](#_Toc194868502)

[2.1.8 Interopérabilité 10](#_Toc194868503)

[2.1.9 Impact environnemental 10](#_Toc194868504)

[2.1.10 Accessibilité 11](#_Toc194868505)

# Introduction

<à rédiger>

# Bonnes pratiques

Le tableau suivant rassemble les bonnes pratiques contribuant à respecter les critères de confiance :

| Critère de confiance | Synthèse des recommandations | Exemples de bonnes pratiques |
| --- | --- | --- |
| Gouvernance responsable | Instaurer un cadre de gouvernance avec des rôles et responsabilités définis, un reporting régulier et des mécanismes d’audit interne pour contrôler l’éthique des processus. | Mettre en place des comités d’éthique, réaliser des audits internes et externes, publier régulièrement des rapports de gouvernance, former les responsables. (Source : AI Act, Chapitre 5 ANSSI, CNIL Fiche 3) |
| Robustesse, fiabilité, sûreté des personnes et des biens | Assurer la robustesse par des tests de résistance systématiques, la mise en place de redondances techniques et des procédures de réponse aux incidents. | Effectuer des tests de charge et de stress, simuler des attaques, établir des plans de continuité d’activité, certifier la sécurité. (Source : AI Act Art. 10, ANSSI Chap. 5.2) |
| Équité et non discrimination | Intégrer des mécanismes de vérification et d’audit des algorithmes pour détecter et corriger les biais. | Réaliser des audits algorithmiques, tester statistiquement les biais, consulter des experts en éthique et inclusion, élaborer des protocoles de correction. (Source : AI Act Art. 11, CNIL Fiche 2) |
| Transparence et explicabilité | Rendre explicites les algorithmes et les processus de décision via une documentation accessible, des interfaces interactives et des explications techniques adaptées aux différents publics. | Publier de la documentation technique, développer des interfaces explicatives, créer des tutoriels et guides d’usage, organiser des ateliers d’information. (Source : AI Act Art. 12, ANSSI Chap. 5.4) |
| Sécurité de l'information | Déployer une stratégie de sécurité robuste incluant des protocoles de chiffrement et une surveillance continue. | Implémenter des protocoles de chiffrement, activer l’authentification multi-facteurs, surveiller les accès en temps réel, réaliser des tests d’intrusion. (Source : AI Act Art. 13, ANSSI Chap. 5.5) |
| Protection des droits et libertés | Mettre en œuvre des politiques de minimisation des données et des mécanismes de consentement explicite. | Chiffrer les données personnelles, établir une politique de conservation limitée, anonymiser les données sensibles, recueillir un consentement clair. (Source : AI Act Art. 14, CNIL Fiche 7) |
| Maintenance et évolutivité | Mettre en place des mécanismes de mises à jour sécurisées et assurer une veille technologique continue. | Automatiser les mises à jour, surveiller les performances, prévoir des cycles d’évolution, assurer la compatibilité ascendante. (Source : AI Act Art. 15, ANSSI Chap. 5.7) |
| Interopérabilité | Standardiser les protocoles et formats de données pour assurer une interopérabilité fluide. | Utiliser des standards ouverts, respecter les API publiques, favoriser la portabilité des données, assurer la documentation des interfaces. (Source : AI Act Art. 16, CNIL Fiche 9) |
| Respect de l’environnement | Intégrer des pratiques visant à réduire l’impact environnemental en optimisant les ressources et la consommation énergétique. | Optimiser les algorithmes, choisir des data centers écologiques, suivre l’empreinte carbone, recycler les équipements. (Source : AI Act Art. 17, RGESN Chap. 9) |
| Accessibilité | Intégrer les principes d’accessibilité numérique dès la conception et tester régulièrement l’interface pour garantir une utilisation inclusive. | Appliquer les normes WCAG, concevoir des interfaces compatibles avec les lecteurs d’écran, proposer plusieurs modes d’interaction, tester avec des utilisateurs en situation de handicap. (Source : AI Act Art. 18, RGAA Chap. 1) |

**Critères de confiance** (Cxx) et explications / risques

**Bonnes pratiques** (BPyy)

### Gouvernance de l’IA

**Garantir les valeurs éthiques**

Impliquer les développeurs, usagers et autorités en matière d’IA pour assurer que les technologies liées à l'IA sont produites et utilisées conformément à des valeurs éthiques, en surveillant les préjugés, la confidentialité et les abus, tout en promouvant l'innovation et la confiance.

**Formaliser les responsabilités des parties intéressées**

Identifier toutes les parties intéressées (utilisateurs, organisations, développeurs d'IA, fournisseurs d'IA, fournisseurs de données, formateurs, institutions, etc.) et formaliser leurs responsabilités dans le cadre du processus mis en œuvre.

### Robustesse

**Vérifier les données d’entrée possibles**

Vérifier les particularités des données futures, par exemple à l’aide de modules spécifiques dans le *pipeline* de données.

**Vérifier la robustesse du modèle**

Vérifier que le modèle ne peut pas être attaqué pour produire un résultat indésirable, par exemple dans le cadre d’un *pipeline* de modèles.

**Sécuriser le système dans sa globalité**

Vérifier que la sécurité est proportionnée aux risques en amont et en aval du système d’IA.

**Évaluer les performances du système**

Faire des mesures quantitatives pour évaluer les performances du modèle dans diverses conditions de stress, notamment les attaques contradictoires, le bruit et les changements de distribution des données (ex : taux de faux positifs/négatifs, étude MSE, distance de Wasserstein, score de Brier).

**Mettre en place une démarche d’amélioration continue**

Entrainer, valider, surveiller, rendre transparent et améliorer en continu le système d’IA, notamment si l’intervention humaine est limitée ou inexistante, afin de réduire les risques, fournir des résultats cohérents et maintenir l’excellence des performances.

**Mettre en place les mesures de sûreté nécessaires**

Tester et valider le système d’IA pour garantir qu'il répond aux normes de sûreté et aux exigences réglementaires dans les applications où la sûreté est primordiale (ex : véhicules autonomes, soins de santé), intégrer plusieurs mesures redondantes pour gérer les pannes sans causer de dommages ni de perturbations significatives, laisser la main à l’humain, en particulier dans le cadre de prises de décisions critiques, pour intervenir, outrepasser ou guider les actions de l'IA si nécessaire.

### Équité (*fairness*)

#### Biais liés à la formulation du cas d’usage

L'équité peut être compromise par manque de cadre au cas d’usage.

**Définir clairement le(s) cas d’usage(s)**

Décrire précisément le cas d’usage, les limitations et les exceptions, et s’assurer que les personnes en charge de la conception et du développement des outils en ont connaissance et les comprennent.

#### Biais liés aux données d’entrée

Risque de prendre des décisions basées sur des données incomplètes ou non équilibrées.

Ex : un système de recommandation de produits en ligne qui ne propose que des produits populaires pour certains groupes démographiques, en excluant les produits moins connus, créant ainsi un biais dans les recommandations.

**Diversifier les données d’entrée**

Utiliser des sources de données diverses et représentatives.

**Nettoyer les données**

Nettoyer les données de manière approfondie.

#### Biais liés aux données d’entrainement

Biais introduits à partir des données d’entrainement et de validation en raison de l'iniquité inhérente aux données :

* biais dans les données de l'échantillon : si l'ensemble de données d’entrainement n'est pas représentatif de la population ;
* identification et transformation des caractéristiques sensibles : si un groupe défavorisé est présent dans l'échantillon, modifier les pondérations du modèle de manière à modifier le résultat pour ce groupe défavorisé ;
* biais dans la représentation des différentes classes ou catégories de données, ce qui peut entraîner des résultats non représentatifs, inexacts ou injustes ;
* biais dans les corrélations entre les caractéristiques ou les variables utilisées dans les données d'entraînement, ce qui peut conduire à des prédictions biaisées.

Ex : un algorithme de prédiction de probabilité de remboursement de prêt, qui accorde systématiquement des scores plus bas aux candidats issus de certaines communautés ethniques, en raison d'une sous-représentation de ces candidats dans les données d'entraînement.

**S’assurer de la qualité des données d’entrainement**

Vérifier la pertinence des données par rapport au problème à résoudre, leur fiabilité (crédibles et/ou de sources objectives), et la légalité de leur collecte et de leur traitement.

**Faire des échantillonnages équilibrés des données d’entrainement**

Respecter les bonnes pratiques d’échantillonnage pour que les échantillons soient équilibrés.

**Corriger les corrélations indésirables**

Identifier les corrélations entre les caractéristiques ou les variables, évaluer leur impact et gérer ceux qui ne sont pas acceptables.

**Collecter de nouvelles données dès que cela est nécessaire**

Recourir à de nouvelles données d’entrée jusqu’à avoir l’assurance que les biais sont suffisamment réduits.

#### Biais liés à l’algorithme d’entrainement / à l’architecture du modèle

Biais introduits à la suite de la conception du modèle, donnant des résultats trompeurs même à partir de données fiables et de qualité :

* construction du modèle : l'architecture du modèle elle-même peut présenter des problèmes inhérents, entraînant des biais tels que le biais de régression, le biais de classification, le biais de *clustering*, etc. Il peut y avoir des erreurs de calcul dans les paramètres du modèle, entraînant des modèles sur-ajustés/sous-ajustés, qui introduisent des biais et du bruit dans les données de sortie ;
* dérive du modèle : le cas d’usage peut évoluer avec le temps, de sorte que le modèle peut devenir obsolète et nécessiter une re-modélisation et un recyclage au fil du temps, réintroduisant les biais mentionnés ci-dessus à chaque étape.

Les principaux risques concernent des décisions discriminatoires ou injustes basées sur des caractéristiques personnelles ou des groupes de personnes spécifiques, la partialité ou les préjugés dans les résultats produits par l'algorithme.

Ex : un algorithme de recrutement basé sur l'IA qui favorise systématiquement les candidats masculins au détriment des candidates féminines, en raison de biais dans les critères de sélection.

**Évaluer la qualité du modèle**

Mesurer l'équité à l’aide de modèles mathématiques, comme l'indépendance statistique, les intervalles de confiance, l'étude de séparation et l'étude de suffisance.

**Évaluer les performances du modèle**

Évaluer les performances de l'algorithme de manière régulière.

**Faire auditer le modèle**

Réaliser un audit pour identifier et corriger les biais.

#### Biais liés aux données de sortie

Risque de produire des **résultats de sortie discriminatoires ou injustes**, qui peuvent avoir un impact négatif sur les utilisateurs ou les parties prenantes concernées

Risque de **renforcer les stéréotypes ou les préjugés** existants à travers les résultats produits par l'IA

Ex : Un système de détection automatique de fraude qui cible de manière disproportionnée certaines catégories de clients, en raison d'un biais dans les résultats de sortie, entraînant des décisions injustes

**Valider les données de sorties**

Valider les résultats produits, si possible de manière continue.

**Obtenir les retours des usagers**

Collecter régulièrement les commentaires et retours d'expérience des usagers, pour détecter et corriger les biais dans les résultats.

### Transparence et explicabilité

**Formaliser les éléments utiles à la transparence**

Élaborer les explications nécessaires à la transparence, et les faire connaître aux personnes chargées de créer et de maintenir les modèles et les flux de données, aux usagers (d’où proviennent les données, quelle forme elles prennent et comment l’organisation les utilise), et aux autorités compétentes. La transparence s’applique aux algorithmes (logique et modèle), aux interactions (via l’interface utilisateur), et à la société (impact social de cette interaction). L’explicabilité peut reposer sur des méthodes agnostiques ou spécifiques, intrinsèques ou *post-hoc*, locales ou globales, *a priori* ou *a posteriori*. Des outils tels que LIME ou SHAP peuvent notamment contribuer à cette explicabilité.

### Sécurité du système d’IA

**Mettre en œuvre les [Recommandations IA de l’ANSSI]**

Appliquer les [Recommandations IA de l’ANSSI] pour les systèmes d’IA générative.

**Homologuer la sécurité du système d’IA**

Évaluer la sécurité du système d’IA dans le cadre d’un processus d’homologation.

### Protection des droits et libertés des personnes concernées

**Mettre en œuvre les [Recommandations IA de la CNIL]**

Appliquer les [Recommandations IA de la CNIL] pour la phase de développement.

**Mettre le traitement en conformité avec la réglementation relative à la protection de la vie privée**

Contacter le correspondant RGPD pour bâtir la conformité du traitement de données à caractère personnel avec la [Loi I&L].

### Maintenabilité et évolutivité

**Adopter un principe de modularité et de réutilisabilité**

Mettre en œuvre une conception modulaire des systèmes d'IA, pour réutiliser/remplacer les composants du *pipeline* sans affecter l'ensemble du système.

**Mettre en place une documentation et un contrôle des versions**

Élaborer une documentation complète et à jour, comprenant des commentaires sur le code, des manuels d'utilisation et une documentation sur le prétraitement des données, la formation des modèles et les processus de déploiement. L'utilisation de systèmes de contrôle de version permet de suivre les modifications, de collaborer avec d'autres et de revenir aux versions précédentes si nécessaire.

**Contrôler la qualité du code**

Appliquer les bonnes pratiques de développement pour produire un code de haute qualité, en employant notamment des conventions de dénomination appropriées, un style cohérent et des algorithmes efficaces, pour faciliter la maintenance et les évolutions.

**Permettre l’évolutivité**

Concevoir le système pour gérer des charges accrues et évoluer selon les besoins. Ceci implique également le maintien des performances et la gestion des ressources.

### Interopérabilité

**Respecter les bonnes pratiques d’interopérabilité**

Appliquer le [RGI] et vérifier la conformité au [CCT] pour favoriser l’interopérabilité des technologies. Le cas échéant, faire les démarches nécessaires pour évaluer et intégrer les outils au [CCT].

**Vérifier l’existence des données dans l’ontologie afférente**

Identifier les sources de données, les données consommées, les traitements de données effectués et les données produites, et vérifier que les données sont déjà référencées dans l’ontologie afférente (ex : ontologie des données du domaine cyber). Le cas échéant, les intégrer dans celle-ci.

### Impact environnemental

**Respecter les bonnes pratiques d’écoconception**

Appliquer le [RGESN] et le [RGIAF], pour réduire globalement les besoins en ressources matérielles et énergétiques et les impacts environnementaux associés. Sont notamment identifiés la consommation d'énergie des centres de données lors de l’entrainement des modèles, l'économie circulaire du matériel, l'empreinte carbone, l'obsolescence et les déchets électroniques.

### Accessibilité

**Respecter les bonnes pratiques d’accessibilité**

Appliquer le [RGAA] ou l’[EN 301 549]. Cela implique notamment de concevoir des interfaces ergonomiques, en tenant compte de facteurs tels que la diversité linguistique et l’accessibilité pour les personnes handicapées. En se concentrant sur l'accessibilité et l'autonomisation des utilisateurs, les développeurs peuvent favoriser une expérience d'IA positive et inclusive pour tous les utilisateurs.