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

Le contenu généré par l’IA peut être incorrect.

Intelligence artificielle

Critères de confiance

Contributeurs

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

<autres contributeurs>

Versions

| Version | Action | Éditeur | État |
| --- | --- | --- | --- |
| 06/04/2025 | Création du document | Matthieu GRALL | Document de travail |
| 10/04/2025 | Transformation des « descriptions » des critères de confiance en « portées » (pour mieux distinguer « portées », qui désignent ce qui est couvert, et « objectifs », qui sont présentés dans les bonnes pratiques), corrections mineures (mises en cohérence avec les autres documents) | Matthieu GRALL | Document de travail |

Documents de référence

| Libellé court | Libellé long |
| --- | --- |
| [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 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) |
| [Principes de l’OCDE sur l’IA] | Principes de l’OCDE sur l’IA (2019)  🡪 [Lien](https://www.oecd.org/fr/topics/sub-issues/ai-principles.html) |
| [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) |
| [Règlement IA] | Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l’intelligence artificielle […] (règlement sur l’intelligence artificielle, 2024)  🡪 [Lien](https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX:32024R1689) |
| [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) |

Ressources utilisées

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

Sommaire

[1 Objet du document 4](#_Toc195099061)

[2 Introduction 4](#_Toc195099062)

[3 Critères de confiance 5](#_Toc195099063)

[3.1 Gouvernance responsable 5](#_Toc195099064)

[3.2 Fiabilité et sûreté 5](#_Toc195099065)

[3.3 Équité 5](#_Toc195099066)

[3.4 Transparence 5](#_Toc195099067)

[3.5 Sécurité des informations 6](#_Toc195099068)

[3.6 Protection des droits et libertés 6](#_Toc195099069)

[3.7 Maintenance et évolutivité 6](#_Toc195099070)

[3.8 Interopérabilité 6](#_Toc195099071)

[3.9 Respect de l’environnement 6](#_Toc195099072)

[3.10 Accessibilité 6](#_Toc195099073)

# Objet du document

**Ce document propose une liste de critères de confiance, qui harmonise les différents critères des nombreux documents de références en matière d’intelligence artificielle (IA)**.

Il s’inscrit dans un ensemble de documents méthodologiques en amélioration continue, destinés à aider les organismes à gérer les risques liés à l’IA, et qui peuvent être utiles ensemble ou séparément :

1. [Exemples de micro-cas d’usages de l’IA] 🡪 Lien ;
2. [Critères de confiance de l’IA] 🡪 Lien ;
3. [Bonnes pratiques de l’IA] 🡪 Lien ;
4. [Méthode de gestion des risques de l’IA] 🡪 Lien.

Il sert de base à la constitution de bonnes pratiques harmonisées, et pourrait également servir à affiner les critères utilisés dans une démarche de gestion des risques.

# Introduction

**Pour obtenir une liste de critères de confiance des systèmes basés sur l’IA, on se heurte à la pluralité des principaux documents de référence**[[1]](#footnote-1) qui comprennent des exigences, règles et recommandations liées à l’IA ou applicables à l’IA. Ils ne sont pas vraiment cohérents, que ce soit en termes de champs d’application, de formulations, de classements, et de langue. **On a donc de nombreuses redondances** dans l’ensemble et de nombreux manques dans chaque, si on souhaite une vision globale.

**Toutefois, les idées convergent toutes, ou se complètent plutôt bien**. On peut donc faire émerger une liste de ces critères de confiance, exhaustive et « non recouvrante », qui traite de l’ensemble des aspects qui peuvent devoir être considérés quand on crée un système basé sur l’IA.

Ainsi, tout système basé sur l’IA devrait :

1. reposer sur une **gouvernance responsable** (notion d’*accountability*) ;
2. assurer la **fiabilité**, et la **sûreté** des personnes et des biens ;
3. être **équitable** (notion de *fairness*) ;
4. garantir la **transparence** ;
5. assurer la **sécurité des informations** ;
6. protéger les **droits et libertés des personnes** ;
7. assurer la **maintenance** et l’**évolutivité** ;
8. permettre l’**interopérabilité** ;
9. respecter l’**environnement** ;
10. assurer l’**accessibilité**.

# Critères de confiance

## Gouvernance responsable

Portée : pilotage éthique, transparent et responsable, incluant la mise en place de mécanismes de contrôle, de *reporting* et de gestion des conflits d’intérêts.

Principaux risques : décisions unilatérales, conflits d’intérêts, manque de supervision, opacité décisionnelle.

## Fiabilité et sûreté

Portée : robustesse, performance, stabilité, résilience, précision, absence d'erreurs, capacité à fonctionner correctement, sans mettre en danger la vie humaine, les biens ou l'environnement, dans des conditions variées, incertaines ou inattendues, même en conditions de stress ou d’attaque.

Principaux risques : dysfonctionnements, accidents, attaques malveillantes, défaillances techniques.

## Équité

Portée : fonctionnement de l’IA sans biais, traitement impartial et équitable des usagers, non-discrimination.

Principaux risques : discriminations ou inégalités d’accès, du fait de biais dans la formulation des cas d’usages[[2]](#footnote-2), liés aux données d’entrée[[3]](#footnote-3), aux données d’entrainement et de validation[[4]](#footnote-4), à l’algorithme d’entrainement / à l’architecture du modèle[[5]](#footnote-5), aux données de sortie[[6]](#footnote-6).

## Transparence

Portée : compréhension des processus et décisions, explicabilité, traçabilité des données, possibilité de contester et vérifier le fonctionnement.

Principaux risques : opacité, incompréhension des mécanismes, perte de confiance.

## Sécurité des informations

Portée : protection de la disponibilité, de l’intégrité et de la confidentialité des données, gestion des risques liés à la sécurité de l’information engendrés par les systèmes d’IA (au-delà du critère de confiance de fiabilité).

Principaux risques : cyberattaques, fuites de données, corruption de systèmes.

## Protection des droits et libertés

Portée : respect de la vie privée, protection des données à caractère personnel et des droits fondamentaux, gestion des risques sur les droits et libertés des personnes concernées engendrés par les systèmes d’IA (au-delà du critère de confiance d’équité).

Principaux risques : atteintes à la vie privée, usage abusif des données personnelles.

## Maintenance et évolutivité

Portée : maintien en conditions opérationnelle et de sécurité, adaptation du système d’IA au fil du temps pour résoudre des problèmes et être utilisé pour de nouveaux besoins.

Principaux risques : obsolescence logicielle, incompatibilités, pertes de performance.

## Interopérabilité

Portée : compatibilité entre différents systèmes d'IA, intégration du système d’IA avec d’autres outils et normes existants, capacité à échanger des informations.

Principaux risques : verrouillage technologique, incompatibilités de formats[[7]](#footnote-7).

## Respect de l’environnement

Portée : maîtrise de l’empreinte écologique globale associée au développement, au déploiement et à l’exploitation de technologies d’IA, maîtrise de la consommation énergétique, gestion des risques sur l’environnement engendrés par les systèmes d’IA.

Principaux risques : consommation excessive d’énergie, émissions de CO₂.

## Accessibilité

Portée : accès aux systèmes d'IA, notamment pour les personnes handicapées, gestion des risques d’inégalités engendrés par les systèmes d’IA.

Principaux risques : exclusion numérique, inégalités d’accès, manque de compatibilité avec les technologies d’assistance.

1. Notamment [Règlement IA], [ISO/IEC 42001], [Lignes directrices IA de l’UE], [Loi I&L], [NIST AI RMF], [Principes de l’OCDE sur l’IA], [Recommandation IA de L’OCDE], [Recommandations IA de l’ANSSI - 2024], [Recommandations IA de l’ANSSI - 2025], [Recommandations IA de la CNIL], [RGAA], [RGESN], [RGI], et [RGIAF]. [↑](#footnote-ref-1)
2. Manque d’équité par défaut de cadrage du cas d’usage. [↑](#footnote-ref-2)
3. Prises de décisions basées sur des données incomplètes ou non équilibrées. [↑](#footnote-ref-3)
4. 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 (ce qui peut conduire à des prédictions biaisées). [↑](#footnote-ref-4)
5. Biais introduits à la suite de la conception du modèle, donnant des résultats trompeurs même à partir de données fiables et de qualité, du fait de la 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/sous-ajustés, qui introduisent des biais et du bruit dans les données de sortie) ou de la 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 impacts potentiels 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. [↑](#footnote-ref-5)
6. 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. [↑](#footnote-ref-6)
7. Comme tout service numérique, ceux qui reposent sur l’IA peuvent ne pas pouvoir interagir avec les autres si les technologies ou les données utilisées ne sont pas cohérentes ou compatibles entre elles. [↑](#footnote-ref-7)