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

Le contenu généré par l’IA peut être incorrect.

Intelligence artificielle

Critères de confiance

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 critères de confiance, qui harmonise les différents critères des nombreux documents de références en matière d’IA**.

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.

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

| 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) |
| [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) |
| [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) |

Sommaire

[1 Introduction 4](#_Toc194867660)

[2 Critères de confiance 5](#_Toc194867661)

[2.1 Gouvernance responsable 5](#_Toc194867662)

[2.2 Robustesse, fiabilité et sûreté 5](#_Toc194867663)

[2.3 Équité et non-discrimination 5](#_Toc194867664)

[2.4 Transparence et explicabilité 5](#_Toc194867665)

[2.5 Sécurité de l'information 5](#_Toc194867666)

[2.6 Protection des droits et libertés 6](#_Toc194867667)

[2.7 Maintenance et évolutivité 6](#_Toc194867668)

[2.8 Interopérabilité 6](#_Toc194867669)

[2.9 Respect de l’environnement 6](#_Toc194867670)

[2.10 Accessibilité 6](#_Toc194867671)

# 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 **robustesse**, la **fiabilité**, et la **sûreté** des personnes et des biens ;
3. être **équitable** et **non-discriminant** (notion de *fairness*) ;
4. garantir la **transparence** et l’**explicabilité** ;
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

**Que recouvre ce critère de confiance ?**

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

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Décisions unilatérales, conflits d’intérêts, manque de supervision, opacité décisionnelle.

## Robustesse, fiabilité et sûreté

**Que recouvre ce critère de confiance ?**

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.

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Dysfonctionnements, accidents, attaques malveillantes, défaillances techniques.

## Équité et non-discrimination

**Que recouvre ce critère de confiance ?**

Veiller à ce que l’IA fonctionne sans biais, garantissant un traitement impartial et équitable de tous les utilisateurs et évitant toute forme de discrimination.

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Biais dans la formulation des cas d’usages, liés aux données d’entrée, aux données d’entrainement, à l’algorithme d’entrainement, à l’architecture du modèle, aux données de sortie, discrimination systémique, inégalités d’accès.

## Transparence et explicabilité

**Que recouvre ce critère de confiance ?**

Communication claire des processus et décisions de l’IA, permettant aux utilisateurs de comprendre, contester et vérifier son fonctionnement, traçabilité des données.

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Opacité, incompréhension des mécanismes, perte de confiance.

## Sécurité de l'information

**Que recouvre ce critère de confiance ?**

Protection de la disponibilité, de l’intégrité et de la confidentialité des données, notamment contre les attaques, gestion des risques liés à la sécurité de l’information engendrés par les systèmes d’IA, tout le long du cycle de vie de ces systèmes.

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Cyberattaques, fuites de données, corruption de systèmes.

## Protection des droits et libertés

**Que recouvre ce critère de confiance ?**

Respect de la vie privée, protection des données personnelles et des droits fondamentaux, gestion des risques sur les droits et libertés des personnes concernées engendrés par les systèmes d’IA.

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Atteintes à la vie privée, usage abusif des données personnelles.

## Maintenance et évolutivité

**Que recouvre ce critère de confiance ?**

Capacité à maintenir en conditions opérationnelles et de sécurité, et à adapter les systèmes d'IA au fil du temps pour résoudre des problèmes ou être appliqués à de nouveaux besoins, en intégrant des mises à jour et des améliorations.

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Obsolescence logicielle, incompatibilités, pertes de performance.

## Interopérabilité

**Que recouvre ce critère de confiance ?**

Compatibilité entre différents systèmes d'IA, intégration des systèmes d’IA avec d’autres outils et normes existants, capacité à échanger des informations (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).

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Verrouillage technologique, incompatibilités de formats.

## Respect de l’environnement

**Que recouvre ce critère de confiance ?**

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

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Consommation excessive d’énergie, émissions de CO₂.

## Accessibilité

**Que recouvre ce critère de confiance ?**

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

**Quels sont les principaux risques que ce critère de confiance contribue à traiter ?**

Exclusion numérique, inégalités d’accès, manque de compatibilité avec les technologies d’assistance.

1. [AI ACT], [ISO/IEC 42001], [Loi I&L], [NIST AI RMF], [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], [RGIAF]. [↑](#footnote-ref-1)