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

Le contenu généré par l’IA peut être incorrect.

Intelligence artificielle

Une méthode 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>

Versions

| Version | Action | Éditeur | État |
| --- | --- | --- | --- |
| 28/03/2025 | Création du document, ajout des documents de références, du tableau périodique de cas d’usages et des critères de confiance | Matthieu GRALL | Document de travail |
| 05/04/2025 | Déplacement du tableau périodique de cas d’usages dans un autre document | Matthieu GRALL | Document de travail |
| 10/04/2025 | Déplacement des critères de confiance et des bonnes pratiques dans d’autres documents, améliorations mineures (mise en cohérence avec les autres documents) | Matthieu GRALL | Document de travail |

Documents de référence

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

| Libellé court | Libellé long |
| --- | --- |
| [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) |
| [Guide d’homologation de l’ANSSI] | Guide d’homologation de sécurité des systèmes d’information, Agence nationale de la sécurité des systèmes d’information (ANSSI, 2025)  🡪 [Lien](https://cyber.gouv.fr/publications/lhomologation-de-securite-des-systemes-dinformation) |
| [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) |
| [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) |

Ressources utilisées

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

Sommaire

[1 Objet du document 4](#_Toc195182875)

[2 Introduction 4](#_Toc195182876)

[3 Étape 1 : Estimer le niveau de risque du projet basé sur l’IA 5](#_Toc195182877)

[4 Étape 2 : l’approche par conformité 6](#_Toc195182878)

[5 Étape 3 : l’approche par scénarios 7](#_Toc195182879)

# Objet du document

**Ce document propose une démarche pour gérer les risques spécifiques aux services numériques qui reposent sur de l’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 a pour vocation à s’inscrire dans les démarches existantes au sein des organismes, notamment les processus d’homologation de systèmes, mais peut également être directement utilisé.

# Introduction

Dans le domaine cyber, on peut autant considérer l’IA comme :

1. **une technologie, comme les autres**, sur laquelle de nouveaux services numériques vont pouvoir reposer, et dont il conviendra d’apprécier et de traiter les risques spécifiques ;
2. **une réelle opportunité d’améliorer nos capacités cyber** dans tous les domaines de lutte ;
3. **un levier démultiplicateur des capacités offensives adverses**, contre lesquelles on va devoir lutter.

Pour les deux premiers points, l’enjeu est le même : **améliorer la confiance envers l’IA** !

Or, un projet basé sur l’IA est susceptible d’engendrer des **risques qui ne se limitent pas à la sécurité de l’information**, et **amplifiés par la frénésie actuelle** qui réduit la prise de recul :

* **sur les organismes**, du fait de défauts de qualité des données et légalité de leur obtention / traitement, et de technologies qui apportent leurs lots de vulnérabilités ;
* **sur les droits et libertés des personnes**, avec notamment des biais sur les données d’entrainement, d’entrée et de sortie, qui peuvent mener à des discriminations ;
* **sur l’environnement**, car les technologies sur lesquelles reposent les outils d’IA sont parfois très gourmandes en ressources.

Ce document propose donc **une méthode de gestion des risques liés à l’IA**, qui :

1. généralise les principes du [Règlement IA] en estimant *a priori* le **niveau de risque sur l’organisme, sur les personnes et sur l’environnement** ;
2. compare les pratiques envisagées aux bonnes pratiques via une **approche par conformité** ;
3. pour les projets susceptibles d’engendrer les risques les plus élevés, explique comment les apprécier et les traiter via une **approche par scénarios**.

Cette méthode :

* s’inscrit dans les démarches d’homologation existantes (cf. [Guide d’homologation de l’ANSSI]) ;
* repose sur la méthode [EBIOS Risk Manager] ;
* respecte donc les principes de gestion des risques (ex : [ISO/IEC 27005]) ;
* contribue à satisfaire les exigences afférentes des systèmes de management (notamment [ISO/IEC 27001] et [ISO/IEC 42001]).

# Étape 1 : Estimer le niveau de risque du projet basé sur l’IA

L’objectif est de **déterminer *a priori* le niveau de risque que le projet est susceptible d’engendrer**.

Pour ce faire, il convient de :

1. **situer le projet** dans chaque colonne de l’échelle suivante :

| **Niveau de risque** | **Impacts potentiels sur les personnes** (cf. [Règlement IA]) | **Impacts potentiels sur l’organisme** en cas de disparition de données, de modification non désirée de données, ou d’accès non autorisé à des données | **Impacts potentiels sur l’environnement** |
| --- | --- | --- | --- |
| 1. Minimal | « Risque minimal ou nul »  Ex : filtres anti-spam, IA gadget | Perturbation très limitée Aucune donnée sensible Rétablissement rapide Aucun impact légal ou réputationnel | Usage local, faible consommation énergétique Pas d'entraînement intensif Empreinte carbone négligeable Impact potentiellement positif |
| 2. Faible | « Risque limité »  Ex : chatbot, IA générative non critique | Dégradation temporaire Données peu sensibles Intervention rapide Risques faibles | Utilisation modérée du cloud Empreinte maîtrisable Usage de données/matériel à faible intensité Compensation possible |
| 3. Élevé | « Risque élevé »  Ex : IA pour santé, emploi, justice | Compromission de données sensibles Interruption prolongée Gestion de crise nécessaire Risques juridiques et réputationnels | Calculs intensifs Stockage/énergie importants Pression sur ressources matérielles Impact environnemental notable |
| 4. Maximal | « Risque inacceptable »  Ex : notation sociale, manipulation, surveillance biométrique de masse | Fuite massive de données critiques Dysfonctionnement généralisé Impact légal/réputationnel majeur Risque pour la pérennité | Modèles massifs (ex : LLM) Consommation énergétique élevée Externalisation des impacts Pas de compensation |

1. **retenir le niveau le plus élevé** de toutes les colonnes ;
2. **déterminer les suites à donner** en fonction de ce niveau :

1. Minimal  🡪 sans suite ;

2. Faible  🡪 étape 2 conseillée ;

3. Élevé  🡪 étape 2, étape 3 conseillée ;

4. Maximal  🡪 étape 2 et 3.

Notes :

* cette approche ne préjuge en rien des obligations et interdictions applicables (ex : « IA à risque inacceptable » interdite par le [Règlement IA]) ;
* le cas échéant, le résultat de cette étape peut être intégré à la stratégie d’homologation.

# Étape 2 : l’approche par conformité

L’approche par conformité, à mettre en œuvre dans le cas d’un niveau 3. Élevé (également conseillée pour 2. Faible ), permet de **gérer les risques standards**, y compris ceux de cause accidentelle, **et les attaques cyber non ciblées**.

L’objectif est d’**évaluer la conformité du projet aux [Bonnes pratiques de l’IA]**, qui contribuent à respecter les [Critères de confiance de l’IA] d’un système qui repose sur l’IA, afin d’éclairer la prise de décision.

Pour ce faire, évaluer chacune des [Bonnes pratiques de l’IA] :

* **si elle est jugée comme applicable :**

* + **si elle est appliquée, comment ?** L’explication fournie doit permettre d’évaluer le respect de la bonne pratique ;

* + **si elle n’est pas appliquée, quelles sont les mesures compensatoires ?** L’explication fournie doit permettre d’évaluer que les mesures prévues sont suffisantes pour atteindre un niveau de confiance aussi bon que si la bonne pratique était appliquée ;

* **si elle est jugée comme non applicable, pourquoi ?** L’explication fournie doit permettre de juger de son inapplicabilité.

La déclaration d’applicabilité en annexe des [Bonnes pratiques de l’IA] peut être utilisée à cet effet.

Notes :

* l’objectif n’est pas de respecter toutes les bonnes pratiques, mais de décrire ce qui est prévu au regard de celles-ci ;
* cette approche correspond à la « déclaration d’applicabilité » de l’[ISO/IEC 27001] et à l’« évaluation du socle de règles » de l’atelier 1 d’[EBIOS *Risk Manager*] ;
* le cas échéant, le résultat de cette étape peut être intégrée au dossier d’homologation.

# Étape 3 : l’approche par scénarios

L’approche par scénarios, à mettre en œuvre dans le cas d’un niveau 4. Maximal (également conseillée pour le niveau 3. Élevé ), permet de **gérer les attaques cyber avancées et ciblées**.

L’objectif est d’**identifier et** d’**apprécier les risques** que le projet est susceptible d’engendrer sur l’organisme, les personnes et l’environnement, de **déterminer les mesures** pour les traiter, et de **présenter les risques résiduels** pour éclairer la prise de décision.

Pour ce faire, **employer [EBIOS *Risk Manager*] sur le projet avec les spécificités suivantes** :

* dans l’atelier 1 :
  + considérer le **service numérique d’IA** comme « objet de l’étude » ;
  + considérer sa **finalité** comme « mission » ;
  + intégrer les **données d’entrée, d’entrainement et de sortie** dans les « valeurs métier » ;
  + intégrer le **système d’IA**, notamment l’algorithme utilisé, dans les « biens supports » ;
  + considérer les **impacts sur l’organisme**, mais aussi **sur les personnes** et **sur l’environnement**, dans les « événements redoutés » ;
* dans l’atelier 4 : considérer les **attaques types d’[ATLAS]** dans les « scénarios opérationnels » ;
* dans l’atelier 5 : considérer les **mesures types d’[ATLAS]** dans les « mesures », ainsi que les recommandations[[1]](#footnote-1) qui n’auraient pas été retenues dans l’étape précédente.

Notes :

* cette approche correspond donc à l’ensemble ateliers d’[EBIOS *Risk Manager*] et aux processus d’ « établissement du contexte », d’ « appréciation des risques » et de « traitement des risques » des normes relatives à la gestion des risques (ex : [ISO/IEC 27005]) ;
* le cas échéant, le résultat de cette étape peut constituer le cœur du dossier d’homologation.

1. [Recommandations IA de l’ANSSI - 2024], [Recommandations IA de l’ANSSI - 2025], [Recommandations IA de la CNIL], etc. [↑](#footnote-ref-1)