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

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 démarche pour gérer les risques spécifiques aux systèmes/services numériques qui reposent sur de l’IA**.

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.

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 |
| --- | --- | --- | --- |
| 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 |
| 06/04/2025 | Déplacement des critères de confiance et des bonnes pratiques dans un autre 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 La méthode en synthèse 5](#_Toc194867525)

[2 La méthode en détail 6](#_Toc194867526)

[2.1 Estimer le niveau de risque du cas d’usage 6](#_Toc194867527)

[2.2 Mettre en œuvre une « approche par conformité » aux bonnes pratiques 7](#_Toc194867528)

[2.3 Mettre en œuvre une « approche par scénarios » pour les cas susceptibles d’engendrer les risques les plus élevés 8](#_Toc194867529)

[Annexes 9](#_Toc194867530)

[Bonnes pratiques 9](#_Toc194867531)

[2.3.1 Gouvernance de l’IA 10](#_Toc194867532)

[2.3.2 Robustesse 10](#_Toc194867533)

[2.3.3 Équité (fairness) 11](#_Toc194867534)

[2.3.4 Transparence et explicabilité 12](#_Toc194867535)

[2.3.5 Sécurité du système d’IA 13](#_Toc194867536)

[2.3.6 Protection des droits et libertés des personnes concernées 13](#_Toc194867537)

[2.3.7 Maintenabilité et évolutivité 13](#_Toc194867538)

[2.3.8 Interopérabilité 13](#_Toc194867539)

[2.3.9 Impact environnemental 13](#_Toc194867540)

[2.3.10 Accessibilité 14](#_Toc194867541)

# La méthode en synthèse

La logique est la suivante : **la rigueur de la réflexion et des mesures est proportionnée aux risques qu’un cas d’usage engendre** (sur l’organisme, les personnes concernées, et l’environnement).

Pour appliquer cette logique générale, la démarche est la suivante :

1. **le niveau de risque est préalablement estimé** à l’aide de critères de qualification (cf. [AI Act]) ; puis, selon cette estimation :
   1. les cas d’usages susceptibles d’engendrer les **risques les plus faibles** se limiteront à une **« approche par conformité »** aux bonnes pratiques permettant de respecter les critères de confiance des systèmes qui reposent sur l’IA ;
   2. les cas d’usages susceptibles d’engendrer des **risques élevés** intègreront en outre une **« approche par scénarios »** dans le cadre de l’étude des risques.

C’est donc une démarche en 2 ou 3 étapes :

Une image contenant texte, capture d’écran, Police, Graphique

Le contenu généré par l’IA peut être incorrect.

Note : cette démarche ne préjuge pas de l’accord des services en charge de la conformité à la réglementation, et encore moins de la conformité du traitement de données à la réglementation. Mais elle permet de leur **fournir les éléments utiles à la mise en conformité**.

# La méthode en détail

## Estimer le niveau de risque du cas d’usage

Une image contenant texte, capture d’écran, Police, Graphique

Le contenu généré par l’IA peut être incorrect.

L’objectif est ici de **qualifier le cas d’usage selon le niveau de risque qu’il est susceptible d’engendrer**.

En se basant sur les catégories identifiées dans l’[AI Act], et notamment au regard des données traitées (types et volumes) et des impacts potentiels sur les personnes concernées, sont considérés comme :

* **cas d’usages susceptibles d’engendrer des risques élevés** les catégories suivantes :

* + « *IA à risque inacceptable* » : ces systèmes contreviennent aux valeurs de l’Union européenne et portent atteinte aux droits fondamentaux. Il s’agit notamment des systèmes de *scoring* social à usage général ;

* + « *IA à haut risque* », correspondant à 8 grands domaines d’activité, notamment :
    - des systèmes de *scoring* social ;
    - des identifications biométriques ;
    - des systèmes de manipulation de contenu ;
    - des infrastructures susceptibles de mettre en danger la vie et la santé des personnes (ex : système de navigation dans les transports) ;
    - des composants de sécurité des produits (ex : IA de chirurgie) ;

* **cas d’usages susceptibles d’engendrer les** **risques les plus faibles** les catégories suivantes :

* + « *IA à risque faible* » : système d’IA interagissant avec les personnes physiques et n’étant ni à risque inacceptable, ni à haut risque (ex : *deepfakes* à vocation artistique, *chatbots*) ;

* + « *IA à risque minime* » : tous les autres systèmes d’IA n’entrant pas dans les autres catégories (ex : filtre anti-spam).

La qualification doit être **argumentée**.

Note : qu’il s’agisse d’un traitement de données à caractère personnel ou non, **l’esprit de cette échelle est utilisé pour qualifier tout cas d’usage reposant sur de l’IA**. Le résultat peut utilement être **intégré à la stratégie d’homologation**.

## Mettre en œuvre une « approche par conformité » aux bonnes pratiques

Une image contenant texte, capture d’écran, Police, Graphique

Le contenu généré par l’IA peut être incorrect.

L’approche par conformité, à mettre en œuvre dans tous les cas d’usages, permet de **gérer les risques standards et non ciblés**, y compris ceux de cause accidentelle.

L’objectif est ici d’**évaluer la conformité du cas d’usage aux bonnes pratiques** (voir l’annexe « Critères de confiance et bonnes pratiques »), qui contribuent à respecter les critères de confiance d’un système qui repose sur l’IA (voir l’annexe « Comparaison des critères à respecter »).

Pour ce faire, chaque bonne pratique est évaluée :

* **si elle est 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 au moins aussi bon que si la bonne pratique était appliquée ;

* **si elle n’est pas applicable, pourquoi n’est-elle par applicable ?** L’explication fournie doit permettre de juger de son inapplicabilité.

L’annexe « Déclaration d’applicabilité » peut être utilisée à cet effet.

Note : 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]. Elle peut ainsi utilement être **intégrée au dossier d’homologation**.

## Mettre en œuvre une « approche par scénarios » pour les cas susceptibles d’engendrer les risques les plus élevés

# Annexes

## Échelle d’estimation

| Niveau | Impact sur la vie privée (cf. [AI Act]) | Impact lié à la sécurité de l’information | Impact environnemental |
| --- | --- | --- | --- |
| 1. Minimal | Risque minimal ou nul Ex : filtres anti-spam, IA gadget Aucune exigence réglementaire | 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 Obligation de transparence | 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 Évaluation stricte et supervision humaine | 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 Interdiction totale dans l’UE | Fuite massive de données critiques Dysfonctionnement généralisé Impact légal/réputationnel majeur Risque pour la pérennité | Modèles massifs (LLM) Consommation énergétique élevée Externalisation des impacts Pas de compensation |