# Déjouer les biais de l’Intelligence Artificielle (IA) par la pensée critique

**Prix** : 1 150 €HT

**Durée** : 1 jour

**Code de Référence** : IA008

Catalogue Intelligence Artificielle

Cette formation d’une journée s’adresse à toute personne souhaitant mieux comprendre les biais présents dans les systèmes d’Intelligence Artificielle (IA), notamment générative, et développer une posture critique face à leurs résultats. Elle permet de distinguer les différents types de biais (données, algorithmiques, interactifs, systémiques) et d’explorer des outils concrets pour les détecter, les analyser et les mitiger.

La matinée est consacrée à l’identification des biais : typologies, démonstrations en direct, cas d’usage concrets, et introduction à des techniques d’analyse critique (méthode QQOQCP, « lateral reading », vérification de sources). Les participants découvrent également les outils de détection des biais dans les modèles de langage (LLM), comme Bias Bench ou Aequitas, qu’ils appliquent lors d’ateliers pratiques.

L’après-midi explore des stratégies concrètes de réduction des biais : prompt engineering, boucles de rétroaction, enrichissement des données. Un point est également fait sur le cadre réglementaire (RGPD, AI Act, ISO 42001) et les responsabilités juridiques et éthiques des organisations. Les participants élaborent
une checklist d’évaluation applicable à leur contexte.

La journée se termine par un audit collectif d’un cas réel, permettant de mobiliser toutes les compétences acquises : identification, mesure, mitigation et restitution stratégique. Cette formation outille les participants pour adopter une approche plus lucide, éthique et responsable dans leurs interactions avec l’IA.

## Objectifs de la formation

A l’issue de cette formation, vous serez capable de :

- Identifier les fondements de l’esprit critique
- Comprendre les biais cognitifs et algorithmiques
- Adopter une posture critique, éthique et stratégique face à l’Intelligence Artificielle Générative (IAG)

## Public

Cette formation s’adresse à toutes les personnes souhaitant renforcer son esprit critique face à l’IA.

## Prérequis

Aucun prérequis

## Programme de la formation

**MATIN (3H30)**

**Introduction aux biais en IA – 30 MIN**

- Définitions clés
- Exemples récents de biais médiatisés
    - Démonstration en direct d’une réponse IA biaisée – Atelier : Identifier un biais simple dans une réponse ChatGPT

**Typologies de biais – 45 MIN**

- Biais liés aux données (échantillonnage, représentativité)
- Biais algorithmiques (optimisation, fonctionnalités latentes)
- Biais d’interaction utilisateur (confirmation, suggestions)
- Biais systémiques (culture, réglementation, marché)
    - Atelier : Cartographier les biais possibles pour un cas métier

**Pensée critique appliquée – 1H**

- Méthodes QQOQCP et « lateral reading »
- Triangulation et vérification de sources
- Grilles d’analyse rapide
    - Atelier : Évaluer la fiabilité d’une réponse à forte incertitude

**Détection de biais dans les LLM – 1H15**

- Tests A/B et « prompts sentinelles »
- Indicateurs : toxicity, demographic parity
- Outils open source : Bias Bench, Aequitas
    - Atelier : Appliquer un outil de scoring sur des réponses générées

**APRÈS-MIDI (3H30)**

**Stratégies de mitigation – 1H15**

- Curations et enrichissement des données
- Prompt engineering anti-biais
- Post-traitement et boucles de rétroaction humaine
    - Atelier : Réécrire un prompt et post-traiter la sortie pour réduire un biais

**Cadre juridique, éthique et responsabilité – 45 MIN**

- RGPD, AI Act, normes ISO/IEC 42001

- Responsabilité éditoriale et transparence
- Principes de sécurité et de gouvernance
    - Atelier : Élaborer une check-list éthique pour son organisation

**Atelier de synthèse : Audit complet d'un cas réel – 1H30**

- Analyse complète d'une interaction IA : détection, mesure et mitigation
    - Activités : Travail en équipe sur un cas réel – Restitution rapide et élaboration d'un plan d'actions.

## Méthodes pédagogiques

Cette formation est rythmée par une alternance d'exposés et de travaux pratiques. Les exercices réalisés lors des travaux pratiques permettent la mise en œuvre des connaissances acquises.

## Méthodes d'évaluation des acquis

Afin d'évaluer l'acquisition de vos connaissances et compétences, il vous sera envoyé un formulaire d'auto-évaluation, qui sera à compléter en amont et à l'issue de la formation.
Un certificat de réalisation de fin de formation est remis au stagiaire lui permettant de faire valoir le suivi de la formation.