ACADÉMIE HUMANITAIRE · HUMANITAIRES JOBS

# Mini MBA IA et action humanitaire : travailler avec Claude et Gemini

## Support de cours · Semaine 4 : Éthique de l'IA et principes humanitaires : ne pas nuire, biais, transparence

Programme de 3 mois exactement (12 semaines, 144 heures) pour utiliser les assistants d'IA Claude (Anthropic) et Gemini (Google) dans le travail humanitaire, de façon utile, sûre et responsable : fonctionnement et limites, formulation des demandes, protection des données, éthique et principes humanitaires, rapports, propositions, analyse de données et MEAL, traduction et communication avec les communautés, recherche et vérification des sources, désinformation et sécurité, productivité et automatisations, politique d'usage et déploiement dans une organisation. Exercices corrigés, trois examens mensuels avec études de cas, et un diplôme vérifiable de l'Académie Humanitaire. Programme indépendant, sans partenariat avec Anthropic ni avec Google.

**https://humanitairejobs.org/fr/academie/mini-mba-ia-claude-gemini/semaine-4/**

Document gratuit de l'Académie Humanitaire. Il peut être imprimé et partagé librement à des fins de formation, sans modification.

# Semaine 4 : Éthique de l'IA et principes humanitaires : ne pas nuire, biais, transparence

**Objectifs**

- Relier l'usage de l'IA aux principes humanitaires (humanité, neutralité, impartialité, indépendance) et au principe de ne pas nuire.
- Repérer les biais d'une réponse d'IA et leurs effets possibles sur les personnes affectées.
- Appliquer des règles claires de transparence, de responsabilité humaine et de respect du droit d'auteur.

## Les principes humanitaires face à l'IA

Les quatre principes humanitaires s'appliquent aussi aux outils numériques :

- **Humanité** : l'outil doit servir les personnes affectées, pas seulement l'efficacité interne. Gagner du temps sur un rapport est utile si ce temps revient aux programmes.
- **Impartialité** : une IA ne doit pas introduire de discrimination dans l'accès à l'aide (par exemple en classant des ménages selon des critères biaisés).
- **Neutralité** : un texte généré peut reprendre, sans qu'on le voie, le vocabulaire d'une partie au conflit. Relisez les noms de lieux, de groupes et d'événements.
- **Indépendance** : dépendre entièrement d'un fournisseur pour une fonction critique est un risque ; gardez la capacité de travailler sans l'outil.

Le principe de **ne pas nuire** (*do no harm*) impose de se demander avant chaque nouvel usage : qui pourrait être lésé si l'outil se trompe, si les données fuient, ou si le résultat est mal compris ?

## Les biais : d'où ils viennent, comment les repérer

Les modèles apprennent sur des textes majoritairement produits dans quelques langues et quelques régions du monde. Il en résulte des **biais** :

- **Biais de langue** : qualité moindre en langues peu présentes sur Internet (de nombreuses langues africaines, asiatiques ou autochtones), et parfois dans certaines variantes de l'arabe ou du français.
- **Biais de perspective** : lecture « vue du Nord » des crises, solutions standardisées, méconnaissance des acteurs locaux et des ONG nationales.
- **Stéréotypes** : représentations genrées ou culturelles (le « médecin » est un homme, la « bénéficiaire » est passive) dans les textes et surtout dans les images générées.
- **Biais de complaisance** : tendance à aller dans le sens de la personne qui pose la question. Si vous demandez « confirme que notre projet est un succès », l'assistant risque de confirmer.

Pour les repérer : demandez la position inverse (« quels arguments contre cette conclusion ? »), faites relire les textes destinés aux communautés par des personnes de ces communautés, et comparez avec des sources locales.

## Décisions sur les personnes : la limite à ne pas franchir

Les décisions qui touchent directement les droits ou l'accès à l'aide d'une personne ne doivent pas être confiées à un assistant d'IA générative :

- sélection ou exclusion de bénéficiaires ;
- évaluation de la crédibilité d'un récit (demande d'asile, plainte, cas de protection) ;
- présélection ou rejet de candidats à un emploi ;
- décision disciplinaire ou d'enquête.

Ces outils peuvent aider à **préparer** (rédiger une grille d'entretien, structurer des critères publiés, résumer une politique), mais la décision revient à des personnes habilitées, selon une procédure documentée, avec possibilité de recours. Dans l'Union européenne, le règlement sur l'intelligence artificielle (AI Act, adopté en 2024 et appliqué progressivement) classe d'ailleurs comme « à haut risque » certains usages comme le recrutement ou l'accès à des services essentiels, avec des obligations renforcées.

> Test : si la personne concernée apprenait que la décision a été prise par une IA, serait-ce acceptable et pourrait-elle la contester ? Si la réponse est non, n'utilisez pas l'IA pour décider.

## Transparence et responsabilité

La confiance est le capital d'une organisation humanitaire. Quelques règles simples la protègent :

- **Responsabilité humaine** : chaque document a un auteur humain responsable de son contenu, qu'il ait été aidé ou non par l'IA.
- **Transparence proportionnée** : indiquez l'usage de l'IA quand il est significatif ou quand le destinataire peut raisonnablement s'y attendre (évaluation, recherche, contenu publié, échanges avec les communautés). Pour une correction orthographique, ce n'est pas nécessaire.
- **Respect des règles des bailleurs et des partenaires** : certains bailleurs et appels à propositions encadrent ou demandent de déclarer l'usage de l'IA dans les candidatures ; lisez les conditions (semaine 6).
- **Pas d'usurpation** : ne générez jamais une image, une voix ou une citation faussement attribuée à une personne réelle, et ne présentez jamais une image générée comme une photo de terrain.

Les lignes directrices éthiques de communication du secteur (dignité, consentement éclairé, pas de misérabilisme) s'appliquent pleinement aux contenus produits avec l'IA.

## Droit d'auteur, plagiat et intégrité

Trois points d'attention :

- **Contenus d'autrui** : demander à l'IA de « réécrire » le rapport d'une autre organisation pour le présenter comme le vôtre reste du plagiat. Citez vos sources.
- **Contenus générés** : le statut juridique des textes et images générés varie selon les pays et évolue. Les conditions d'utilisation de Claude et de Gemini précisent les droits de l'utilisateur sur les résultats ; pour un usage commercial ou une publication importante, consultez-les.
- **Intégrité académique et professionnelle** : pour les examens de ce Mini MBA, comme pour une évaluation de recrutement, utiliser l'IA pour répondre à votre place est une fraude. L'objectif est que vous maîtrisiez les compétences, pas l'outil seul.

Pour les photos, la règle de l'organisation reste : pas de photo d'autrui sans droits, pas de photo d'enfant identifiable sans consentement, et une image générée est clairement signalée comme telle.

## Coût environnemental et équité d'accès

L'IA consomme de l'électricité et de l'eau dans les centres de données. Les estimations par requête varient beaucoup selon les études et les modèles ; l'impact global du secteur est en forte croissance. Pour une organisation engagée sur le climat, des réflexes raisonnables :

- utiliser l'IA quand elle apporte une vraie valeur, pas par réflexe pour des tâches triviales ;
- préférer un document bien préparé à des dizaines d'essais successifs ;
- intégrer l'usage du numérique dans la politique environnementale.

**Équité d'accès** : les équipes de terrain ont souvent une connexion faible, des appareils anciens et moins d'accès aux abonnements payants. Une organisation qui réserve l'IA au siège et aux personnels internationaux creuse les inégalités. Une politique juste prévoit l'accès, la formation et les langues pour le personnel national et les partenaires locaux.

## Exercice d'application : analyser trois usages au regard de l'éthique

**Situation.** Une ONG envisage trois usages : A) générer des images « réalistes » d'enfants malnutris pour une campagne de collecte, faute de photos autorisées ; B) demander à l'IA de classer 600 candidatures à un poste de chauffeur et de rejeter automatiquement les 500 moins bien notées ; C) utiliser l'IA pour proposer une première version des messages de sensibilisation au choléra en français facile, ensuite testés auprès de groupes de la communauté et traduits par des locuteurs natifs.

**Travail demandé.** Pour chaque usage, dites s'il est acceptable et pourquoi, et proposez une alternative si nécessaire.

**Corrigé.** A) Non acceptable : une image générée présentée comme réelle trompe le public, porte atteinte à la dignité et renforce le misérabilisme ; alternative : photos avec consentement éclairé, illustrations clairement signalées, ou messages centrés sur les solutions. B) Non acceptable : décision sur des personnes sans contrôle humain, risque de biais et de discrimination, absence de recours ; alternative : critères publiés appliqués par un panel, l'IA pouvant aider à rédiger la grille ou un test pratique. C) Acceptable : l'IA prépare un premier jet, la communauté et des locuteurs natifs valident ; c'est un usage proportionné, transparent et vérifié.

## À retenir

- Les principes humanitaires et le principe de ne pas nuire s'appliquent à l'usage de l'IA.
- Les biais touchent la langue, la perspective, les stéréotypes et la complaisance ; cherchez la contradiction.
- Aucune décision sur les droits ou l'accès à l'aide d'une personne n'est confiée à une IA générative.
- Un auteur humain reste responsable ; signalez l'usage significatif de l'IA ; jamais d'image générée présentée comme réelle.
- Pensez au coût environnemental et à l'équité d'accès pour le personnel national et les partenaires.

## Quiz

1. Demander à l'IA « confirme que notre projet est un succès » expose surtout au biais :
   A. De langue
   B. De complaisance
   C. De date
   D. De format

2. Laquelle de ces tâches ne doit PAS être confiée à une IA générative ?
   A. Rédiger une grille d'entretien
   B. Résumer une politique RH publiée
   C. Décider quels ménages sont exclus de l'aide
   D. Corriger l'orthographe d'un rapport

3. Une image générée d'une situation de terrain peut être utilisée :
   A. Comme photo réelle si elle est réaliste
   B. Jamais, en aucun cas
   C. Seulement pour les bailleurs
   D. Si elle est clairement signalée comme générée et respecte la dignité

4. Le principe d'indépendance invite à :
   A. Garder la capacité de travailler sans l'outil pour les fonctions critiques
   B. Utiliser un seul fournisseur
   C. Ne jamais utiliser l'IA
   D. Laisser l'IA décider

5. Répondre aux examens de ce Mini MBA avec l'IA à sa place est :
   A. Recommandé
   B. Une fraude
   C. Obligatoire
   D. Sans importance

**Corrigé du quiz :** 1. B · 2. C · 3. D · 4. A · 5. B