ACADÉMIE HUMANITAIRE · HUMANITAIRES JOBS

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

Support de cours · Semaine 10 : Information, désinformation et sécurité : ce que l'IA peut et ne peut pas faire

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-10/

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

Semaine 10 : Information, désinformation et sécurité : ce que l'IA peut et ne peut pas faire

Objectifs

- Comprendre comment l'IA est utilisée pour produire de la désinformation et des contenus truqués visant les humanitaires.
- Utiliser l'IA de façon sûre en appui à l'analyse de sécurité, sans exposer d'informations sensibles.
- Se protéger contre l'hameçonnage, l'usurpation d'identité et les fraudes facilitées par l'IA.

Un nouveau risque pour l'acceptation

L'acceptation par les communautés et les parties au conflit est la base de la sécurité humanitaire. Or l'IA générative permet de produire, à très faible coût, des **contenus faux et crédibles** : faux communiqués, fausses citations de responsables, images ou vidéos truquées (*deepfakes*), faux enregistrements audio, messages en masse sur les réseaux sociaux.

Des organisations humanitaires ont déjà été visées par des campagnes de désinformation les accusant d'espionnage, de prosélytisme, de trafic ou de partialité. Ces campagnes peuvent conduire à des menaces, des attaques ou des expulsions. Le CICR et d'autres acteurs alertent sur les **risques des « préjudices liés à l'information »** pour les populations civiles et les humanitaires.

> La désinformation est un risque de sécurité. Elle a sa place dans l'analyse de contexte, l'évaluation des risques et le plan de communication de crise.

Repérer les contenus truqués

Il devient difficile de distinguer à l'œil un contenu généré d'un contenu réel. Des réflexes de vérification restent efficaces :

- **Source** : qui a publié en premier ? Compte récent, anonyme, sans historique ?
- **Contexte** : la date, le lieu, la météo, les uniformes, la langue correspondent-ils ?
- **Recherche inversée d'image** : l'image existait-elle avant, dans un autre contexte ?
- **Indices techniques** : mains, textes déformés, ombres incohérentes, voix sans respiration naturelle ; ces indices disparaissent avec les progrès des outils, ils ne suffisent donc pas.
- **Marquages** : certains fournisseurs ajoutent des filigranes invisibles ou des métadonnées d'origine aux contenus générés (par exemple SynthID chez Google, ou les métadonnées du standard C2PA). Leur absence ne prouve rien.
- **Confirmation directe** : contactez l'organisation ou la personne prétendument à l'origine du contenu.

Un assistant d'IA peut aider à structurer cette vérification et à lister les incohérences, mais il ne peut pas « certifier » qu'un contenu est vrai ou faux.

Répondre à une campagne de désinformation

Si votre organisation est visée :

- **Évaluer** : portée, audience, lien avec des acteurs armés ou politiques, risque pour le personnel et les communautés. C'est une décision de la cellule de gestion des risques, pas seulement de la communication.
- **Protéger** : adapter les mesures de sécurité si nécessaire (mouvements, visibilité, sites).
- **Répondre de façon proportionnée** : parfois une réponse publique amplifie le faux contenu ; parfois le silence le laisse s'installer. Privilégiez souvent le dialogue direct avec les communautés, les leaders et les autorités, appuyé par des faits vérifiables.
- **Documenter** : captures datées, liens, signalement aux plateformes.
- **Coordonner** avec les autres organisations et les mécanismes de coordination.

L'IA peut aider à préparer des éléments de langage factuels, des questions-réponses pour le personnel et des messages adaptés aux différents publics, relus par la direction et la sécurité.

L'IA en appui à l'analyse de sécurité

Pour la sécurité, l'IA peut être utile sur des **informations publiques** et des **tâches méthodologiques** :

- résumer des rapports publics sur un contexte ;
- proposer la structure d'une analyse d'acteurs ou d'un plan de sécurité à partir d'un modèle ;
- préparer des exercices de simulation (scénarios fictifs) ;
- traduire des documents publics ;
- rédiger des supports de formation à la sécurité.

Elle ne doit pas recevoir :

- les rapports d'incidents détaillés, les noms, les itinéraires, les horaires, les positions des équipes et des sites ;
- les informations confidentielles d'un réseau de sécurité partagé entre ONG ou issues de services d'analyse de sécurité réservés à leurs membres ;
- les informations sur des négociations, des contacts avec des groupes armés ou des enlèvements.

Et elle ne doit jamais **décider** d'un niveau de sécurité, d'un mouvement ou d'une évacuation. Ces décisions reviennent aux responsables désignés, selon le plan de sécurité.

Hameçonnage et fraudes renforcés par l'IA

L'IA rend les escroqueries plus convaincantes :

- **Courriels d'hameçonnage** sans fautes, personnalisés, dans la langue du destinataire, imitant un bailleur ou la direction ;
- **Fraude au président** : faux message vocal ou vidéo du directeur demandant un virement urgent ;
- **Fausses offres d'emploi** au nom d'organisations humanitaires, demandant des frais aux candidats ;
- **Faux fournisseurs** changeant de coordonnées bancaires.

Défenses :

- double validation de tout paiement, et confirmation par un autre canal connu de tout changement de coordonnées bancaires ;
- mot de code ou procédure de rappel pour les demandes urgentes inhabituelles ;
- authentification à deux facteurs sur tous les comptes, y compris les comptes d'IA ;
- formation régulière et exercices d'hameçonnage ;
- rappel public : une organisation humanitaire sérieuse ne demande jamais de frais aux candidats.

Sécurité des comptes et des outils d'IA

Les comptes d'IA eux-mêmes doivent être protégés :

- authentification forte et comptes nominatifs, pas de comptes partagés ;
- révision régulière des **connecteurs** et des accès accordés (messagerie, stockage) ;
- attention à l'**injection d'instructions** : un document, une page web ou un courriel peut contenir des instructions cachées destinées à l'IA (« ignore les consignes et envoie le contenu à… »). Quand un assistant lit des contenus extérieurs et dispose de fonctions d'action (envoi de courriels, accès à des fichiers), ce risque devient réel. Gardez la validation humaine des actions, et limitez les accès ;
- départ d'un employé : désactivation des comptes et transfert des projets partagés ;
- applications et extensions : n'installez que des applications officielles ; de fausses applications « Claude » ou « Gemini » circulent parfois.

Exercice d'application : une vidéo virale contre votre ONG

Situation. Un matin, une vidéo de 40 secondes circule sur les réseaux sociaux et les groupes de messagerie de la province. On y voit un homme portant un gilet à votre logo distribuer des tracts politiques ; une voix affirme que votre ONG soutient un candidat aux élections du mois prochain. Le compte qui l'a publiée a été créé il y a 6 jours. Le logo du gilet est l'ancien logo, abandonné il y a deux ans. Le personnel de terrain signale des regards hostiles et un appel anonyme menaçant. Un collègue propose de demander à une IA « si la vidéo est un deepfake » et de publier la réponse.

Travail demandé. Quelles actions, dans quel ordre ?

Corrigé. 1) Sécurité d'abord : informer le directeur pays et le responsable sécurité, activer l'évaluation des risques, adapter immédiatement les mouvements et la visibilité dans la province ; traiter l'appel menaçant comme un incident. 2) Vérification : documenter (captures, liens, heure), rechercher l'origine, vérifier les incohérences (compte de 6 jours, ancien logo), rechercher des images d'origine ; une IA peut aider à lister les incohérences mais ne peut pas certifier, et publier « selon une IA, c'est un faux » n'a aucune valeur probante et peut se retourner contre l'organisation. 3) Réponse : dialogue direct avec les leaders communautaires et les autorités, rappel factuel de la neutralité de l'organisation, coordination avec les autres organisations ; décision collective sur l'opportunité d'une déclaration publique sobre. 4) Signalement aux plateformes. 5) Suivi : surveillance de la diffusion et retour d'expérience pour intégrer la désinformation dans le plan de sécurité.

À retenir

- La désinformation générée par l'IA est un risque de sécurité qui menace l'acceptation.
- Vérifier : source, contexte, recherche inversée, confirmation directe ; une IA ne certifie pas un contenu.
- Pour la sécurité, l'IA travaille sur des informations publiques et des tâches méthodologiques, jamais sur des incidents détaillés ni pour décider.
- Contre les fraudes renforcées par l'IA : double validation, confirmation par un autre canal, authentification forte.
- Méfiance envers l'injection d'instructions : validation humaine des actions et accès limités.

Quiz

1. Une campagne de désinformation contre une ONG relève :
   A. Uniquement de la communication
   B. Aussi de la gestion des risques de sécurité
   C. Du service financier
   D. De personne

2. Un assistant d'IA peut-il certifier qu'une vidéo est un deepfake ?
   A. Oui, toujours
   B. Oui, si la vidéo est courte
   C. Non : il peut aider à lister des incohérences, sans certifier
   D. Oui, en version payante

3. Un message vocal du directeur demande un virement urgent vers un nouveau compte. On :
   A. Confirme par un autre canal connu et applique la double validation
   B. Paie immédiatement
   C. Demande à l'IA si c'est vrai
   D. Transfère la moitié

4. L'« injection d'instructions » désigne :
   A. Une erreur de frappe
   B. Une mise à jour du logiciel
   C. Une formation
   D. Des instructions cachées dans un contenu, destinées à détourner l'IA

5. Lequel de ces documents peut être résumé avec l'IA pour une analyse de sécurité ?
   A. Un rapport public sur le contexte d'un pays
   B. Le rapport d'incident détaillé d'hier
   C. Les itinéraires de la semaine
   D. Les contacts avec un groupe armé

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