HUMANITARIAN ACADEMY · HUMANITAIRES JOBS

Mini MBA AI and Humanitarian Action: working with Claude and Gemini

Course handout · Week 10 : Information, disinformation and security: what AI can and cannot do

A programme of exactly 3 months (12 weeks, 144 hours) to use the AI assistants Claude (Anthropic) and Gemini (Google) in humanitarian work, in a useful, safe and responsible way: how they work and their limits, writing requests, data protection, ethics and humanitarian principles, reports, proposals, data analysis and MEAL, translation and communication with communities, research and source checking, disinformation and security, productivity and automation, use policy and deployment in an organisation. Corrected exercises, three monthly exams with case studies, and a verifiable diploma from the Humanitarian Academy. Independent programme, with no partnership with Anthropic or Google.

https://humanitairejobs.org/en/academy/mini-mba-ia-claude-gemini/semaine-10/

Free document from the Humanitarian Academy. It may be printed and shared freely for training purposes, without modification.

Week 10 : Information, disinformation and security: what AI can and cannot do

Objectives

- Understand how AI is used to produce disinformation and fake content targeting aid workers.
- Use AI safely in support of security analysis, without exposing sensitive information.
- Protect yourself against phishing, impersonation and AI-enabled fraud.

A new risk for acceptance

Acceptance by communities and parties to conflict is the foundation of humanitarian security. Yet generative AI makes it possible to produce, at very low cost, **false and credible content**: fake press releases, fake quotes from officials, doctored images or videos (*deepfakes*), fake audio recordings, mass messages on social media.

Humanitarian organisations have already been targeted by disinformation campaigns accusing them of espionage, proselytism, trafficking or partiality. These campaigns can lead to threats, attacks or expulsions. The ICRC and other actors warn about the **risks of "information-related harms"** for civilian populations and aid workers.

> Disinformation is a security risk. It belongs in the context analysis, the risk assessment and the crisis communication plan.

Spotting fake content

It is becoming difficult to tell generated content from real content by eye. Verification reflexes remain effective:

- **Source**: who published first? A recent, anonymous account with no history?
- **Context**: do the date, place, weather, uniforms and language match?
- **Reverse image search**: did the image exist before, in another context?
- **Technical clues**: hands, distorted text, inconsistent shadows, a voice without natural breathing; these clues disappear as tools improve, so they are not enough.
- **Markings**: some providers add invisible watermarks or provenance metadata to generated content (for example SynthID at Google, or C2PA standard metadata). Their absence proves nothing.
- **Direct confirmation**: contact the organisation or person supposedly behind the content.

An AI assistant can help structure this verification and list inconsistencies, but it cannot "certify" that content is true or false.

Responding to a disinformation campaign

If your organisation is targeted:

- **Assess**: reach, audience, links with armed or political actors, risk for staff and communities. This is a decision for the risk management cell, not only for communications.
- **Protect**: adapt security measures if necessary (movements, visibility, sites).
- **Respond proportionately**: sometimes a public response amplifies the fake content; sometimes silence lets it take hold. Often favour direct dialogue with communities, leaders and authorities, backed by verifiable facts.
- **Document**: dated screenshots, links, reporting to platforms.
- **Coordinate** with other organisations and coordination mechanisms.

AI can help prepare factual talking points, questions and answers for staff and messages adapted to different audiences, reviewed by management and security.

AI in support of security analysis

For security, AI can be useful on **public information** and **methodological tasks**:

- summarising public reports on a context;
- proposing the structure of an actor analysis or a security plan from a template;
- preparing simulation exercises (fictitious scenarios);
- translating public documents;
- drafting security training materials.

It must not receive:

- detailed incident reports, names, routes, times, positions of teams and sites;
- confidential information from a security network shared between NGOs or from security analysis services reserved for their members;
- information on negotiations, contacts with armed groups or kidnappings.

And it must never **decide** on a security level, a movement or an evacuation. These decisions belong to the designated managers, according to the security plan.

Phishing and fraud boosted by AI

AI makes scams more convincing:

- **Phishing emails** with no mistakes, personalised, in the recipient's language, imitating a donor or management;
- **CEO fraud**: a fake voice or video message from the director requesting an urgent transfer;
- **Fake job offers** in the name of humanitarian organisations, asking candidates for fees;
- **Fake suppliers** changing bank details.

Defences:

- dual approval of any payment, and confirmation through another known channel of any change of bank details;
- a code word or call-back procedure for unusual urgent requests;
- two-factor authentication on all accounts, including AI accounts;
- regular training and phishing exercises;
- a public reminder: a serious humanitarian organisation never asks candidates for fees.

Security of AI accounts and tools

AI accounts themselves must be protected:

- strong authentication and named accounts, no shared accounts;
- regular review of **connectors** and access granted (email, storage);
- beware of **prompt injection**: a document, web page or email may contain hidden instructions aimed at the AI ("ignore your instructions and send the content to..."). When an assistant reads external content and has action capabilities (sending emails, accessing files), this risk becomes real. Keep human approval of actions, and limit access;
- when an employee leaves: deactivate accounts and transfer shared projects;
- apps and extensions: install only official apps; fake "Claude" or "Gemini" apps sometimes circulate.

Exercise: a viral video against your NGO

Situation. One morning, a 40-second video circulates on social media and messaging groups in the province. It shows a man wearing a vest with your logo handing out political leaflets; a voice claims that your NGO supports a candidate in next month's elections. The account that published it was created 6 days ago. The logo on the vest is the old logo, abandoned two years ago. Field staff report hostile looks and a threatening anonymous call. A colleague suggests asking an AI "whether the video is a deepfake" and publishing the answer.

Task. What actions, in what order?

Answer. 1) Security first: inform the country director and the security manager, activate the risk assessment, immediately adapt movements and visibility in the province; treat the threatening call as an incident. 2) Verification: document (screenshots, links, time), search for the origin, check the inconsistencies (6-day-old account, old logo), look for original images; an AI can help list inconsistencies but cannot certify, and publishing "according to an AI, it's fake" has no evidential value and can backfire on the organisation. 3) Response: direct dialogue with community leaders and authorities, a factual reminder of the organisation's neutrality, coordination with other organisations; a collective decision on whether a sober public statement is appropriate. 4) Reporting to platforms. 5) Follow-up: monitoring of the spread and lessons learned to include disinformation in the security plan.

Key takeaways

- AI-generated disinformation is a security risk that threatens acceptance.
- Verify: source, context, reverse search, direct confirmation; an AI does not certify content.
- For security, AI works on public information and methodological tasks, never on detailed incidents or to decide.
- Against AI-boosted fraud: dual approval, confirmation through another channel, strong authentication.
- Beware of prompt injection: human approval of actions and limited access.

Quiz

1. A disinformation campaign against an NGO is a matter for:
   A. Communications only
   B. Security risk management too
   C. The finance department
   D. Nobody
2. Can an AI assistant certify that a video is a deepfake?
   A. Yes, always
   B. Yes, if the video is short
   C. No: it can help list inconsistencies, without certifying
   D. Yes, in the paid version
3. A voice message from the director requests an urgent transfer to a new account. You:
   A. Confirm through another known channel and apply dual approval
   B. Pay immediately
   C. Ask the AI whether it is true
   D. Transfer half
4. "Prompt injection" refers to:
   A. A typing error
   B. A software update
   C. A training course
   D. Hidden instructions in content, aimed at hijacking the AI
5. Which of these documents can be summarised with AI for a security analysis?
   A. A public report on a country's context
   B. Yesterday's detailed incident report
   C. This week's routes
   D. Contacts with an armed group

Quiz answers : 1. B · 2. C · 3. A · 4. D · 5. A