HUMANITARIAN ACADEMY · HUMANITAIRES JOBS

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

Course handout · Week 4 : AI ethics and humanitarian principles: do no harm, bias, transparency

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-4/

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

Week 4 : AI ethics and humanitarian principles: do no harm, bias, transparency

Objectives

- Link the use of AI to the humanitarian principles (humanity, neutrality, impartiality, independence) and to the do no harm principle.
- Spot the biases in an AI answer and their possible effects on affected people.
- Apply clear rules on transparency, human responsibility and respect for copyright.

The humanitarian principles and AI

The four humanitarian principles also apply to digital tools:

- **Humanity**: the tool must serve affected people, not only internal efficiency. Saving time on a report is useful if that time goes back to programmes.
- **Impartiality**: AI must not introduce discrimination in access to aid (for example by ranking households according to biased criteria).
- **Neutrality**: generated text may, unnoticed, adopt the vocabulary of one party to a conflict. Review the names of places, groups and events.
- **Independence**: depending entirely on one provider for a critical function is a risk; keep the ability to work without the tool.

The **do no harm** principle requires asking before each new use: who could be harmed if the tool is wrong, if the data leaks, or if the result is misunderstood?

Bias: where it comes from, how to spot it

Models learn from texts mostly produced in a few languages and a few regions of the world. This results in **biases**:

- **Language bias**: lower quality in languages with little presence on the Internet (many African, Asian or Indigenous languages), and sometimes in certain varieties of Arabic or French.
- **Perspective bias**: a "Global North" reading of crises, standardised solutions, little knowledge of local actors and national NGOs.
- **Stereotypes**: gendered or cultural representations (the "doctor" is a man, the "beneficiary" is passive) in texts and especially in generated images.
- **Sycophancy bias**: a tendency to agree with the person asking. If you ask "confirm that our project is a success", the assistant is likely to confirm.

To spot them: ask for the opposite position ("what arguments go against this conclusion?"), have texts for communities reviewed by people from those communities, and compare with local sources.

Decisions about people: the line not to cross

Decisions that directly affect a person's rights or access to aid must not be entrusted to a generative AI assistant:

- selection or exclusion of beneficiaries;
- assessing the credibility of an account (asylum claim, complaint, protection case);
- shortlisting or rejecting job applicants;
- disciplinary or investigation decisions.

These tools can help **prepare** (draft an interview grid, structure published criteria, summarise a policy), but the decision belongs to authorised people, following a documented procedure, with the possibility of appeal. In the European Union, the Artificial Intelligence Act (AI Act, adopted in 2024 and applied gradually) classifies certain uses such as recruitment or access to essential services as "high-risk", with stronger obligations.

> Test: if the person concerned learned that the decision had been made by an AI, would that be acceptable, and could they challenge it? If the answer is no, do not use AI to decide.

Transparency and accountability

Trust is the capital of a humanitarian organisation. A few simple rules protect it:

- **Human responsibility**: every document has a human author responsible for its content, whether or not they were helped by AI.
- **Proportionate transparency**: disclose the use of AI when it is significant or when the recipient can reasonably expect it (evaluation, research, published content, exchanges with communities). For a spelling check, it is not necessary.
- **Respect for donor and partner rules**: some donors and calls for proposals regulate or require disclosure of AI use in applications; read the terms (week 6).
- **No impersonation**: never generate an image, voice or quote falsely attributed to a real person, and never present a generated image as a field photo.

The sector's ethical communication guidelines (dignity, informed consent, no pity-based imagery) apply fully to content produced with AI.

Copyright, plagiarism and integrity

Three points to watch:

- **Other people's content**: asking AI to "rewrite" another organisation's report to present it as yours is still plagiarism. Cite your sources.
- **Generated content**: the legal status of generated texts and images varies between countries and is evolving. The terms of use of Claude and Gemini specify the user's rights over the outputs; for commercial use or a major publication, consult them.

- **Academic and professional integrity**: for the exams of this Mini MBA, as for a recruitment test, using AI to answer in your place is fraud. The aim is for you to master the skills, not the tool alone.

For photos, the organisation's rule remains: no photo of others without rights, no photo of an identifiable child without consent, and a generated image is clearly labelled as such.

Environmental cost and equity of access

AI consumes electricity and water in data centres. Estimates per request vary widely between studies and models; the sector's overall impact is growing fast. For an organisation committed on climate, reasonable habits:

- use AI when it brings real value, not by reflex for trivial tasks;
- prefer one well-prepared document to dozens of successive attempts;
- include digital use in the environmental policy.

Equity of access: field teams often have weak connections, old devices and less access to paid subscriptions. An organisation that reserves AI for headquarters and international staff widens inequalities. A fair policy provides access, training and languages for national staff and local partners.

Exercise: analysing three uses from an ethical standpoint

Situation. An NGO is considering three uses: A) generating "realistic" images of malnourished children for a fundraising campaign, for lack of authorised photos; B) asking AI to rank 600 applications for a driver position and automatically reject the 500 lowest-scoring ones; C) using AI to propose a first version of cholera awareness messages in easy French, then tested with community groups and translated by native speakers.

Task. For each use, say whether it is acceptable and why, and propose an alternative if necessary.

Answer. A) Not acceptable: a generated image presented as real misleads the public, undermines dignity and reinforces pity-based imagery; alternative: photos with informed consent, clearly labelled illustrations, or messages focused on solutions. B) Not acceptable: a decision about people without human control, risk of bias and discrimination, no appeal; alternative: published criteria applied by a panel, with AI possibly helping to draft the grid or a practical test. C) Acceptable: AI prepares a first draft, the community and native speakers validate it; it is a proportionate, transparent and verified use.

Key takeaways

- The humanitarian principles and the do no harm principle apply to the use of AI.
- Biases affect language, perspective, stereotypes and sycophancy; look for contradiction.
- No decision on a person's rights or access to aid is entrusted to a generative AI.
- A human author remains responsible; disclose significant AI use; never present a generated image as real.
- Consider the environmental cost and equity of access for national staff and partners.

Quiz

1. Asking AI "confirm that our project is a success" mainly exposes you to:
   A. Language bias
   B. Sycophancy bias
   C. Date bias
   D. Format bias

2. Which of these tasks must NOT be entrusted to a generative AI?
   A. Drafting an interview grid
   B. Summarising a published HR policy
   C. Deciding which households are excluded from aid
   D. Correcting the spelling of a report

3. A generated image of a field situation can be used:
   A. As a real photo if it is realistic
   B. Never, under any circumstances
   C. Only for donors
   D. If it is clearly labelled as generated and respects dignity

4. The principle of independence encourages you to:
   A. Keep the ability to work without the tool for critical functions
   B. Use a single provider
   C. Never use AI
   D. Let AI decide

5. Answering the exams of this Mini MBA with AI in your place is:
   A. Recommended
   B. Fraud
   C. Mandatory
   D. Unimportant

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