10 Principles for Responsible AI Use in Child Protection in Humanitarian Action

THE ALLIANCE FOR CHILD PROTECTION IN HUMANITARIAN ACTION

These ten principles provide a foundation for the safe, ethical, and responsible use of artificial intelligence (AI) in child protection in humanitarian action. They are intended to guide practitioners and organisations[1] in considering the risks, responsibilities, and safeguards associated with AI use, while keeping children's safety and well-being at the centre.

1 Safety and do no harm

Before any AI use that may affect a child or family, ask whether it could put them at greater risk. If you have real doubt that an output is accurate, appropriate, or safe to act on, do not use it.

2 Data protection

Never enter personal data about individuals (including names, identification number, individual case details, audio-visual materials, family or referral information, and information that could identify someone indirectly) into general-purpose AI tools, including free tools such as free ChatGPT, Gemini, or Google Translate. Give a tool only the data the task needs. Aggregated or anonymised programme data may be used where it has been assessed as appropriate and does not pose identification, confidentiality, protection, or security risks.

3 Human primacy in child protection decisions

AI must support, never replace, professional judgement. Every decision rests on the practitioner's expertise and knowledge of the child, family, and context. AI is one input and never the deciding factor.

4 Accuracy and verification

Check every AI output before using it in practice. Some errors are convincing enough to fool experts.[2] Treat AI summaries, analysis, referral information, legal guidance, risk assessments, and translations as drafts. If you cannot judge an output yourself, such as a translation into a language you do not read, have a qualified person check it.

Explore the full guidance.

5 Fairness and non-discrimination

Do not use AI in ways that deepen bias against marginalised children or communities. LLMs carry systematic biases of gender, race, ethnicity, religion, and nationality, which fall hardest on groups least represented in AI development.[3,4] Check that outputs are linguistically and culturally appropriate before using them.

6 Transparency and accountability

Be able to explain how AI was used in any decision affecting a child or family, and name the professional responsible. Children and families have a right to know, in terms they understand, when AI is used in processes that affect them.

7 Data security

Keep sensitive data confidential, accurate, and accessible only to authorised individuals when needed, especially in conflict settings where exposed data can be used to target children and families. Commercial tools may store data on third-party servers, use it for training, or leak it in a breach. Confirm what protections a tool has and record where it processes data, since local law governs who can access it. Data that looks anonymous can still be re-identified.

8 Defined purpose and proportionality

Use AI only for a clear, legitimate purpose, and never for surveillance, profiling, or anything beyond the original purpose. Consider whether AI is needed, including its environmental and equity costs. Weigh the benefits against the risks. If the time saved does not justify the added risk, do the work without AI.

9 Child-centred and participatory

The best interests of the child come first in every decision about AI. Judge AI by its effect on children and families, and give them a say in decisions that affect them. AI must never replace the direct human relationship in child protection. Seek meaningful, informed consent where appropriate before children or families interact directly with AI tools. Explain clearly how AI will be used, what information will be involved, and any potential risks. Where consent cannot be obtained, including in acute emergencies, follow applicable safeguarding and data protection requirements, use only non-identifying information, document the reasons, and inform those affected when it is safe and appropriate to do so. Information and opportunities for participation should also be adapted to the child's age and evolving capacities, wherever appropriate. For younger children, information and consent processes should involve caregivers or other legal guardians, in accordance with applicable laws and safeguarding requirements.

10 Local and cultural relevance

AI tools must suit the language, culture, and context where they are used, and must never override practitioners' local knowledge. Most systems are trained mainly on English and perform worse in other languages, with the greatest risk in translation (Section 1.5). Before relying on a tool in a language, check what is known about its performance, and treat a language with no evidence as high risk.

1 Principles 1–7 and 10 adapt the trustworthiness characteristics of the NIST AI Risk Management Framework to child protection. Principle 8 comes from the IASC Operational Guidance on Data Responsibility (2023), and Principle 9 from the UN Convention on the Rights of the Child (Articles 3 and 12).
2 C. Rathkopf, "Hallucination, Reliability, and the Role of Generative AI in Science" (2025), arXiv preprint.
3 I. O. Gallegos et al., "Bias and Fairness in Large Language Models: A Survey," Computational Linguistics 50, no. 3 (2024).
4 United Nations High-Level Advisory Body on Artificial Intelligence, Governing AI for Humanity: Final Report (2024).