LiGHT EPFL ICRC

# From Principles to Practice

## AI in Humanitarian Organizations

AI Alignment Guide, 2026

## Acknowledgements

**The EPFL and the ICRC are very grateful to the academic experts and humanitarian colleagues who made this project possible by participating and sharing insights that have been integrated in this report.**

**This work was funded by the Science and Technology for Humanitarian Action Challenges (HAC) programme from the Engineering for Humanitarian Action (EHA) initiative, a partnership between the ICRC, EPFL and ETH Zurich. EHA initiatives are managed jointly by the ICRC, EPFL EssentialTech Centre and ETH Zurich's ETH4D.**

# Tables of Contents

# Foreword

Like many other sectors, the humanitarian sector is confronted with the immense promises, and equally immense questions, brought about by rapid advances in "Artificial Intelligence" (AI). Many organisations are exploring how AI based systems and tools can help them be more efficient in addressing overwhelming needs in a context of reduced humanitarian budgets. From automated data analysis for better programming, to enhanced information management, leaner logistics and engagement with affected populations at scale, AI innovation projects are spreading across the sector.

It can be difficult for humanitarians to manage at the necessary pace the many risks and ethical questions that using AI for humanitarian purposes generates. The challenge is daunting, but humanitarians can create the space, tools, methods and safeguards they need to ensure that leveraging AI to alleviate human suffering does not inadvertently produce harmful or counter-productive consequences. There are many ethical guidelines and resources already available to help organisations mitigate that risk. But it can be difficult to navigate such a wealth of guidance and to meaningfully transpose it into practice – ethical values can be difficult to translate into lines of code.

To contribute to those efforts, the ICRC partnered with the Ecole Polytechnique Fédérale de Lausanne (EPFL) in the context of the Swiss Humanitarian Action Challenge (HAC) to develop a 'Humanitarian-AI Alignment' research project. The partnership brought together humanitarian and technical expertise to develop guidance for practitioners to explore, assess and improve the alignment of AI systems with the humanitarian principles of humanity, impartiality, neutrality and independence. Using the ICRC AI Policy as a starting ground, the project focused on the potential of Large Language Models (LLMs) to support humanitarian organisations' activities. It aimed at positioning the humanitarian principles with methodological approaches to AI to define synergies and tensions in the humanitarian context.

A series of practical workshops helped identify key challenges and opportunities. Practitioners from the humanitarian and technology sectors were brought together to help build a common understanding and language to co-create a common working framework between humanitarian and AI professionals. These conversations generated ideas of practical questions and pathways to embed humanitarian principles in the technical development and use of LLMs for humanitarian activities.

The following pages describe the findings of this project. These include recommendations for the diverse streams of actions related to the development, deployment, and use of AI tools, through a constructivist recognition of the tensions – as trade-offs – related to the application of humanitarian principles in urgent and sensitive contexts.

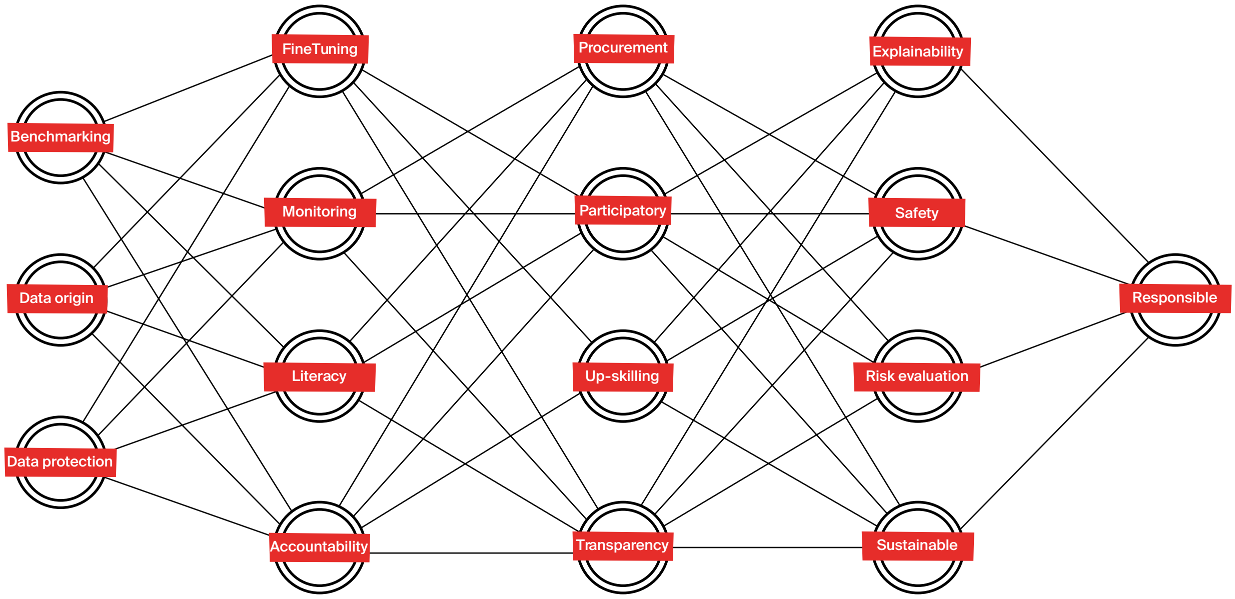
This document is created to guide and inspire humanitarian and AI experts. It collects useful foundations and resources to foster learning and reframing of humanitarian practices. By addressing humanitarian AI challenges through the prism of the humanitarian principles, it acknowledges the inherent difficulties and leverages them to advance AI humanitarian practice through a responsible approach.

## How to use this document

This document focuses on AI systems such as LLMs used to support humanitarian activities and programs. It focuses on AI design and governance questions on which humanitarian organisations can exert some level of control. It does not address broader societal risks related to AI systems.

This guide does not provide one-size-fits all solutions or checklists. It is meant to help generate the foundations of a 'mindset' and identifies how the humanitarian principles can be considered in these different phases to better align the development and use of AI systems with humanitarian ethics. It provides a set of thematically clustered 'step-by-step' recommendations to help structure conversations, evaluation and decision-making, and identifies some of the most relevant existing guidance resources related to each stage of the process. The considerations are not linked to specific technical approaches so that they remain relevant in a fast-paced technology landscape.

While specifically tailored for IT specialists, AI experts, product owners and project managers in humanitarian organizations, this guidance is meant to create a shared reference to support a multidisciplinary framing of AI. The different steps identified in the guidance below all require the systematic involvement of different categories of professionals and experts to support a holistic and cross-organisational understanding and integration of AI systems.

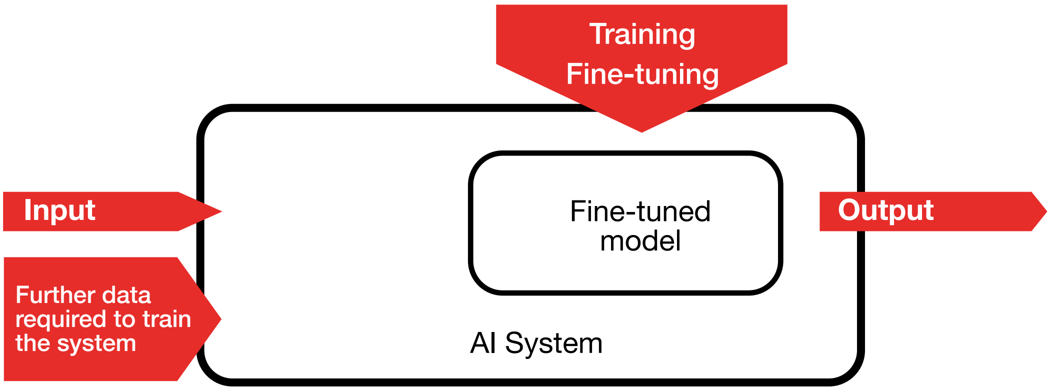


# 1 Definitions and Overview

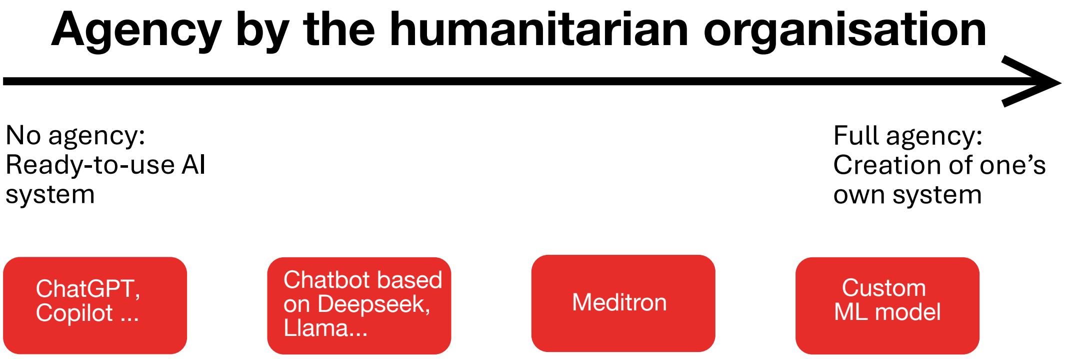
**For the purpose of this document, AI is defined as: “system, tool or technology that involve the use of computer systems to carry out tasks that would ordinarily require human cognition, planning or reasoning„[1]**

# 1 Definitions and Overview

The guidance below focuses on advanced machine learning (ML) based AI systems, including LLMs and multimodal generative AI systems. It distinguishes between AI models (i.e., algorithms trained on large amount of data) and AI systems (i.e., products including end-user interfaces adapted to the deployment context):

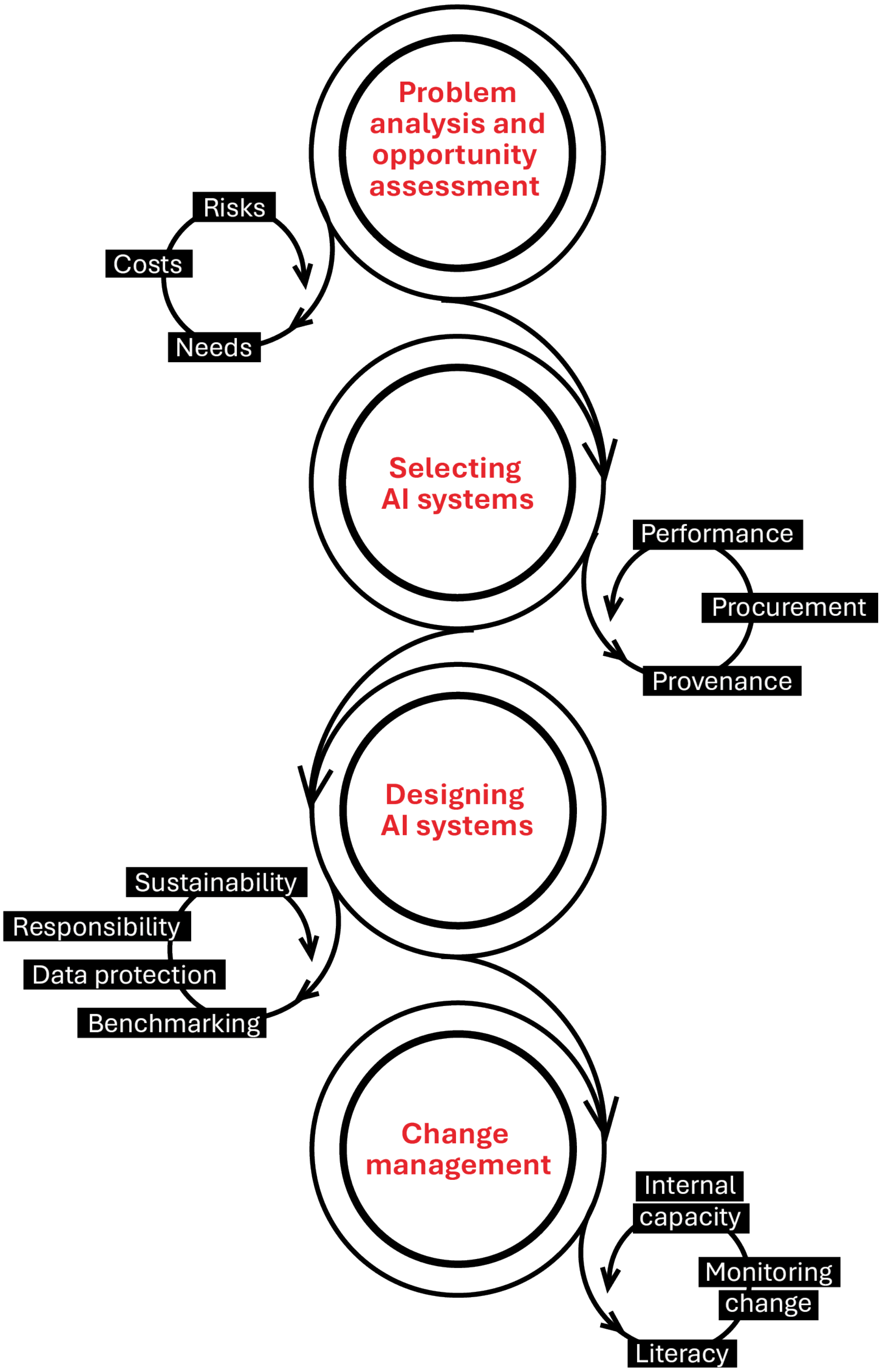


AI systems are categorized by their design along a continuum of different levels of implementation agency, ranging from not adaptable off-the-shelf AI systems (i.e., no agency) to the internal development of AI systems (i.e., full agency).


Agency by the humanitarian organisation
No agency: Ready-to-use AI system
Full agency: Creation of one's own system
ChatGPT, Copilot ...
Chatbot based on Deepseek, Llama...
Meditron
Custom ML model

The thematically clustered step-by-step process presented below identifies the key stages of AI systems development and deployment and highlights how the humanitarian principles can be used to identify the key practical considerations relevant to each stage. This process starts with reflections about the relevance and suitability of AI solutions and covers the life-cycle of AI project management until the deployment and integration of the tools/systems in existing processes and operations.

[1] https://www.icrc.org/en/publication/building-responsible-humanitarian-approach-icrcs-policy-artificial-intelligence, p. 2


Problem analysis and opportunity assessment
Risks
Costs
Needs
Selecting AI systems
Performance
Procurement
Provenance
Designing AI systems
Sustainability
Responsibility
Data protection
Benchmarking
Change management
Internal capacity
Monitoring change
Literacy

# ② The Humanitarian Foundations

**Humanitarian action is guided by the principles of humanity, impartiality, neutrality and independence. They represent the ethical and operational framework that humanitarian organisations should follow in delivering their mandate and activities.**

The humanitarian Principles reflect the values and core working modalities of humanitarian actors. They are also a useful compass and analytical prism to understand and navigate the complexity of humanitarian environments, and to evaluate the risks and limitations of the tools and methodologies used.

The tables below summarizes the key elements of each humanitarian principle and how they relate to the use of AI in the humanitarian context.

These principles are also helpful to help make sense of the challenges, risks and opportunities attached to AI solutions, and how they align, or not, with humanitarian ethics. Embedding them – in a dynamic and contextual fashion – in the methodology, decision-making systems and technical processes related to the development and use of AI systems is a good way to inject ethical safeguards into practice.

## Humanity

Human suffering must be addressed whenever it is found. The purpose of humanitarian action is to protect life and health and ensure respect for human beings.

This is an essential and 'supreme' principle capturing the objective and ultimate goal of humanitarian action. It is the trigger of humanitarian action and a moral imperative.

It is about alleviating suffering (and 'doing no harm' while doing so) and contributing to improve the protection and respect for people's safety, dignity and rights.

**Relevant considerations in the AI domain**

- Consider the cost/benefit analysis and balance the expected result and outcome of using an AI solution in responding to the suffering of population.
- Consider not only the scale and quantitative aspects of the expected output, but also its quality and impact on people and contexts.
- Identify the potential negative impact of using the AI solution on the dignity, autonomy, agency, and rights of the end-users or those affected by the solution.
- Identify the potential negative impact of using digital and automated interfaces instead of humans to interact with users and affected (and often vulnerable) populations. This should include consideration of human contact, empathy, and compassion.

**Examples**

- Evaluate whether a high cost AI tool is the best response to address the needs of a small community of about 100 people.
- Evaluate the cost/benefit of deploying a mental health chatbots to help provide support to an isolated community with no access to mental health support vs. the cost/benefit of contracting a mental health professional with expertise of the local context.
- Evaluate and addressing the difficulty of understanding the functioning of an automated self-registration system for people with low level of digital literacy.
- Assess the pros and cons of replacing a staffed accountability to affected population helpdesk with an automated chatbot.

## Impartiality

Humanitarian action must be carried out on the basis of needs alone, giving priority to the most urgent cases of distress and making no distinctions on the basis of nationality, race, gender, religious belief, class or political opinions.

This is a substantive but also derivative and operational principle guiding 'how, to whom and where' humanitarian organizations should carry out and prioritize their activities and assistance. It is a positive principle that triggers and helps organize action.

This principle requires making an objective assessment based on considerations of non-discrimination, equity and proportionality.

In practice, it means that humanitarian organizations must prioritize responding to people facing the greatest needs and vulnerabilities, without any form of discrimination based on religious beliefs, nationality, race, political opinions or social distinctions. This also means taking into consideration how existing societal or systemic discrimination can impact impartial action.

It also requires humanitarian responders to base their decision only on this objective assessment of needs, without being influenced by personal considerations or feelings.

**Relevant considerations in the AI domain**

- Identify and mitigate the risks of data-based and algorithmic biases and discriminations integrated in AI systems.
- Consider the contextual and individual needs, requirements and vulnerabilities of ends users and affected populations, including in terms of (digital) literacy, language and ability to understand the relevance, functioning, impact and limits of AI solutions – with a focus on avoiding risks of exclusion.
- Consider the need to be able to understand and explain if and how the use of AI solutions impacts your ability to justify programmatic decisions, outputs and outcomes in line with an objective assessment of needs.

**Examples**

- Verify whether a facial recognition system for indigenous population is trained over a representative datase.
- Identify the accessibility limitation of a Help Chatbot not trained on local dialects.
- Evaluate if/how an automated decision-system with proprietary algorithm can be analysed to explain how the data input informs the output

## Neutrality

Humanitarian actors must not take sides in hostilities or engage in controversies of a political, racial, religious or ideological nature.

This is a derivative and operational principle related to the ability of humanitarian organizations to carry out their activities in often polarized and divided environments.

This principle enables access to reach people in need on 'all sides' and to maintain a dialogue with all the authorities and actors who have influence over affected populations and the ability of humanitarian organisations to operate.

It is both a negative (i.e., abstaining from taking sides or making personal judgement) and positive principle (i.e., neutrality requires action to be seen and understood) that is closely linked to the perception of the identity, working modalities and independence of humanitarian organisations.

Neutrality must be conceived in the particular circumstances of each operational context, but also more generally, to avoid perception of double standards.

### Relevant considerations in the AI domain

- Analyse the potential political, socio-cultural or religious values and objectives that may be integrated in the design and functioning of the AI solutions considered.
- Consider the potential political, religious or socio-cultural leanings and actions of the AI solution providers to avoid the risk that using their products or services may project a perception of association or alignment with them or their objectives.

### Examples

- Identify potential race- or gender-based discriminative responses from an off-the shelf chatbot system.
- Evaluate the potential perception issues related to the use of an AI system from a tech company provider that supports a party to the conflict at play in the context where you operate.

## Independence

Humanitarian action must be autonomous from the political, economic, military or other objectives that any actors may hold with regard to areas where humanitarian action is being implemented.

This is a derivative and operational principle related to the ability of humanitarian organizations to carry out their activities in often polarized and divided environments. It can be considered as the 'visible' side of neutrality and is closely related to it.

It is about ensuring that humanitarian organisations' decisions and actions are taken autonomously and independently from potential pressures or agendas of military, political or socio-economic actors.

### Relevant considerations in the AI domain

- Analyse how the use and integration of a considered AI solution can impact the ability of humanitarian organizations to take autonomous decisions and avoid being perceived as associated, or favoring, the interests and objectives of military, political or socio-economic actors (see also neutrality above).
- Think about the long-term costs and risks attached to the use and integration of the considered AI solutions vis-à-vis the operational autonomy and agency of the organization, in particular around 'vendor lock-in' risks.

### Examples

- Evaluate the perception risks attached to use AI systems connected to a social service program provided by a government known to persecute a local minority.
- Identify the risks of being forced to adopt further tools and services attached to the automatic upgrade of a proprietary commercial AI systems.

# ③ Learning Cycles

An iterative approach to learning that continuously adapts to feedback is critical at every step. From the problem analysis and opportunity assessment, proofs-of-context and piloting help refine ideas, approaches, and risk identification.

## 3.1. Continuous adaptation and iterative approach

An iterative approach to learning that continuously adapts to feedback is critical at every step. From the problem analysis and opportunity assessment, proof-of-concepts and piloting help refine ideas, approaches, and risk identification. As the project moves to implementation, methodologies such as Machine Learning Operations (MLOps) play a critical role in ensuring the ongoing assessment and optimization of AI system performance. By embedding continuous monitoring and learning cycles at every stage, organizations can enhance their accountability and the resilience and effectiveness of AI systems. This ensures a good integration of AI solutions in the existing human, technical and financial resources of an organization.

for further information, click here to access more resources

## 3.2. Participatory and Multidisciplinary approach

**#Humanity**

Involving affected people in the choice and design of humanitarian solutions that have an effect on their life, security and rights is critical to support their autonomy and sense of dignity. It is an additional element that can support transparency and accountability, and their informed consent.

**#Impartiality**

Having inclusive methodologies of engagement and participation that take into account existing discriminations and power dynamics within local communities is very important to ensure that participatory approaches are fair and equitable. They require specific design efforts to ensure that the needs of people with specific vulnerabilities are taken into account and addressed.

Creating effective mechanisms to involve  affected individuals and relevant experts at each stage of the AI development and project management cycle is fundamental. It is what helps ensure as much as possible that solutions are relevant and tailored to specific needs (i.e., see problem analysis and opportunity assessment below) and what facilitates efficient change management, transparency and accountability – among other things.

Considering the technical complexity and cross-organizational impact that the development and use of AI tools can have, ensuring the participation of all relevant stakeholders to support informed decision-making at each stage of the process is critical. For example, front lines operators will help contextualise the problem analysis and design phase; data scientist and data protection officers will help strengthen the quality of the datasets required; cybersecurity experts will help integrate information security across the solution's lifecycle; legal advisers will help manage regulatory compliance risks; and communication experts help facilitate internal change management and transparency and accountability with affected communities. A participatory approach also helps addressing possible tensions by identifying trade-offs and possible mitigation strategies.

for further information, click here to access more resources

## 3.3. Transparency and Accountability

**#Humanity**

The ability to explain and be accountable for the use and consequences of AI solutions is essential to respect the dignity of the people humanitarian organizations work with and for. Given the power imbalances at play in the humanitarian context, making concrete and effective efforts to help users/affected people understand what solutions are chosen and why, and what impact they may have on their lives is critical to ensure respect for their autonomy and agency, and to help them make choices that are as informed as possible.

**#Impartiality**

Being able to document the choices and decisions made around the development and use of AI solutions can help humanitarians better explain how they align with the principles of impartiality, and to justify the outcomes of their activities and methods in line with the objective of addressing the most urgent needs while avoiding risks of discriminations.

Transparency and accountability are fundamental to ensuring a responsible humanitarian approach to AI. These are essential requirements that must be embedded and reflected at each stage of the AI lifecycle and project management to facilitate change management, decision-making and the sustainability of the AI solutions in the organization.

Technical transparency and accountability both require a systematic documentation of the analysis and decisions taken at each stage of the AI development and use, with a focus on cost/benefits analysis, risks/opportunities evaluation, trade-offs and ethical considerations. They also require proactive human communication and sharing of the relevant documentation with stakeholders, decision-makers and responsibility holders. In many circumstances, a combination of internal and external auditing mechanisms can significantly support transparency and accountability.

for further information, click here to access more resources

# 4 Problem analysis and opportunity assessment

Because AI is a ‘multi-purpose’ technology, adopting a ‘problem-driven approach’ to assess its relevance in each situation is required.

## 4.1 Needs, relevance and usefulness

### #Humanity

Ensuring that AI solutions are useful and relevant to address the needs of affected people is required to ensure alignment with the principle of humanity.

### #Impartiality

Being able to demonstrate how the use of the AI solution helps better organizing existing resources and having a more impactful outcome on the needs of the most vulnerable people is also necessary to justify operational decisions in line with the principle of impartiality.

### #Independence

A problem-driven approach is helpful to rationalize the decision to use an AI solution, and to mitigate the potential perception that these solutions are driven by external considerations such as donor or partner-driven interests.

The starting point of any AI solution related discussion should be 'what is the problem we face and are trying to solve?' and then 'what is the best way to solve this problem, and if so how (be it an AI system or not), when, where and at what cost?'.

There is often a risk of humanitarian innovation being driven by a solution-based approach: 'now that this technology is here, what can we do with it?'. This can lead to launching costly and energy-consuming projects that end up having little or no real impact on operational efficiency or the lives of the people humanitarians are here to serve. It is easy to fall into the trap: first investing in an AI system with a vague idea of the purpose, and trying to find a problem to be solved with it afterwards. This 'techno-solutionist' approach can also lead to continuously looking for technological solutions to technologically created problems. It is a common mistake across the humanitarian innovation space, but it is not inevitable.

Developing or using AI systems must be shown to contribute to the organization's objectives, and to align with its values and working modalities. Following a "problem-first" approach and careful opportunity assessment can help prevent developing or acquiring tools with no clear purpose or insufficient return-on-investment. The key rationale for this assessment is to repeatedly ask oneself: "is the AI system helping solve the problem(s) we face and, if so, with which consequences?".

for further information, click here to access more resources

## 4.2. Risks evaluation

### #Humanity

Risk evaluation is an essential step to identify and mitigate the risks of potential negative impact of AI solutions on the safety, dignity and resilience of affected people. This helps support the 'do no harm' approach that is required by the principle of humanity.

### #Independence

Identifying the risks that the use and deployment of AI solutions increase external dependencies and reduce the operational autonomy of the organization can help strengthen alignment with the principle of independence.

AI systems can be categorized along a risk-based approach, an angle taken by regulatory instances like the European Union (see the EU AI Act). Identifying the general risk and sensitivity level of the envisioned AI system is an important part of the problem analysis and relevance assessment, as some AI systems may be excluded from use in the humanitarian context due to the nature and impact of the risks entailed, or the unavailability of effective safeguards.

For a humanitarian organization, the risk level will relate to the potential impact on its core missions, objectives, and principles. For instance, an AI solution that can potentially cause harm and impact the life, safety or dignity of populations; or relate to the provision of assistance activities, such as food and water, healthcare or shelter, will be considered high risk. This includes AI solutions for needs assessments, programmatic planning, or operational delivery, as their use could have significant consequences on affected populations. On the other end of the spectrum, AI solutions related to purely internal administrative processes, such as automated archiving of documents, will typically be considered low risk as their impact on the organization or the people it serves will be indirect.

The risks associated with an AI system may extend beyond its core functionality, such as 'vendor lock-in' when using a third-party solution that may not be maintained or updated. These longer-term risks can emerge later in the project life cycle and persist after the AI system's retirement, for example around the absence of alternatives beyond an AI system that reaches end-of-life. Beyond identification and assessment, specific measures can be envisaged at this stage to mitigate the relevant risks throughout the lifecycle of the AI system.

for further information, click here to access more resources

## 4.3. Costs assessment

**#Independence**

A holistic and long-term cost assessment will support a more sustainable management of AI solutions that mitigate and reduce the financial and organizational dependencies that can affect the organization's capacity to maintain an independent approach.

The prevailing marketing narratives around AI are often anchored in the promise of increased efficiencies and cost-savings on organizational processes and human resources. The pervasiveness of this marketing discourse can create a confirmation bias and a short-term approach to costs management.

While cost management is an important requirement for responsible humanitarian action, it is important to ensure that cost-efficiency deliberations do not override humanitarian considerations of humanity and impartiality. Additionally, practice and experiences from the private and public sector illustrate that cost-efficiency is not a given with AI solutions, which often require significant investments to ensure safety, quality and sustainability in each organisational or operational context.

At this stage, it is therefore important to adopt a long-term and holistic approach that prioritises operational impact and integrates a quantification of all the direct and indirect costs and investments required by the envisioned AI system. Beyond direct IT development, acquisition, licensing and procurement fees, the cost assessment must consider the significant human, financial and technical resources required to ensure maintenance and updates, cybersecurity and data protection measures, quality control, organisational integration and staff training, as well as energy and environmental costs.

for further information, click here to access more resources

# 5 Selecting and designing AI systems

Humanitarian organizations often do not have the resources to fully develop their own AI systems in-house. They regularly acquire AI systems developed by commercial companies or open source models that have already been trained.

## 5.1. Assessing provenance of external assets

Whether commercially acquired or open source, AI models limit the organization's level of agency at the design and development level because of the complexity and opacity of supporting datasets and training procedures. While open-source models may offer some advantages in that respect, they also require significant resources and competences to be analysed, assessed, finetuned, and implemented.

Similarly to risk assessments, the scope of the due diligence will vary based on the positioning, purpose and reach of the AI system (e.g., backbone infrastructure, front-office application). Additionally, the complexity of a model, even if open source, might prevent meaningful quality checks on the training data and algorithms. The opacity of the policies and practices of the companies providing the solutions may also be an obstacle to a precise and reliable evaluation of other relevant aspects (e.g., corporate social responsibility, data protection or environmental costs).

## 5.2. Procurement: risks and due diligence

**#Humanity**

Ensuring that the AI solution providers are not involved in harmful corporate practices that have a negative impact on local communities (e.g., labour practices, environmental impact) or societies (e.g., legal compliance or political activism) can help preventing that the organisation is seen as supporting or contributing to harmful products and practices (i.e. 'do no harm' requirement).

**#Neutrality #Independence**

Assessing the corporate practices, policies and governance of the companies providing the AI solutions is helpful to identify potential reputational risks that could affect the perception of the neutrality of the organization. Partnering with companies that provide AI systems to military actors, or that are involved in political activities, is particularly risky and should be avoided.

The evaluation of an AI system's provenance must be holistic and dynamic, as different relevant dimensions can change overtime depending on the nature and technical evolutions of the tools, or the practices and policies of their providers (i.e., commercial models) or sources (i.e., open-source models).

The corporate activities, policies and practices of the companies providing the considered solutions should be investigated to identify the human (e.g., labour practices and respect for international standards), societal (e.g., regulatory, intellectual property/copyrights and data protection compliance) and environmental (e.g., carbon footprint and mitigation of environmental impact) costs involved in their value chain. While it can be challenging to assess these different dimensions along the AI value and supply chains, many providers make relevant information available to their customers and shareholders. The absence of relevant information is a negative indicator, and in that case, efforts should be made to obtain such information from the provider or third parties (i.e., media or civil society organisations).

This evaluation is essential to ensure that organisations have a good understanding of the potential perception and reputational risks involved in those commercial or partnership relationships. A 'scoring' system can be helpful to weight and compare the different human, societal and environmental costs behind different AI solutions and decide if, to what extent and how humanitarian organisations can partner with or use the products of their providers.

Humanitarian organizations should be particularly cautious when acquiring the products and services or partnering with companies that are involved in the military, defense and security sectors, or that have previously been identified for harmful practices (e.g., violation of data protection regulation, involvement with dubious intermediaries and business partners, etc.). Using their products or services may have a negative impact on the perception of the humanitarian identity, neutrality and independence of the organization. These perceptions issues can in turn have a concrete impact on trust, access and security, especially in conflict settings where these companies are involved in supporting one side against another. This risk is significant and must be carefully weighed in the due diligence process. Whenever possible, alternative providers should be preferred.

for further information, click here to access more resources

## 5.3. Design for performance

In the continuation of the opportunity assessment conducted previously, the general goal is to make a useful tool that has a positive impact for affected people, users or the organisation (and ideally for all). This means that the design and development of the AI solution must focus on responding to the problem identified through a participatory approach. Those affected must be able to bring their experience, and contribute to the definition of  the system's parameters and requirements to ensure adaptation to the specificities of their needs and to the local environment in which the solution is meant to be deployed and used.

### 5.3.1. Data quality

**#Humanity #Impartiality**

Ensuring that data and algorithmic biases do not replicate or amplify existing systemic or societal discriminations is critical to ensure that all people in need are taken into account in the design and deployment of the AI solutions. This enables respecting the dignity inherent to every human being, in line with the principles of humanity, and to ensure that those most vulnerable are effectively integrating in the assessment and response to needs, in line with the principle of impartiality.

**#Independence #Neutrality**

Some data gaps and biases included in datasets are resulting from the political opinions of preferences of the companies or people behind them (e.g., voluntary exclusion of gender or ethnic minorities from a population sample). Relying on such datasets for humanitarian related activities may be perceived as being aligned with those opinions, and therefore jeopardize the perception of the neutrality or independence of the organization.

The quality of the output of a system is strongly related to the quality of the data that is fed into the system and on which it is trained (i.e., 'garbage in, garbage out'). Incomplete, obsolete or biased data can significantly limit the performance of a system and jeopardize the relevance and quality of its outputs. Strong data quality, protection and governance mechanisms and safeguards are critical to the successful deployment and use of AI systems.

The quality of AI systems is directly dependent on the implementation and compliance with data management best practices throughout their lifecycle. Strong data governance practices are critical to the performance and sustainability of AI systems, and in return they must adapt to the new data created by AI systems.

When developing their own AI systems, humanitarian organizations should pay specific attention to ensure that supporting datasets are representative of the situation where the systems will be used. This is particularly important as existing datasets may not reflect the realities of the contexts in which the AI system will be deployed. This requires understanding the power dynamics and discriminations at play in the local environment, to identify the potential blind spots or biases of the datasets and ensure that categories of population that may be excluded from them are adequately included and represented.

The participation of users and the support of a network of local experts and partners can be particularly helpful in carrying out this exercise and identify solutions to assess representativeness from a local perspective. These assessments may not always be possible when using pretrained models or off-the-shelf due to the opacity around training practices, in which case specific attention should be put on benchmarking performance.

for further information, click here to access more resources

## 5.4. Benchmarking performance

**#Impartiality**

Benchmarking performance is a useful way to support the ability to justify the choice and use of a particular solution. It can help provide transparency and explainability, in particular regarding potential data and algorithmic biases that are relevant for the principle of impartiality.

Using benchmarks is necessary to assess the performance and relevance of AI systems. It is important that the benchmark used adequately reflects real-world conditions (e.g., available resources, knowledge, competences, sensitivity) of the tasks for which the AI system will be employed – as many may rely on decontextualised versions. This is particularly relevant when the quality of supporting datasets cannot be evaluated, as discussed above. Extensive testing in extraordinary conditions of use before, during and after deployment is required when AI solutions are likely to be deployed in emergency situations and highly volatile humanitarian settings.

It is essential to understand that, while necessary, the benchmarking exercise is not a 'silver bullet' that can address all the performance relates issues of an AI solutions. Benchmarking will often remain theoretical and not necessarily reflect real-use operational contexts and circumstances. Using public benchmarks (external or internals) may lose value over time as developers will be able to specifically train and customize models to perform well on the benchmark. If internal and non-public benchmarks are developed and used, organizations must integrate accountability requirements and ensure their capacity to publicly justify technology-related decisions and choices.

for further information, click here to access more resources

## 5.5. Responsibility by design

Humanitarian organizations and staff must always be able to justify their choice and use of an AI system and be responsible for their impact. This is especially important when using AI systems which have an impact on the delivery of assistance to affected persons and on the rights of the users – whether they are proprietary off-the-shelf or customized models.

As discussed above, transparency and accountability measures require continuous and proactive investments within the organisation. But technological design solutions can also help embed transparency, explainability and responsibility in the core system of the tools considered. Responsibility by design integrated through appropriate levels of controls can help the user better understand the functioning of the AI system, assess its weaknesses and limitations, and critically consider its output and impact. Integrating monitoring and takedown protocols with AI solutions processes is also useful to ensure that a problematic tool can be suspended or removed when necessary.

### 5.5.1. Understandability

**#Impartiality**

The ability to explain the functioning and impact of AI systems is essential to the principle of impartiality and the ability of humanitarian organisations to justify the rationale between humanitarian needs assessment and the design of responses. Algorithmic opacity and the 'black box' effect can constitute significant obstacles to this requirement.

The AI system must be designed to enable and support users to understand its functioning and the relationship between inputs and outputs. User interfaces that systematically rely on interactive engagement and provide adaptive options and explanatory functions will support a continuous learning process and facilitate the feedback loop and participatory requirements. The design of an AI system can embed automatic functions that systematically support the user in identifying its weaknesses and limitations. These elements should also be adapted to the needs of the users and the risks-based sensitivity of the AI system.

for further information, click here to access more resources

### 5.5.2. Incentivizing critical thinking

**#Independence**

Ensuring that humanitarian staff are responsible for their decisions, and able to justify them, is important to support the autonomy of the organization and its ability to justify its choices independently from the potential interference of external actors.

Through adapted training, design and governance measures, an AI system can automatically incentivize and help the user to identify trade-offs, challenge and critically assess its methodology and output. Designing user interface to enhance 'human in the loop' elements can significantly strengthen performance, accountability and responsibility.

To address the inherent tension between control and agency, and speed and easiness of use, AI solution designers need to consider interface elements that can help complexify and slow down decision-making to enhance critical reflection at each stage of the use process. Creating features that design the tool as a critical thinking-partner will support humanitarians in understanding the risks and limitations of using the system without taking the time to think about the consequences; rushed and impatient usage can and should be discouraged through appropriate design.

From a governance perspective, organisational measures and incentives should be created to encourage, support and train staff to ensure that they have the adequate literacy, time and opportunity to effectively review and contribute to the AI system's output. Access to complex and more risky AI tools should be restricted to specifically trained and authorized staff. Such measures are essential to support responsibility and accountability at the individual and organisational levels.

for further information, click here to access more resources

### 5.5.3. Monitoring and takedown protocols

**#Humanity**

Being able to identify and stop using problematic AI solutions is important for organisations who need to be able to justify that the use of their limited resources is effectively used to alleviate human suffering (and not wasted in harmful, useless or counter-productive solutions).

AI systems owners should implement protocols and design features enabling the continuous performance monitoring of the AI systems. Interfaces can be designed to encourage and facilitate user feedback and 'flagging' mechanisms when performance or quality issues are encountered. For sensitive AI systems, clear technical and organizational steps should be defined before hand to be able to take-down problematic AI systems and ensure that contingency plans and resources are in place to maintain operational delivery.

for further information, click here to access more resources

## 5.6. Data protection and information security by design

### 5.6.1. Personal data protection by design

**#Humanity**

In humanitarian contexts, the abuse and misuse of personal data can lead to serious risks for the safety and dignity of vulnerable populations (e.g., attacks, persecutions, arrests, killings). Ensuring that everything is done to protect the security and confidentiality of the personal data humanitarian organisations use in support of their activity is essential to align with the 'do no harm' approach.

In the context of humanitarian activities, many applications rely on the use of sensitive personal data to improve the delivery or design of programs. The collection, processing and management of personal data must strictly follow data protection rules to ensure that the rights of data subjects are respected and that the use of sensitive data in AI systems does not cause harm. The ICRC Handbook on Data Protection in Humanitarian Action provides a complete overview and guidance on how to manage the collection and processing of personal data to ensure a responsible approach to AI in support of humanitarian activities, including through technological design measures.

### 5.6.2. Information security

Beyond sensitive personal data, organizations may deal with other types of sensitive information that needs to stay private or shared under specific rules to third party providers. The design and choices of solutions affect information security and must be in line with the guidelines of the organization, for example regarding the approach to cybersecurity for self-hosted software or on the choice of cloud-based AI solutions.

Technical approaches such as federated learning help preserve information security when training AI models. By training models collaboratively without centralizing data, these techniques can also be leveraged to promote participatory approaches and collaborations where the data stays in the hands of its owners. They also reduce risks linked to centralized vulnerabilities.

for further information, click here to access more resources

## 5.7. Sustainability by design

**#Humanity**

In line with the 'do no harm' approach, humanitarian organisations must take every possible measure to mitigate the negative impact that their activities and operations (including the use of AI solutions) can have on the safety and dignity of the population they serve. This includes mitigating the negative environmental impact that AI solutions can have, as environmental degradation and climate change are known to increase vulnerabilities and contribute to violence and conflict dynamics. The economic and environmental sustainability of AI systems are particularly relevant from a humanitarian, ethical and organizational perspective.

### 5.7.1. Economic sustainability

The ability of an organisation to anticipate, manage and finance the short-, mid- and long-term costs of the integration of an AI system into its operations is essential to ensure that the organisation preserves its autonomy, agency and accountability in the use of its funding. This enables the organization to protect its effectiveness, professionalism and reputation, and to preserve the trust of the population it serves, as well as the confidence of its donors and partners. The procurement due diligence and cost evaluation stages (see above), and design choices that ensure robust functionality, resilience and interoperability over time, are critical to support the economic sustainability of a solution.

### 5.7.2. Environmental sustainability

The development and use of AI solutions have a significant environmental cost; the carbon footprint of the value and supply chains of AI systems or the electricity and water consumption of the data centre and cloud computing services behind them are difficult to fully measure part of that cost. As such, the adoption and use of AI solutions can be significant in humanitarian organizations' efforts to embed environmental considerations and impact mitigation objectives in their activities and operations.

Ensuring that this cost is integrated at the procurement and cost evaluation stages is important to allow the organisation to have a good understanding and visibility of those aspects, and be able to manage competing commitments (e.g., environmental footprint reduction). Decision making on these important trade-offs is an important aspect of AI governance and responsibility. The organization is responsible to embed these considerations in institutional practices and ensure that policy and operational mechanisms are in place. Using specific thematic tools, such as the Climate and Environment Charter for Humanitarian Organization, can be helpful to inform environmental sustainability practices in the AI domain. It is also important to consider that certain AI tools may help organisation to achieve their environmental objectives and mitigation strategies.

for further information, click here to access more resources

# 6 Change management

AI systems are socio-technical ensembles that must integrate into the cultural identity and human elements of an organization.

While most change management challenges are applicable to all digital technologies and not specific to AI, it is important to anticipate how organizational policies and practices will shape the integration and performance of AI systems – and how these will shape staff and organizational behaviours in return.

## 6.1. Monitoring change

It is important to have systems in place to monitor the impact of the deployment of AI systems, in particular when those have a wide impact or important role in the organisation (in line with the risk-based approach discussed above). This should include productivity but also knowledge and behavioural dimensions as the use of certain AI systems can have a negative impact on users' cognitive abilities and agency and other human elements that are important strengths of an organisation. This monitoring requires specific skills and expertise that should be embedded in AI project management.

## 6.2. Preserving internal capacity to achieve humanitarian missions

AI systems impacting core missions of the humanitarian organization should be particularly monitored, from a perspective of capacity-preservation and of professional up-skilling for the staff members. Within the organization, AI systems may move decision-making away from local structure towards central offices. Organizations should assess which specific skills and local knowledge risk being lost through AI integration, document shifting roles across all staff levels and locations, put specific prevention and mitigating measures in place, and provide proactive communication and training opportunities to affected staff.

## 6.3. Building humanitarian literacy on using this technology

The pace of change of AI technologies presents competency challenges that might disempower humanitarians in assessing and evaluating whether and how AI can best support organisations in addressing some of the most pressing needs. Having in place strategies that continuously support professional staff in developing knowledge and on AI skills would build readiness and resilience in strategically defining actions that best harness the opportunities of these technologies, by also weighing the necessary trade-offs. Continuous professional training should be tailored to humanitarian needs to also mitigate the risk of losing necessary operational and strategic knowledge that could be otherwise enhanced by this technology (e.g., presenting different scenarios on a task, rather than providing an answer). Here, humanitarian principles can be crucial in guiding literacy strategies on AI use; this would help move beyond the efficiency framework that often limits understanding of how this technology can be best tailored to humanitarian mindsets, goals, and vocation.

for further information, click here to access more resources

## Further Resources

### 3.1. Continuous adaptation and iterative approach

**ELRHA.** (n.d.). Artificial intelligence for humanitarians: Synthesis report. Retrieved from https://www.elrha.org/resource/artificial-intelligence-for-humanitarians-synthesis-report

**ELRHA.** (n.d.). Monitoring humanitarian innovation. Retrieved from https://www.elrha.org/resource/monitoring-humanitarian-innovation

**Signpost AI.** (n.d.). Towards responsible humanitarian AI: Guidelines from research and practice. Retrieved from https://www.signpostai.org/research/responsible-humanitarian-ai-guidelines-lessons-from-the-field

**UK Humanitarian Innovation Hub & University College London.** (n.d.). Designing and deploying AI tools to support humanitarian practice: A practical guide. Retrieved from https://www.ukhih.org/documents/125/Designing_and_Deploying_AI_Tools_to_Support_Humanitarian_Practice_A_Practical_Guide-V5.pdf

### 3.2. Participatory and Multidisciplinary approach

**CARE International UK.** (n.d.). AI and the Global South: Making AI more ethical and effective through inclusive participation. Retrieved from https://www.careinternational.org.uk/news-stories/ai-and-the-global-south-making-ai-more-ethical-and-effective-through-inclusive-participation/

**UK Humanitarian Innovation Hub.** (2024, December). Humanitarian AI Unpacked – December 2024. Retrieved from https://www.ukhih.org/news/humanitarian-ai-unpacked-december-2024/

**CDAC Network.** (n.d.). Policy brief: Addressing power dynamics in participatory AI for crisis-affected communities (beta version). Retrieved from https://www.cdacnetwork.org/resources/addressing-power-dynamics-in-participatory-ai-for-crisisaffected-communities/

### 4.1 Needs, relevance and usefulness

**Mazzucato, M., & Valletti, T.** (2025, February 11). Governing AI for the public interest. Project Syndicate. Retrieved from https://www.project-syndicate.org/c/q3E36UC (registration may be required)

**NetHope, Inc.** (2024). The guide to usefulness of existing AI solutions in nonprofit organizations. Retrieved from https://nethope.org/toolkits/the-guide-to-usefulness-of-existing-ai-solutions-in-nonprofit-organizations/

**United Nations Office for the Coordination of Humanitarian Affairs (OCHA).** (2014). Humanitarian innovation: The state of the art. Retrieved from https://www.unocha.org/publications/report/world/humanitarian-innovation-state-art

### 4.2. Risks evaluation

**International Committee of the Red Cross.** (2023, July 27). Understanding digital risks in armed conflict. Retrieved from https://blogs.icrc.org/law-and-policy/2023/07/27/digital-risks-in-armed-conflict/ (icrc.org)

**European Union.** (2024). Regulation (EU) 2024/1689 on artificial intelligence (AI Act). Retrieved from https://eur-lex.europa.eu/eli/reg/2024/1689/oj (eur-lex.europa.eu)

**National Institute of Standards and Technology.** (2023). Artificial intelligence risk management framework (AI RMF 1.0). Retrieved from https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10 (nist.gov)

**United Nations System.** (2024). Framework for a model policy on the responsible use of artificial intelligence in UN System organizations. Retrieved from https://unsceb.org/sites/default/files/2024-11/Framework%20for%20a%20Model%20Policy%20on%20the%20%20Responsible%20Use%20of%20AI%20in%20UN%20System.pdf (unsceb.org)

### 4.3. Costs assessment

**World Economic Forum & GEP.** (2023). Adopting AI responsibly: Guidelines for procurement of AI solutions by the private sector. World Economic Forum. Retrieved from https://www.weforum.org/publications/adopting-ai-responsibly-guidelines-for-procurement-of-ai-solutions-by-the-private-sector/

**United Nations Framework Convention on Climate Change.** (2025, July 11). AI and climate action: Opportunities, risks and challenges for developing countries. UNFCCC. Retrieved from https://unfccc.int/news/ai-and-climate-action-opportunities-risks-and-challenges-for-developing-countries.

## 5.2. Procurement: risks and due diligence

**Center for Inclusive Change.** (n.d.). Risk management framework for procuring AI systems. Retrieved from https://www.inclusivechange.org/ai-governance-solutions/rmf-for-ai-procurement/

**MERL Tech.** (2025). Tool for assessing AI vendors. Retrieved from https://merltech.org/resources/tool-for-assessing-ai-vendors/

**World Economic Forum.** (n.d.). AI procurement in a box. Retrieved from https://www.weforum.org/publications/ai-procurement-in-a-box/

**Scottish AI Playbook.** (n.d.). Responsible AI procurement. Retrieved from https://www.scottishaiplaybook.com/responsible-ai-procurement

**United Nations Human Rights Council Working Group on Business and Human Rights.** (2025). Artificial intelligence procurement and deployment: Ensuring alignment with the Guiding Principles on Business and Human Rights (A/HRC/59/53). Retrieved from https://digitallibrary.un.org/record/4087141

## 5.3.1. Data quality

**Signpost AI.** (n.d.). How AI learns, and what it misses: Why data selection matters in humanitarian action. Retrieved from https://blogs.icrc.org/law-and-policy/2025/08/14/how-ai-learns-and-what-it-misses-why-data-selection-matters-in-humanitarian-action/

**Signpost AI.** (n.d.). Humanitarian AI: A literature survey. Retrieved from https://www.signpostai.org/research/humanitarian-ai-a-literature-survey

**ReliefWeb.** (n.d.). Humanitarian "Do No Harm": Plugging gaps in data governance. Retrieved from https://reliefweb.int/report/world/humanitarian-do-no-harm-plugging-gaps-data-governance

**NetHope, Inc.** (n.d.). Nonprofit data governance toolkit guide. Retrieved from https://nethope.org/toolkits/data-governance-toolkit-a-guide-to-implementing-data-governance-in-nonprofits/

**Humanitarian Data Science and Ethics Group (HUM-DSEG).** (2020). A framework for the ethical use of advanced data science methods in the humanitarian sector. Retrieved from https://www.hum-dseg.org/sites/g/files/tmzbdl1476/files/2020-10/Framework%20for%20the%20ethical%20use.pdf

**Sekara, V., Karsai, M., Moro, E., Kim, D., Delamonica, E., Cebrian, M., ... & Garcia-Herranz, M.** (2023). Are machine learning technologies ready to be used for humanitarian work and development?. arXiv preprint arXiv:2307.01891.

## 5.4. Benchmarking performance

**Reuel, A., Hardy, A., Smith, C., Lamparth, M., Hardy, M., & Kochenderfer, M. J.** (2024). What makes a good AI benchmark? Stanford University Human-Centered Artificial Intelligence. Retrieved from https://hai.stanford.edu/policy/what-makes-a-good-ai-benchmark

**Ma, A.** (2025, August). Is your AI benchmark lying to you? The AI Forum. Retrieved from https://www.theaiforum.org/news-were-reading/ai-needs-better-benchmarks

Note: Few resources are available on the benchmarking of AI systems for humanitarian applications. Initiatives like the MOOVE provide a platform to align and benchmark with real-world, humanitarian standards.

### 5.5.1. Understandability

Pizzi, M., Romanoff, M., & Engelhardt, T. (2020). AI for humanitarian action: Human rights and ethics. International Review of the Red Cross, 102(913), 145-180.

### 5.5.2. Incentivizing critical thinking

El-Assady, M., & Moruzzi, C. (2022). Which biases and reasoning pitfalls do explanations trigger? Decomposing communication processes in human–AI interaction. IEEE Computer Graphics and Applications, 42(6), 11-23. Retrieved from https://ieeexplore.ieee.org/abstract/document/9887997

Data Friendly Space. (2025, November 4). The Human in the Loop: How Oversight Turns AI into a Humanitarian Ally. ReliefWeb. https://reliefweb.int/report/world/human-loop-how-oversight-turns-ai-humanitarian-ally.

### 5.5.3. Monitoring and takedown protocols

Khadka, R., & Shah, H. (2024). MLOps as enabler of trustworthy AI. AI and Ethics. Retrieved from https://digitalcollection.zhaw.ch/server/api/core/bitstreams/6fa4cbfb-fa07-4eab-9bc9-59b465bcee12/content

### 5.6.2. Information security

Matthias, A., & Nouwens, M. (2022). Artificial intelligence. In F. K. Mayer & R. van der Hoff (Eds.), Handbook on data protection in humanitarian action. Cambridge University Press. Retrieved from https://www.cambridge.org/core/books/handbook-on-data-protection-in-humanitarian-action/artificial-intelligence/84780BB35FAFF04F403AB7A7AA5C0934

FLock.io. Federated learning: The future of privacy-preserving public sector AI. https://www.flock.io/blog/federated-learning-the-future-of-privacy-preserving-public-sector-ai

### 5.7.2. Environmental sustainability

Here I Am Studio. (n.d.). Why humanitarian tech needs a business plan: Lessons from the private sector. Retrieved from https://hereiamstudio.com/insights/why-humanitarian-tech-needs-a-business-plan

Hugging Face. (n.d.). AI + environment primer. Retrieved from https://huggingface.co/ai-environment-primer

### 6.3. Building humanitarian literacy on using this technology

Spencer, S., W., (2025). Humanitarian AI revisited: Seizing the potential and sidestepping the pitfalls (Network Paper No. 89). https://odihpn.org/wp-content/uploads/2024/05/HPN_Network-Paper89_humanitarianAI.pdf

Humanitarian Leadership Academy. (2025). Initial insights report: How are humanitarians using artificial intelligence in 2025? Retrieved from https://www.humanitarianleadershipacademy.org/resources/initial-insights-report-how-are-humanitarians-using-artificial-intelligence-in-2025/

Sphere Association. (n.d.). AI literacy training for humanitarians. Retrieved from https://spherestandards.org/ai-literacy-training

# References

Amoroso, D. and G. Tamburrini (2021). “Toward a Normative Model of Meaningful Human Control over Weapons Systems.” Ethics & International Affairs 35(2): 245-272.

Bullock, J. B. (ed.) (2024). The Oxford handbook of AI governance. New York, Oxford University Press.
Coppi, G., R. Moreno Jimenez and S. Kyriazi (2021). “Explicability of humanitarian AI: a matter of principles.” Journal of International Humanitarian Action 6(1): 19.

Delgado, F., S. Yang, M. Madaio and Q. Yang (2023). “The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice.” Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization: 1-23.

Díaz-Rodríguez, N., J. Del Ser, M. Coeckelbergh, M. López de Prado, E. Herrera-Viedma and F. Herrera (2023). “Connecting the Dots in Trustworthy Artificial Intelligence: From AI Principles, Ethics, and Key Requirements to Responsible AI Systems and Regulation.” Information Fusion 99: 101896.

El-Assady, M. and C. Moruzzi (2022). “Which Biases and Reasoning Pitfalls Do Explanations Trigger? Decomposing Communication Processes in Human-AI Interaction.” IEEE Comput Graph Appl 42(6): 11-23.
Ferrara, E. (2024). “Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies.” Sci 6(1): 3.

Narayanan, A. and S. Kapoor (2025). “AI as Normal Technology.” Knight First Amend. Inst.
Sai, S., U. Mittal, V. Chamola, K. Huang, I. Spinelli, S. Scardapane, Z. Tan and A. Hussain (2024). “Machine Un-learning: An Overview of Techniques, Applications, and Future Directions.” Cognitive Computation 16(2): 482-506.

Spencer, S. W. (2024). “Humanitarian AI revisited - Seizing the potential and sidestepping the pitfalls “ Humanitarian Practice Network: 1-33.

Yeung, K., A. Howes and G. Pogrebna (2020). “AI Governance by Human Rights–Centered Design, Deliberation, and Oversight: An End to Ethics Washing”, in M. D. Dubber, F. Pasquale and S. Das (ed.) The Oxford Handbook of Ethics of AI. Oxford, Oxford University Press: 77-108.

Information Commissioner’s Office, The Guidance on AI and Data Protection, 2023, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/
Gisli Olafsson, Nethope - The guide to usefulness of existing AI solutions in nonprofit organizations, Part 1, Humanitarian AI, 2024.

SAFE AI Project, Co-Design vs.User-Centred Design for AI Solutions, 2025.

SAFE AI Project, FCDO SAFE AI Roundtable: Building a Responsible AI Framework for Humanitarian Action in a Rapidly Changing Landscape, 2025

Advancing Responsible Development and Deployment of Generative AI: A UN B-Tech foundational paper, 2023
Generative AI for Humanitarians, Digital Humanitarian Network, 2023, https://humanitarianlibrary.org/resource/generative-ai-humanitarians.

Humanitarian AI revisited: Seizing the potential and sidestepping the pitfalls, Humanitarian Practice Network, 2024, https://odihpn.org/en/publication/humanitarian-ai-revisited-seizing-the-potential-and-sidestepping-the-pitfalls/

Humanitarian AI Code of Conduct, NetHope, 2024, https://nethope.org/toolkits/humanitarian-ai-code-of-conduct/

Nonhuman humanitarianism: when ’AI for good’ can be harmful, Mirca Madianou, Information, Communication& Society, 2021, VOL. 24, NO. 6, 850–868, https://doi.org/10.1080/1369118X.2021.1909100
How to Design AI for Social Good: Seven Essential Factors, Luciano Floridi, Josh Cowls, Thomas C. King, Mariarosaria.