HUMANITARIAN LAW & POLICY

How AI learns, and what it misses: why data selection matters in humanitarian action

August 14, 2025, Analysis / Emerging Voices / Humanitarian Action / Humanitarian Principles / New Technologies / Special Themes / Technology in Humanitarian Action

16 mins read

Maria Haas
Researcher, European Centre on Privacy and Cybersecurity (ECPC), Maastricht University

Artificial intelligence (AI) is entering humanitarian action at pace, shaping decisions from crisis forecasting to frontline response. Yet, AI systems are themselves shaped by the data they learn from — and much of that data was never designed with humanitarian use in mind.

In this post, part of the Emerging Voices series, Maria Haas, an Associate at the ICRC Data Protection Office, explores how misaligned training data can embed blind spots and bias into the systems humanitarian organizations rely on, quietly undermining the humanitarian principles of humanity, impartiality, neutrality, and independence. She argues that aligning AI with humanitarian purposes must begin with a critical look at the data that shapes what and how these systems learn — and what they fail to see and integrate.

Humanitarian principles are more than moral ideals. They are *operational foundations*, and the principles of humanity and impartiality particularly are the very purpose behind humanitarian action. They guide humanitarian actors through complex, volatile environments and enable access, trust, and the protection of those most at risk. But what happens when the digital tools we use to uphold and support the operationalization of these principles are trained on datasets — and shaped by design choices — that do not align with them?

This is one of the core dilemmas of using AI in humanitarian settings, where the danger lies not necessarily in ill intent, but in misaligned data and the assumptions it encodes. From *chatbots* that interact with affected populations, to models that help anticipate needs through *predictive analytics*, support *healthcare* delivery in the field, or *detect patterns of violence* in armed conflict, AI is rapidly shaping humanitarian response, promising a more effective and fair use and distribution of scarce humanitarian resources.

Yet, AI systems are never fully neutral. They are designed, developed, and deployed in a specific context, by specific actors, and with specific purposes in mind. Moreover, like any machine learning (ML) process,[1] they follow a *lifecycle* — from problem formulation and data selection to training, (re-)evaluation, and deployment. At the heart of this process lies the data on which the model is trained. Static models draw once from datasets, locking in patterns that persist over time; dynamic models continue drawing, adjusting their outputs as new data and feedback flows in. In any case, what a system learns at this stage — or fails to learn — fundamentally shapes its capacities, scope, and limitations.

Many humanitarian organizations are already grappling with these challenges. The ICRC, for instance, has issued internal guidelines and recently a dedicated *AI Policy* aimed at responsible use of AI. Yet, as essential as these frameworks are, they often leave unaddressed the concrete decisions — like data selection — that shape humanitarian responses and affect the extent to which they align with humanitarian principles.

Just as a person's worldview is shaped by the stories they hear and the experiences they absorb, an AI system's "understanding" is shaped by the data it consumes. If that data lacks cultural nuance or the texture of lived experience, the system cannot fabricate them on its own. And when trained on unrepresentative or decontextualized data sources, models may internalize and quietly reproduce distorted views at scale.

In a sector governed by foundational principles, such distortions are more than technical imperfections. While "humanitarian AI" covers a wide range of tools — many of which resemble applications in other sectors, such as logistics or finance, with fewer data-related risks — this analysis focuses on systems embedded in core humanitarian operations, where any data-based distortion risks misguiding priorities, doing or amplifying harm, weakening the trust upon which humanitarian action depends, and ultimately working against the purpose they were intended to serve.

Missing what matters: what data can't capture

Human suffering must be addressed where it is found. This tenet lies at the heart of the principle of humanity — the moral and operational foundation of humanitarian action. It is more than an abstract ideal: it is a call to protect life, uphold

dignity, and act with compassion in the face of suffering.

In his foundational work *Un Souvenir de Solferino* (A Memory of Solferino), Henri Dunant did not write about "needs assessments" or indicators. He bore witness to the suffering of wounded soldiers, as fellow humans in pain:

The poor wounded men that were being picked up all day long were ghastly pale and exhausted. Some, who had been the most badly hurt, had a stupefied look as though they could not grasp what was said to them; they stared at one out of haggard eyes, but their apparent prostration did not prevent them from feeling their pain. Others were anxious and excited by nervous strain and shaken by spasmodic trembling. Some, who had gaping wounds already beginning to show infection, were almost crazed with suffering. They begged to be put out of their misery, and writhed with faces distorted in the grip of the death-struggle. (1862, *p. 44*)

For Dunant, suffering was not a category to be measured, but a reality to be met with presence, empathy, and care. Humanity, as a principle, is not only about *what* is done, but *how* it is done — the way assistance is delivered, the attention paid to dignity, and the willingness to understand pain that cannot be quantified.

Today's AI systems, however, are trained on data — not experience. When that data is scraped from *social media*, *market trends*, or *open-source repositories* not designed with humanitarian use in mind, models may adopt assumptions that strip away the nuance, emotional weight, and lived realities that humanitarian responses must be built upon. The markers inferred from such data are inherently reductive: compressing complexity into classification, often *trading texture for efficiency*.

Algorithms may *simulate aspects of empathetic response*, but the subtlety of trauma or grief, cultural meanings, and the silent dimensions of dignity often elude systems trained solely on textual or behavioural data. AI-generated content tools such as chatbots built on large language models (LLMs), for instance, may appear fluent, but are essentially "*stochastic parrots*," stitching together statistically likely outputs based on correlations in training data without genuine understanding of meaning or context. A chatbot meant to offer psychosocial support may fail to recognize distress cues that don't conform to the specific cultural expressions it has been trained on, suggest irrelevant resources, or hallucinate reassuring answers detached from the user's reality.

While human support has its own limitations, these risks must be carefully weighed against available alternatives, especially in high-stakes environments where misunderstanding can *increase* human suffering and ultimately undermine the very principle of humanity upon which humanitarian action is built.

Training on the wrong signals: how AI learns misaligned priorities

If the previous section highlights what AI systems *cannot* encode — empathy, dignity, the lived experience of suffering — this section turns to what they *do* encode. In humanitarian contexts, political repression, conflict, and infrastructural collapse often mean that disaggregated or context-sensitive information is *unavailable, inaccessible, or outdated*. And even when data exists, it may reflect only parts of a population, omitting those who are less digitally connected, harder to reach, or excluded from formal systems. If AI models learn from such incomplete inputs, they risk reinforcing blind spots and overlooking those most at risk.

That said, humans are not inherently better at identifying needs or patterns. In some cases, AI systems can enhance humanitarian analysis by revealing counterpoints or gaps that may have escaped humanitarian actors' minds at a given time. Problems arise when AI is used beyond the limits of what its data can meaningfully support, when flawed inputs are treated as comprehensive, and model outputs are trusted without critical scrutiny. These risks are compounded when tools are deployed without proper safeguards — an issue addressed in more detail elsewhere.

These challenges are exacerbated by the *funding pressures* currently facing the humanitarian sector. With key humanitarian data systems — such as FEWS NET, a critical early warning tool for hunger crises — now facing *serious uncertainty*, and *no comparable alternatives* in place, there is a risk of a chain reaction where the collapse of interdependent data streams not only delays immediate response by potentially leaving crises undetected, but also degrades the availability and quality of future training data.

While more data might seem like the answer, humanitarian actors are generally bound by data protection principles, including data minimization. *The principle matters* precisely because necessity must be critically assessed: collecting more data may seem justified in crisis settings, but without clear limits and safeguards, it risks exposing individuals to additional harm and undermining their dignity.

Moreover, data *scarcity* is only one part of the problem. Very often, the issue rather lies in *bad data*. As the expression goes: *bias in, bias out*. Even seemingly rich datasets often obscure the political and structural conditions that shape them. *Far from neutral*, data can reflect entrenched inequalities, geopolitical priorities, and visibility biases influenced by donor interests and media attention. Consider systems trained on geospatial data from conflict or disaster zones — *satellite imagery*, *drone footage*, *Call Detail Records (CDRs)*, or *location-tagged social media posts*. If such data clusters around regions with greater media coverage or where donor-funded projects have historically been concentrated, AI models may begin to associate urgency not with humanitarian need, but with historical visibility — reinforcing cycles of over-recognition and neglect.

These patterns are further reinforced by the architecture of open-source repositories, which often privilege *digitally connected areas*, amplifying Americentric and Eurocentric worldviews. Remote, marginalized, or linguistically diverse communities are frequently underrepresented, and models trained on such data *tend to perform worse* for these communities.

In politically sensitive contexts, the risks multiply. Available data may be curated or influenced by actors with vested interests such as *governments* or *parties to a conflict*. Even media accounts, often well-intentioned and aimed at holding power to account, can be shaped by the *perspectives* and *assumptions* of those reporting them. AI systems trained on such sources risk absorbing and reproducing external narratives, rather than the needs and realities of affected communities.

Bias in, principles out: how flawed data undermines the foundation of humanitarian action

When partial or flawed data becomes the foundation for humanitarian decision-making, it can quietly undermine a set of fundamental *humanitarian principles*.

Impartiality — the commitment to serve based on needs alone — is at risk when AI is used for needs assessments and/or to direct humanitarian action to specific communities but the training data overrepresents groups or regions that are more digitally visible or historically well-funded, leading models to perform better for them than for others who may be equally or more in need but underrepresented in the data.

Neutrality can be compromised when training data reflects politicized narratives. Additional risks emerge when data is sourced through partnerships with companies like *Palantir*, *NASA*, or *commercial satellite firms* that may also serve non-humanitarian clients, raising the risk that data collected for humanitarian purposes may be repurposed for military or surveillance objectives — blurring humanitarian boundaries and threatening perceived neutrality.

Independence falters when humanitarian organizations become *reliant* on data produced or maintained by actors with external agendas who are not bound by humanitarian principles.

When any of these principles are compromised, the consequences strike at the heart of humanitarian legitimacy and the trust upon which humanitarian action depends. If affected populations even just perceive AI tools as biased, politicized, or disconnected, they may lose trust in the humanitarian actor and disengage entirely. The less trusted a system and the less people engage in it, the less data it can ethically collect, and the worse it performs.

However, at a time when humanitarian needs are growing and resources are scarce, the relevant question might not be whether an AI system is perfect, but whether it offers meaningful support in complex humanitarian situations. Many existing non-AI systems, where they exist at all, face similar constraints: limited data, human error, resource shortages, inconsistent access to expertise. In some cases, AI may offer imperfect but valuable support. A medical language model like *Meditron*, for instance, currently considered by the ICRC, may not perform flawlessly across all settings, but could still improve the existing (direly lacking) access to critical medical knowledge by practitioners in under-resourced environments. The question, then, is not whether AI systems are bias-free, but whether their risks are acceptable (and appropriately considered, communicated, re-evaluated) in light of the alternatives. There rarely is an easy answer, but evaluating AI tools in isolation, without reference to the systems they might complement or replace, risks missing the practical trade-offs that humanitarian actors must *navigate*.

From principle to practice: accountability starts with the data

In practice, most humanitarian organizations aren't building advanced AI systems from scratch. Instead, they often rely on external models, especially for the most advanced tools like LLMs, typically sourced from a handful of dominant commercial providers with the *resources* and infrastructure to develop them. These are often trained on massive, proprietary datasets, with little visibility into what data was included, how it was selected, or whose voices it reflects. Even humanitarian-focused projects often rely on these providers, whether by drawing on existing models (such as the LLMs from *OpenAI, Anthropic, and Google*), or by partnering through initiatives like Microsoft's *AI for Good Lab* or Meta's *Data for Good* platform.

This raises particular responsibilities regarding procurement decisions: They must go beyond technical performance to include an evaluation of the training data itself – its composition, quality, origin, and relevance to humanitarian realities.

Here, data protection frameworks offer concrete principles and requirements that can *guide* and help this exercise. Even when personal data is anonymized or aggregated, *data protection principles* like risk assessment, transparency, and accountability can help ensure that harm is minimized and that fairness and transparency are built into system design from the start. Risk assessments pertaining to the training data should scrutinize accuracy, representational gaps, and potential proxies for sensitive attributes that may inadvertently replicate bias or marginalize the very communities they aim to support. Explainable AI (XAI) can *support these efforts* by making AI systems understandable to their human operators, helping trace how certain outputs were generated, identify potential logic flaws, and enable oversight by keeping the *humanitarian in the loop*.

But identifying gaps depends not just on explainability, but also on the quality of *evaluation* (or test) data – which is *distinct from training data* and used to test the model independently after development. How we evaluate models shapes where they are trusted, accepted, and deployed. If test data lacks diversity or skews toward more visible populations, it can mask the very gaps the model perpetuates and create a false sense of confidence in the model's reliability. Ensuring that evaluation sets reflect humanitarian realities, and are not merely engineered for high benchmark scores, is therefore just as critical as scrutinizing the data a model is trained on.

Aligning AI with humanitarian purpose

As AI tools become more embedded in humanitarian action, the Fundamental Principles of humanity, impartiality, neutrality, and independence must serve not just as ideals but as active benchmarks for evaluating the design, development, and deployment of AI systems – including and especially the data that shapes them.

Not all risks can always be eliminated, but they must be clearly identified, weighed against alternatives, communicated, and acted upon. In some cases, humanitarian actors must be prepared to walk away when a model, however technically advanced and well-designed, is simply misaligned with the values or operational realities of humanitarian work.

There are different approaches to improving this alignment. Sector-wide coordination may be one, if it holds any weight with private-sector actors. Others include *early* engagement with tech providers, cross-disciplinary collaboration, and partnerships with academia, civil society, or local actors. Pursuing greater technological independence is another option, though one that comes with trade-offs. Humanitarian organizations must advocate for shared ethical and technical standards, translate them into new contexts, and insist on responsible data governance. At the same time, they must build internal capacity, because without technical literacy, meaningful oversight becomes impossible.

Humanitarian action begins long before the deployment of any service. It begins with intention – with the questions asked, the risks weighed, and the values embedded from the start. This is as true for algorithms as it is for field operations. If AI is to serve the humanitarian mission, its foundations must be built on the same principles it aims to uphold, ensuring that we know what we build on before we let it shape the lives of those we aim to serve.

References

[1] While this piece focuses on machine learning systems, including deep learning models based on neural networks like Large Language Models (LLMs), these represent just one family of approaches within the broader field of AI. Other techniques, such as symbolic AI, including rule-based systems and Bayesian networks, remain relevant in certain applications. However, ML has become the dominant approach in humanitarian contexts due to its scalability and pattern recognition, making it the central reference point for this analysis. See, *https://www.cambridge.org/core/books/handbook-on-data-protection-in-humanitarian-action/artificial-intelligence/84780BB35FAFF04F403AB7A7AA5C0934*.

See also:

- Joanna Wilson, *AI, war and (in)humanity: the role of human emotions in military decision-making*, February 20, 2025.
- Elke Schwarz, *The (im)possibility of responsible military AI governance*, December 12, 2024.
- Erica Harper, *Will AI fundamentally alter how wars are initiated, fought and concluded?*, September 26, 2024.
- Matthias Klaus, *Transcending weapon systems: the ethical challenges of AI in military decision support systems*, September 24, 2024.
- Jimena Sofía Viveros Álvarez, *The risks and inefficacies of AI systems in military targeting support*, September 4, 2024.

Tags: Artificial Intelligence, Fundamental Principles, humanitarian action, humanitarian principles, machine learning, protection, Techplomacy

You may also be interested in:

International humanitarian law and connectivity disruptions during armed conflict

13 mins read

Analysis / Emerging Voices / Humanitarian Action / Humanitarian Principles / New Technologies / Special Themes / Technology in Humanitarian Action

Tilman Rodenhäuser

"Without information and telecommunication, people don't know where to go for safety," the ICRC reported ...

Offline and in danger: the humanitarian consequences of connectivity disruptions

11 mins read

Analysis / Emerging Voices / Humanitarian Action / Humanitarian Principles / New Technologies / Special Themes / Technology in Humanitarian Action Cléa Thouin

As people around the world become increasingly reliant on digital and telecommunications networks to access ...