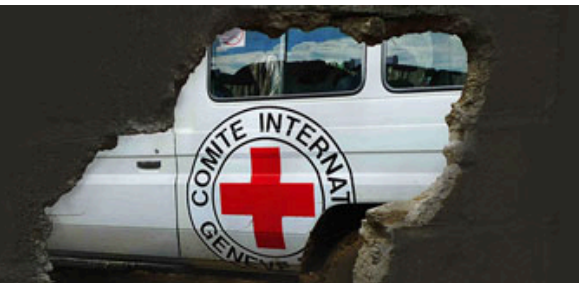
HUMANITARIAN
LAW & POLICY



Trying to square the circle: the ICRC AI Policy

November 28, 2024, Accountability / Analysis / Humanitarian Action / New Technologies / Technology in Humanitarian Action

16 mins read

Pierrick Devidal
Policy Adviser, ICRC

The development of artificial intelligence (AI) technologies brings significant opportunities, and risks, for principled humanitarian action. While AI innovations advance at a pace that seemingly defies human capabilities to manage them responsibly, humanitarian organizations are chasing 'AI for Good', and struggling to find effective safeguards.

In this post, ICRC Senior Policy Adviser Pierrick Devidal reflects on some of the lessons from the ICRC's experience in building its recently adopted AI Policy, with the hope that it can inform other efforts to build an ethical and responsible approach to the use of AI in the humanitarian sector.

It is hard to think straight about AI. Especially as a humanitarian.

On the one hand, the *hype and the hope* around the *AI wave* have triggered a general 'fear of missing out' (*FOMO*) and pushed everyone, including humanitarians, to jump on the bandwagon. There is a *rush* to find use cases for an all-purpose technology that brings immense promises, including in places and for people affected by armed conflict and humanitarian crises.

Contending with continuously unmet needs across the world, humanitarians are particularly susceptible to the prophecies of *techno-solutionism*: when a new technological solution brings *potential* to improve the effectiveness of our work, we get excited. And in the hope that it can make a difference for the many people suffering in front of us, sometimes we get *over*-excited. This natural pro-innovation bias is amplified by the ambient *techno-determinism*. There is a strong impression that AI is not a choice, but a *diktat*, a non-negotiable requirement for a contemporary humanitarianism that cannot afford to lag behind technological developments.

In other words, there is no time to think, and not using AI would be irresponsible.

On the other hand, the *moral panic* and dystopian fears around the perils of AI can be paralyzing for humanitarian professionals. It can be hard to grasp the relevance of AI technologies for communities living in remote areas where food is lacking and where mere connectivity remains a distant reality. The alienating *language* and technical opacity of the AI conversation are also difficult for humanitarians to connect to. While the *list* of humanitarian AI innovations is growing – from predictive analysis supporting program planning and needs assessments, to chatbots managing information requests and engagement with affected populations – debates around the *risks and responsibilities* of using AI are multiplying in humanitarian forums. The challenges at play are daunting.

In other words, there is no time to think, and using AI would be irresponsible.

We have seen these push-and-pull factors play out at the ICRC over the last two years. When we started to map our AI uses 18 months ago, the excitement behind colleagues' enthusiasm to explore AI for humanitarian impact was very palpable. What was also clear was the matching sense of fear and uncertainty vis-à-vis a technology that many of us did not really understand.

Talking to colleagues across the organization, we realized how deeply AI was already embedded in the ICRC's operations (e.g., from information management to logistics support systems) and that it was driving many innovation *projects*. We also realized the disparities in the digital literacy and ability of our staff to understand how AI functions, and what it means in humanitarian terms. While some had advanced expertise in AI, others could not see the basic risks attached to it. While some wanted to accelerate, be daring and competitive, others felt concerned and wanted to slow down or opt-out.

What emerged from this initial mapping was a pressing need for guidance to support ICRC staff to learn and explore if, how, when and where AI can help them implement the ICRC's mandate.

Designing a policy to make humanitarian sense of AI was challenging. We turned to humanitarian principles as an *ethical compass*, and to provide us with a consensual starting point from which to approach the humanitarian AI dilemma.

It still wasn't easy.

Deconstructing our biases about AI

The main challenge we faced was being able to deconstruct the cognitive biases that prevent us from thinking objectively about AI.

The first one comes from the binary framework that often characterizes the debate around AI –whether it will save or destroy the world. Those confrontational and polarized visions of AI are not useful. It is not AI that will save or destroy the world, *but humans*. Our first decision was therefore to keep a *human-centered* approach to AI as the red thread in our policy.

The second bias is a combination of fascination and emotional biases that drive us to be hopeful vis-à-vis any new technology that might help us achieve our goals. In practice, this cognitive bias often materializes into a solution-driven approach that takes AI as the starting point. The question becomes: what can we do with it? But AI is *not* a new technology, and humanitarians are not here to innovate for the sake of *innovation*. Using our *problem-driven* and human-centered approach, we took humanitarian needs, instead of AI, as the starting point. The question became: what is the humanitarian problem, and *can* AI help us respond to it – ethically, safely – and if so when, where and how?

The third bias is the belief that technologies, including AI, are 'just tools' that do not carry intrinsic values or issues. This implies that everything depends on us, and that if we use it to do 'good' (i.e., humanitarian action), AI will be good. But technologies have *never* been *neutral*, and AI is a great example of their *political* and '*dual use*' nature. Aware that if we are not interested in politics, politics is interested in us – including through AI – we placed the humanitarian principles of neutrality and independence at the center of our approach to help us manage those tensions.

Acknowledging and addressing those different cognitive traps was very helpful to create a clearer shared thinking space to deconstruct complexity and build our own, *humanitarian*, approach to AI.

Towards a humanitarian approach to AI

The principle of humanity drives the humanitarian endeavor, and ICRC action. It is a useful *bottom line* to deal with AI. To avoid falling into the general productivity-driven perspective of AI, our policy adopts a value-driven prism of human rights and vulnerabilities. We are committed to doing what we can, including using AI, to alleviate suffering and protect the rights and dignity of people affected by conflict and other humanitarian crises. To do so responsibly, effective safeguards must be identified to mitigate the risk that AI solutions add algorithmic forms of abuse or discrimination to the lives of people already suffering – *before* those solutions are deployed.

This precautionary approach means carefully weighing the risks of using AI in each situation, and using AI *only* if and when it can make a tangible positive difference in the lives of affected people or on the ICRC's ability to protect and assist them. This means avoiding AI if it erodes our ability to interact directly and be physically present. It also means abstaining from using AI when effective safeguards are not available to ensure we '*do no harm*'.

Effective safeguards must assess what AI solutions are used for, how they impact affected people, and what type of data they are relying on – in particular if they involve personal, confidential or sensitive data. AI brings a magnifying glass to data availability, quality and reliability, and algorithms will only ever be as good as the data behind them. To get humanitarian AI right, we also need to get data protection right. And despite tremendous progress over the past *ten* years, there is still a lot of *work* to do in that domain.

The increased polarization and strategic *competition* around AI at the global level is also turning humanitarian use of AI into a tricky political issue. Using AI or not, for what and with whom, have become political choices that can influence how we are perceived by people and communities, and by governments and parties to conflict. For humanitarians, managing that perception and the trust, access and security that comes with it is critical. It is therefore urgent to upgrade our due diligence capabilities and ensure that procurement processes adequately address the perception risks related to what AI solutions we adopt, and to which provider we get them from.

The regulatory vacuum myth: there is no need to reinvent the wheel

The ongoing narrative often creates the impression that AI is emerging from a vacuum. As is often the case with technological innovations, there is a sense that because AI is 'new' and its uses are multiplying so quickly, we do not have the right policy tools or normative frameworks to deal with it.

This assumption is not accurate. We already have plenty of legal norms, ethical principles and innovation good practices that apply to AI: *international human rights* and *humanitarian law*, national and international legislation on *data protection*, *product liability*, *intellectual property*, and innovation *ethics*, for instance. Using those norms and standards is critical to regulate AI systems *before* they are launched, instead of making up new rules to justify their misalignment retroactively.

The other good news is that AI experts have already identified the *essential principles* (and *traps*) that define what a responsible and ethical approach to AI looks like. *Many* ethical guidelines have emerged in the past few years to help align AI with human values. As humanitarians, our *task* is therefore not to reinvent the AI ethical wheel, but rather to define how we transpose those principles to the humanitarian domain. This requires translating AI concepts into the language of humanitarian action, so that humanitarians are able to speak and reflect about AI with a grammar that makes sense to them. The ICRC AI Policy is such a translation effort.

Equipped with the right language, it becomes easier to engage with AI experts, discuss challenges and learn from their practice to explore solutions. We proactively sought the support of external experts in machine learning and AI ethics to help us cross-check our approach and distinguish the known and unknown unknowns. It made a big difference. There is nothing like collective *human* intelligence to make sense of *artificial* intelligence.

It is all about context

One of the many things we've learned through these conversations is that you always need to think about AI in context: the *external* context of the places where AI solutions are intended to be used or to have effects; and the *internal* context and the eco-system in which they are meant to be deployed, taking into account existing digital infrastructures and data landscape, organizational strengths and weaknesses, socio-cultural and human factors. Thinking about what AI can do *for* humanitarian action is not enough. You also need to think what it can do *to* humanitarian action.

In a context of stagnating humanitarian budgets, rushing towards potentially cost-saving solutions makes sense. But the calculus tends to be done on the short-term savings, while AI investments should be considered over a long-time horizon. What you may eventually save in automating human labor should be balanced against: the required investment in recruiting AI experts, acquiring additional computing capacities, the staff time required to assess precaution and mitigating measures, the training needed so that staff using AI do so in line with policy commitments, and the risk of *vendor lock-in* effects. AI cost effectiveness is a potential, but it is a fool's promise if not assessed against the holistic cost.

The analysis required to assess costs and benefits demands a multidisciplinary approach that brings critical thinking from across the organization and along the humanitarian value chain. Understanding systemic and operational constraints, what users need and their operational setting, or where data is an opportunity and where it is a threat, takes time and effort. It requires relying on both those who have technical AI expertise, and on those who can anchor proposed AI solutions in both operational realities and humanitarian ethics.

The building blocks of responsible AI

To build a responsible approach for the ICRC, we used AI *ethical principles* that have emerged as minimum international standards. These are based on the frameworks mentioned above, and boil down to the following core guidance for ICRC staff making decisions about AI.

The precautionary principle

The relevance of the *AI precautionary principle* is quite evident considering the many remaining questions around the impact and risks of AI in the humanitarian domain. Combined with an approach based on proportionality, it helps us focus on using AI to resolve existing problems when there is evidence that it will bring a positive impact, in context.

'Do no Harm'

Placing safety and security at the center is a '*do no harm*' requirement of humanitarian ethics. In practice, this means explicitly identifying the type of data that will feed AI systems, and the nature and potential impact of the risks attached to their use. *Differentiated* data protection and cybersecurity risk mitigation systems should be put in place to match the sensitivity of the data being processed and the consequences of the risks for the end-user. While AI solutions for back-end and administrative processes can be low risks, those on the front-end that have an impact on program delivery or on interactions with affected populations can have deep consequences on their lives and need to be managed as such. Protecting data is, literally, protecting people.

Most commercial AI systems are '*black boxes*'

Transparency

, making it nearly impossible to fully analyze their inner working. But in the humanitarian sector, transparency and 'explainability' are particularly *important*. It is critical that humanitarian organizations make a proactive and continuous effort to be transparent and to explain – at minimum – when, why and how they use AI to carry out their activities. The litmus test is that, if you are not able to do so, you should probably not use it. Open-source AI solutions can help in that respect, because they offer a greater level of transparency than closed proprietary models.

Finding the right way to explain our approach to the people we serve, in a language that they can understand, can be challenging. And it is established that humanitarian organizations already have a *deficit* in terms of accountability to affected populations. There is a significant risk that the use of AI could undermine ongoing efforts and the progress achieved in this domain if our *accountability* gets lost in algorithmic *translation*. This is why the principles of responsibility and accountability are critical when you want to use AI for humanitarian purposes.

Governance and accountability

A human-centered approach also demands that AI tools are regularly reviewed by qualified professionals throughout their lifecycle. It requires us to upgrade our toolbox and structure our governance and control mechanisms to ensure that our use of AI is compliant with applicable laws and regulations, and with humanitarian *ethics*. This means bringing

legal experts to support engineers and decision-makers, and upskilling existing governance structures to manage the implications of AI for humanitarian action. It means having actionable complaint and remedy processes in place, and being able to offer effective and accessible alternative systems for users who cannot, or do not want to, rely on AI solutions. And it also means considering the negative impact that they may have on our ability to *reduce* the environmental *impact* of humanitarian activities and to enhance their *localization*.

It is important to remember that responsibility and accountability come together. They are interdependent and can only be effective through meaningful transparency and continuous *engagement* with relevant stakeholders and users. The human feedback loop is essential to avoid the risks of *collapse* that comes with the AI one.

The start of a learning journey

Building an AI policy to guide our organization and help colleagues navigate AI challenges was not an easy task. And yet, adopting a policy is only the beginning of the ICRC's humanitarian AI journey.

We have a long way to go, but we are already providing an AI e-learning program to help ICRC staff acquire the data and digital literacy they need to explore AI solutions for humanitarian impact. The ICRC Digital Governance Board will use the AI policy to oversee compliance and implementation, and we will continue to engage with external experts and peers to learn and improve.

It is possible that some of our AI initiatives will fail or fall short of expected outcomes. Learning by doing and making mistakes is fine, as long as it does not negatively affect the people we work with and for, or come at a disproportionate cost.

The critical challenge will be not to confuse the urgency to *think* about AI with an urgency to *use* AI. In that sense, the ICRC AI policy is meant to be aspirational and to provide the foundations we need to address the humanitarian AI challenge ethically and successfully. As humanitarians, it is critical we steer away from the dehumanization, abuse and discrimination that bad AI might bring on the people we work with and for.

See also:

- Erica Harper, *Will AI fundamentally alter how wars are initiated, fought and concluded?*, September 26, 2024
- Roxana Radu, Eugenia Olliaro, *Not child's play: protecting children's data in humanitarian AI ecosystems*, December 14, 2023
- Pierrick Devidal, *'Back to basics' with a digital twist: humanitarian principles and dilemmas in the digital age*, February 2, 2023

Tags: AI, artificial intelligence, ethics, humanitarian action, humanitarian policy, new technologies

You may also be interested in:

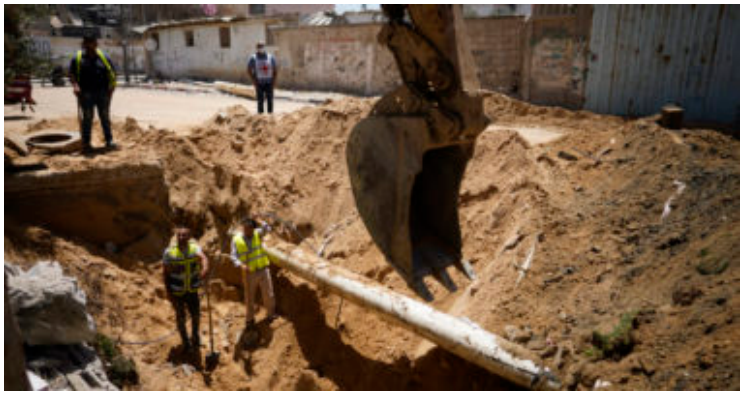
Climate action in conflict and fragile settings: closing the implementation gap

12 mins read

Accountability / Analysis / Humanitarian Action / New Technologies / Technology in Humanitarian Action

Catherine-Lune Grayson & Amir Khouzam

Communities in conflict-affected areas are direly impacted by growing climate risks and shocks. Over the ...



Protecting essential service personnel is a vital part of humanitarian action

9 mins read

Accountability / Analysis / Humanitarian Action / New Technologies / Technology in Humanitarian Action

Marnie Lloydd, Peter Herby, Caroline Baudot & Tobias Ehret

Water and wastewater pipelines, electricity lines and telecommunication installations permeate contemporary urban landscapes and form ...