ROME HUMANITARIAN CONGRESS 2025

Sant'Anna School - Working Papers

Preventing Child Recruitment in Conflict Zones: The Potential and Pitfalls of Ethical AI in Humanitarian Practice

Written by Safa Yagoub

Abstract:

The recruitment of children by armed groups remains one of the most persistent violations of international law in protracted conflicts. In regions such as North Darfur, where formal child protection systems have collapsed, humanitarian actors face the dual challenge of sustaining community-based protection while navigating the rise of new technologies. This paper examines the potential and risks of Artificial Intelligence (AI) in supporting child recruitment prevention strategies, grounded in both legal frameworks (the Convention on the Rights of the Child, International Humanitarian Law, and the Paris Principles) and insights from previous fieldwork in Darfur. By drawing on feminist and community-led approaches, the study assesses how tools such as predictive analytics, satellite imagery, and low-bandwidth digital platforms could strengthen early warning systems and risk monitoring. At the same time, it identifies pitfalls, including bias, surveillance, and cultural misalignment, that may undermine trust and protection outcomes. Using North Darfur as a reference case, this paper proposes a rights-based and participatory framework for ethical AI design that emphasizes local agency, accountability, and humanitarian principles. The analysis contributes to current humanitarian debates by showing that AI can only be effective when embedded in the social fabric of affected communities, never as a substitute for it.

Keywords: CAAFAG, Ethical AI, Humanitarian Access, Darfur, IHL, Child Protection, Community-Based Approaches, Conflict-Affected Communities.

Table Of Contents

Executive Summary

Preventing the recruitment of children in conflict zones requires more than technical innovation; it depends on rights-based, community-led strategies that reflect local realities. This paper argues that while AI tools hold potential for strengthening early warning systems, monitoring risks in displacement camps, and supporting education or psychosocial interventions, they can only play a supporting role. The North Darfur context demonstrates this balance: despite the collapse of formal systems, communities continue to rely on Madrasa safe spaces, extended kinship networks, and women-led protection groups. These structures offer resilience that digital tools should adapt to, not replace. The analysis also highlights serious pitfalls, including dataset bias, cultural insensitivity, exclusion from digital spaces, and a lack of transparency. The central conclusion is one of cautious optimism: AI can enhance child protection efforts when guided by humanitarian principles, legal safeguards, and participatory design, but without such safeguards, it risks reinforcing inequalities and undermining trust. The paper calls for responsible innovation in humanitarian practice—where communities remain the primary agents of protection, and technology is carefully governed as a complementary tool.

1. Introduction

As humanitarian actors confront increasingly complex conflicts, rising displacement, and the pressure to do more with less, the use of AI is rapidly entering humanitarian policy debates[1]. From predictive analytics to satellite surveillance and automated triage systems, AI-based tools are being explored as ways to enhance targeting, improve early warning systems, and strengthen data management in crisis settings[2]. However, these technologies are rarely developed with the lived realities of conflict-affected communities in mind—particularly in contexts like north Darfur, where formal child protection systems have collapsed, and community-led mechanisms have emerged in their place[34].

The recruitment of children by armed actors remains one of the most persistent violations in Darfur and similar protracted crises[5]. Prevention efforts have traditionally relied on legal frameworks, reintegration programs, and education[6]. Yet, these have often fallen short due to security constraints, limited funding, and cultural gaps in design[7]. As digital technologies become more accessible, they present new opportunities—but also new ethical and operational risks—for reimagining prevention strategies, especially in hard-to-reach or insecure areas[8].

1.1. The Problem of Child Recruitment in Conflict Zones

In many conflict-affected settings, children face an impossible choice between survival and safety[9]. Armed groups often target them for recruitment, exploiting their vulnerabilities and the absence of protective systems[10]. Recruitment is not always forced at gunpoint—it can occur through promises of food, protection, or revenge[11]. Once recruited, children are exposed to extreme violence, denied education, and stripped of their rights and agency.

The issue of Children Associated with Armed Forces and Armed Groups (CAAFAG) and its patterns have evolved. Protracted conflicts, increasing displacement, and weakened state institutions continue to

[1] Ana Beduschi, “Harnessing the Potential of Artificial Intelligence for Humanitarian Action: Opportunities and Risks,” *International Review of the Red Cross* 104, no. 919 (2022), pp. 1149–1169, doi:10.1017/S1816383122000261. Available at: Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks.

[2] Michael Pizzi, Mila Romanoff & Tim Engelhardt, “AI for Humanitarian Action: Human Rights and Ethics,” *International Review of the Red Cross* 102, no. 913 (2020), pp. 150–151, doi:10.1017/S1816383121000011. Available at: AI for humanitarian action: Human rights and ethics.

[3] Victor H. Mlambo, Siphesihle Edmond Mpanza & Daniel N. Mlambo, “Armed Conflict and the Increasing Use of Child Soldiers in the Central African Republic, Democratic Republic of Congo and South Sudan: Implications for Regional Security,” *Journal of Public Affairs* 19, no. 2 (2019), pp. 125–138, doi:10.1002/pa.1896. Available at: (PDF) Armed conflict and the increasing use of child soldiers in the Central African Republic, Democratic Republic of Congo, and South Sudan: Implications for regional security.

[4] Safa Yagoub, “Beyond the Reintegration: The Role of Women in Preventing CAAFAG in North Darfur,” *Journal of Social and Political Sciences* 8, no. 3 (2025): 177–178, https://doi.org/10.31014/aior.1991.08.03.592.

[5] UNICEF, “Sudan: Up to 6000 Child Soldiers Recruited in Darfur,” *ReliefWeb* (23 December 2008), accessed 25 September 2025, https://reliefweb.int/report/sudan/sudan-6000-child-soldiers-recruited-darfur-unicef.

[6] UNICEF, *The Paris Principles: Principles and Guidelines on Children Associated with Armed Forces or Armed Groups* (February 2007), p. 7, https://www.unicef.org/mali/media/1561/file/parisprinciples.pdf.

[7] Michelle Legassicke, Dustin Johnson & Catherine Gribbin, “Definitions of Child Recruitment and Use in Armed Conflict: Challenges for Early Warning,” *Civil Wars* 26, no. 3 (2024), pp. 430–454, doi:10.1080/13698249.2023.2167042.

[8] Tino Kreutzer, Solveig ten Berge, Marie de Briant, Susan Ngigi, Jesse Kamstra, Alex Schade & Simon M. Opiyo, “Ethical Implications Related to Processing of Personal Data and Artificial Intelligence in Humanitarian Crises: A Scoping Review,” *BMC Medical Ethics* 26, no. 49 (2025), p. 12, https://doi.org/10.1186/s12910-025-01189-2.

[9] ICRC, “Child Soldiers,” *ICRC Casebook: International Humanitarian Law* (updated 2013), accessed 26 September 2025, https://casebook.icrc.org/a_to_z/glossary/child-soldiers.

[10] Legassicke, Johnson & Gribbin, “Definitions of Child Recruitment and Use”, pp. 3–4.

[11] Mlambo, Mpanza & Mlambo, “Armed Conflict and the Increasing Use of Child Soldiers,” sec. 3.2.

erode protection structures[12]. Despite international frameworks—such as the Paris Principles and the UN Convention on the Rights of the Child—millions of children remain at risk, especially in regions with limited humanitarian access[13].

In this context, exploring new tools to prevent recruitment is both urgent and delicate. Among these, AI has emerged as a powerful but complex instrument[14]. When responsibly deployed, it may help actors identify risks early and reach children before harm occurs[15]. However, AI also brings risks of exclusion, misuse, and unintended harm—particularly in fragile and culturally diverse settings.

1.2. North Darfur as a Case Study: Conflict, Displacement, and Child Protection Risks

North Darfur presents a deeply layered protection crisis. Years of armed conflict, displacement, and economic collapse have dismantled formal child protection systems[16]. Today, the region is marked by alarming child protection trends: widespread recruitment by armed groups, early and forced marriages, increasing child labor, and the emergence of child-headed households following the deaths of parents[17].

Formal schooling has all but disappeared in several areas. In this vacuum, communities have built informal protection systems. Madrasa classes—religious schools traditionally focused on Quranic education—sometimes serve as safe spaces for children, offering both learning and a degree of protection from recruitment or exploitation. Community Protection-Based Networks (CBPNs)—including both women and men—monitor risks and relay concerns. Clan systems and host communities continue to support displaced households, often with minimal external support. Yet, these community-led mechanisms operate under immense strain.

The field realities of North Darfur underscore the stakes of digital experimentation. Technological solutions should navigate limited infrastructure, low smartphone penetration, and gendered access barriers. They should also align with customary norms that remain central to how communities resolve conflict and maintain social order.

1.3. Understanding AI: Technologies and Definitions

AI refers to a set of computational systems designed to perform tasks traditionally requiring human intelligence, such as reasoning, learning, or decision-making[18]. In the humanitarian sphere, relevant AI applications include predictive analytics, natural language processing, satellite image classification, and automated alert systems. For example, tools can be trained to detect early signs of recruitment activity

[12] Mlambo, Mpanza & Mlambo, *"Armed Conflict and the Increasing Use of Child Soldiers"*, p. 7.

[13] Yutaka Arai-Takahashi, "War Crimes relating to Child Soldiers and Other Children in Armed Conflict: An Incremental Step toward a Coherent Legal Framework?," *QIL Zoom-in* 60 (2019), pp. 28–32, https://www.qil-qdi.org/wp-content/uploads/2019/09/03_Child-Soldiers_ARAI_FIN-2.pdf.

[14] Beduschi, *"Harnessing the Potential of Artificial Intelligence for Humanitarian Action,"* pp. 1150–1152.

[15] Kreutzer et al., *"Ethical Implications Related to Processing of Personal Data and Artificial Intelligence in Humanitarian Crises,"* p. 12.

[16] Aaron Martin, Mathias Fiedler & Julie Guenat, "Digitisation and Sovereignty in Humanitarian Space: Technologies, Territories and Tensions," *Geopolitics* 28, no. 3 (2023), pp. 1362–1397, at 7–8, https://doi.org/10.1080/14650045.2022.2047468.

[17] Mlambo, Mpanza & Mlambo, *"Armed Conflict and the Increasing Use of Child Soldiers,"* p. 8.

[18] ICRC, *Handbook on Data Protection in Humanitarian Action*, 2nd ed. (Geneva: International Committee of the Red Cross, 2020), ch. 17, p. 274, https://www.icrc.org/en/data-protection-humanitarian-action-handbook.

by analyzing patterns in social media posts, movement data, or humanitarian needs assessment[19]. In practice, this could mean identifying online campaigns by armed groups targeting children, flagging unusual spikes in child movement towards recruitment hubs, or detecting sudden drops in school attendance recorded in needs assessments. AI-driven language models can also monitor local radio transcripts or messaging apps for recruitment slogans, while image recognition applied to satellite photos may reveal new encampments or training sites where children are present.

Chatbot-based systems can gather information from affected populations, while platforms like KoboToolbox support digital data collection even in low-connectivity environments. While these tools may offer new avenues for early warning, ethical dilemmas arise concerning consent, representation, and the risks of surveillance or bias embedded in AI design and deployment[20]. As AI is increasingly normalized in humanitarian practice, critical reflection is needed on what problems it is assumed to solve and for whom.

1.4. Rise of AI in Humanitarian Responses

Humanitarian actors have begun adopting AI technologies to enhance efficiency, target interventions, and inform rapid decision-making in crises[21]. From using satellite data to map displacement trends to deploying machine learning for food security forecasts, AI is reshaping humanitarian workflows. Child protection programming, however, remains at the early stages of integrating AI. Interest is growing in the use of AI for real-time risk monitoring, behavioral trend analysis, and location-based vulnerability mapping. Yet, such tools may also be co-opted for securitized purposes, especially in regions with weak data protections or authoritarian governance. In Sudan and similar contexts, the introduction of predictive models—without safeguards for transparency or community oversight—risks exacerbating the very harms humanitarian action seeks to prevent. As the humanitarian sector experiments with AI, a robust ethical framework is necessary to balance innovation with accountability.

These trends make a robust ethical framework indispensable for any AI use in child protection (see Section 2)

1.5. Research Questions and Scope

This paper seeks to understand how AI can be ethically used to prevent child recruitment in conflict settings, with North Darfur as a reference case. It asks:

- How can AI be ethically applied to support the prevention of child recruitment in conflict-affected areas like North Darfur?
- What are the potential benefits of AI in enhancing early warning systems, community-based protection, and humanitarian access?
- How can AI be ethically applied to support the prevention of child recruitment in conflict-affected areas like North Darfur?

[19] Zara Rahman & Julia Keseru, *Predictive Analytics for Children: An Assessment of Ethical Considerations, Risks, and Benefits* (Innocenti Working Paper 2021-08, UNICEF Office of Research – Innocenti, 2021), pp. 27–28 (lists data sources incl. social media and models for forecasting movement/risks) and pp. 15–16 (population-based needs assessment, incl. mobility tracking of displaced populations), https://www.unicef.org/innocenti/media/5161/file/UNICEF-Predictive-Analytics-Working-Paper-2021.pdf.

[20] Tino Kreutzer, Karolina MacLachlan, Lars Bromley & Kristin Bergtora Sandvik, "Ethical Implications Related to Processing of Personal Data and Artificial Intelligence in Humanitarian Crises: A Scoping Review," *BMC Medical Ethics* 26, no. 49 (2025), pp. 1–15, https://doi.org/10.1186/s12910-025-01189-2.

[21] Beduschi, *"Harnessing the Potential of Artificial Intelligence for Humanitarian Action,"* pp. 1152–1153.

- What are the potential benefits of AI in enhancing early warning systems, community-based protection, and humanitarian access?

1.6. Methodological Approach

The analysis draws on mixed qualitative methods, combining prior fieldwork in North Darfur (focused on the reintegration of CAAFAG) with secondary literature, global case studies on AI in humanitarian contexts, and insights from expert interviews. The methodology is guided by a protection-first lens, acknowledging the complexity of verifying child recruitment risks in volatile environments. Limitations include the absence of large-scale datasets from conflict zones like North Darfur and the ethical challenges of relying on algorithmically filtered information. Special attention is paid to the influence of training data and design logic in shaping AI outcomes—underscoring that technical systems cannot be divorced from the sociopolitical contexts in which they operate.

This methodology informs the analysis of opportunities (Section 3) and pitfalls (Section 4), and prepares the case focus on North Darfur (Section 5), leading to the design framework (Section 6)

2. Conceptual Framework: AI Ethics and Humanitarian Practice

This section addresses the first research question: How can AI be ethically applied to support the prevention of child recruitment in conflict-affected areas such as North Darfur? Ethical integration of AI into child protection programming requires more than technical expertise, demands a principles- and context-aware approach grounded in humanitarian values. The following subsections outline the conceptual foundations of the paper, combining humanitarian principles, feminist and community-centered perspectives, rights-based and culturally sensitive design, binding legal frameworks, and the clear delineation of human versus machine responsibility. Together, these elements form the ethical guardrails within which AI should operate in conflict-affected settings.

2.1. Humanitarian Principles and Do No Harm

The application of AI in humanitarian operations should uphold the principles of humanity, neutrality, impartiality, and independence. Humanity means that AI systems should ultimately serve the goal of alleviating suffering, such as identifying risks of child recruitment early so that protective action can be taken before harm occurs. Also, neutrality requires that AI tools avoid reinforcing political or military agendas; for instance, predictive models should not be used by state or armed actors to securitize children under the guise of protection. While impartiality demands that AI-supported interventions prioritize children most at risk, Independence requires that AI-driven analysis remain autonomous from donor or government influence; so that child protection decisions are based on humanitarian needs rather than political priorities.

In child protection, these principles require that interventions minimize harm, avoid exacerbating vulnerabilities, and remain sensitive to the complex dynamics of conflict. The "Do No Harm" approach becomes especially critical when predictive models are used to prioritize high-risk areas. Algorithmic predictions can influence resource allocation, but without context-specific oversight, they risk overlooking less visible yet equally vulnerable populations[22].

In volatile contexts such as North Darfur, prioritizing one community over another based on incomplete or biased datasets could unintentionally heighten protection risks. AI should therefore be an analytical support tool, not a replacement for the nuanced judgment of humanitarian actors. AI systems are not neutral; every algorithm reflects the values, priorities, and blind spots of its designers. Recognizing this

[22] Kreutzer et al., "Ethical implications related to processing of personal data and artificial intelligence in humanitarian crises: a scoping review," p. 12.

value-laden nature is essential from the outset, as bias is not simply an unfortunate byproduct but a predictable consequence of development choices[23]. This reality reinforces the moral responsibility to interrogate design decisions before deployment.

Therefore, upholding these principles is not only about preventing harm but also about who is involved in shaping AI tools. This leads directly to the importance of integrating feminist and community-centered perspectives.

While humanitarian principles such as impartiality and 'do no harm' set the ethical foundation, they do not explain how affected communities themselves shape and guide interventions. This makes it necessary to turn to feminist and community-centred approaches, which bring those perspectives to the forefront.

2.2. Feminist and Community-Centered Approaches

Feminist and community-centered approaches position affected populations –particularly women-led networks and grassroots structures—as co-creators of both programme design and technological tools. Their involvement ensures that AI systems are informed by local knowledge and capable of capturing subtle, culturally specific indicators of recruitment risk.

For example, women-led early warning groups in Sudan may detect shifts in daily routines, community gatherings, or language use that signal recruitment pressure—insights often absent from formal datasets. Embedding such knowledge into AI design not only strengthens accuracy but also reinforces community ownership protection strategies.

Involving communities in co-creation directly connects to the next consideration: ensuring that AI systems respect rights and remain culturally sensitive to the contexts in which they operate. Inclusivity on its own, however, is not enough. To ensure these approaches translate into practice, AI systems should be designed in ways that respect rights and adapt meaningfully to diverse cultural contexts.

2.3. Rights-Based and Culturally Sensitive Design

Rights-based design embeds the protection of fundamental rights into every stage of AI development, ensuring that technology upholds — rather than undermines — children's rights under international law. In practice, this means actively preventing discriminatory outcomes by recognising and correcting structural biases present in training datasets, which may reflect gender inequality, racial discrimination, or colonial legacies.

Culturally sensitive design further ensures that risk assessment models align with local norms and values. In some cases, introducing what can be termed "desirable bias" — such as giving greater algorithmic weight to alerts from trusted community actors, or discounting intelligence from militarised or politically influenced sources — can help safeguard ethical outcomes. By intentionally embedding humanitarian and child-protection values into AI systems, practitioners can create tools that are both context-aware and ethically robust.

While rights-based and culturally sensitive design focus on the how of development, they should also operate within clearly defined legal boundaries that guide responsible AI use in humanitarian settings.

Placing rights and cultural sensitivity at the centre also requires grounding in law. International humanitarian and human rights law, along with emerging tech norms, provide the necessary legal foundations for ethical design.

[23] Beduschi, *"Harnessing the Potential of Artificial Intelligence for Humanitarian Action,"* pp. 1156–1157.

2.4. Legal Foundations: CRC, IHL, Paris Principles, Tech Norms

International legal frameworks such as the CRC, IHL, and Paris Principles establish binding obligations to protect children affected by armed conflict[24]. AI ethics norms, including the OECD Principles on AI and UNESCO's Recommendation on the Ethics of AI, provide complementary guidance for responsible technology use in humanitarian contexts[25].

Within these frameworks, four core principles (legality, necessity, proportionality, and non-discrimination) are especially relevant to AI systems to comply with existing child protection laws[26], avoiding practices that might infringe on rights under the CRC or IHL. Closely linked is the principle of necessity, which restricts the deployment of AI to situations where it is essential for child protection, such as detecting recruitment patterns, rather than for broader or experimental uses. Proportionality then serves as a safeguard, requiring that any intrusion into children's lives, such as the monitoring of digital communications or movement patterns, should be strictly balanced against the risks to their privacy and dignity. Finally, non-discrimination reinforces that AI should be designed and applied in ways that do not reproduce social bias; in practice, this means ensuring vulnerable groups like girls, who are often recruited for domestic or sexual exploitation, are not overlooked simply because they are less visible in datasets.

Taken together, these principles form a coherent legal and ethical baseline; they not only guide how AI should be designed and deployed but also highlight the continuing importance of human oversight. While AI can assist in risk identification and prevention, ultimate accountability under the CRC and IHL remains with human actors. Humanitarian agencies should therefore ensure that AI tools strengthen, rather than weaken existing obligations to protect children in conflict settings.

This legal requirement for accountable human oversight leads to the final element of the framework: clarifying the boundaries between human and machine responsibility. Even with strong legal frameworks, a final challenge remains: deciding how much responsibility rests with humans versus machines when decisions affect children's lives. This balance is the focus of the next subsection.

2.5. Human vs Machine Responsibility

AI can process vast datasets, identify patterns, and predict potential recruitment risks, but it cannot bear ethical or legal responsibility. In humanitarian contexts, the principle of "human-in-the-loop" is essential: human actors should interpret AI outputs, validate them against contextual realities, and make final decisions.

Clear chains of accountability ensure that when harm occurs—such as wrongful risk categorisation or neglect of certain communities—responsibility can be traced to decision-makers, not to the technology itself[27]. Delegating protection judgments to algorithms without oversight risks eroding humanitarian accountability and masking political inaction or resource constraints.

In North Darfur, informal community protection systems—often led by women or youth-demonstrate that meaningful protection is relational, not computational. By this, it means that protection depends on trust, dialogue, and lived social bonds that grow from cultural knowledge and shared experiences. These forms of protection cannot be captured or reproduced by algorithms, because they rely on human relationships and community trust. While AI can assist by highlighting potential risks, it cannot replace

[24] ICRC Casebook, *"Child Soldiers,"* https://casebook.icrc.org/a_to_z/glossary/child-soldiers.

[25] Pizzi, Romanoff & Engelhardt, *"AI for Humanitarian Action: Human Rights and Ethics,"* passim.

[26] Marelli, *Handbook on Data Protection in Humanitarian Action*, ch. 17.

[27] Kreutzer et al., *"Ethical implications related to processing of personal data and artificial intelligence in humanitarian crises: a scoping review,"* p. 12.

the human accountability and cultural understanding that form the foundation of effective child protection.

Ethical AI should therefore complement, not replace, the nuanced judgment, trust-building, and culturally grounded decision-making of humanitarian actors. A participatory and transparent governance structure is essential to ensure that machine-generated insights are always filtered through the lens of human rights, contextual understanding, and community consent.

By grounding AI use in humanitarian principles, community participation, rights-based design, legal obligations, and human accountability, this framework offers a safeguard against the risks of technological determinism. It ensures that AI tools serve as enablers—rather than substitutes—for the deeply human work of protecting children in conflict zones.

3. Opportunities: The Potential Role of AI in Preventing Child Recruitment

While Section 2 explored the ethical foundations of AI in humanitarian practice, this section addresses the second research question: What potential roles can AI play in preventing child recruitment in conflict zones? The focus is on how AI can enhance early detection, risk monitoring, education, and psychosocial support, as well as how local actors can shape the technology to align with community needs. These opportunities are not theoretical; they build on both existing humanitarian practices and emerging innovations that, if designed with care, could strengthen child protection efforts in volatile environments such as North Darfur.

3.1. Early Warning Systems and Predictive Tools

Early warning system (EWS) are vital for identifying emerging risks of child recruitment and enabling timely, preventive action. In North Darfur, where state-led protection structures have collapsed, EWS should adapted to low-connectivity, resource-constrained settings. Offline-first tools such as KoboToolbox allow community-based networks to collect data without internet access and later sync it for analysis[28]. Their adaptability to protection-specific questionnaires makes them well-suited for rural and displacement contexts.

Yet predictive models cannot operate in a vacuum. They should be built with desirable bias—intentionally embedding humanitarian and child-protection priorities—to detect both visible threats and hidden vulnerabilities, such as the risks faced by girls in secluded households. Without context-specific oversight, algorithmic predictions risk prioritizing certain communities over others, especially when data is incomplete or biased. In volatile contexts like North Darfur, this could heighten protection risks. AI should therefore remain an analytical support tool, never a substitute for the nuanced judgment of humanitarian actors.

Feasibility also matters. Chatbot-based check-ins via SMS or social media, while promising for real-time distress detection, are hindered by Darfur's unstable connectivity. Even where satellite internet, such as Starlink, is available, it is often limited to humanitarian offices or wealthier households, excluding the most vulnerable. Predictive tools should therefore be grounded in community-owned systems that combine offline functionality with strong local validation.

[28] KoboToolbox, *"About KoboToolbox"* (2025), https://www.kobotoolbox.org.

3.2. Monitoring Risks in Displacement Camps

Displacement camps across North Darfur remain high-risk environments for child recruitment, particularly given the collapse of formal education systems, increasing child-headed households, and pervasive gender-based violence. In such fragile settings, monitoring child protection risks is challenging due to mobility constraints, safety concerns, and communication barriers. Nonetheless, humanitarian actors, especially through Community-Based Protection Networks (CBPNs), have developed low-tech but effective methods such as entry and exit interviews, focus group discussions (FGDs), post-distribution monitoring (PDM), and complaint mechanisms[29].

AI can enhance these systems by enabling hybrid monitoring models that merge human intelligence with technological tools. For instance, satellite imagery cross-checked with CBPN reports can help confirm patterns such as sudden displacement, school closures, or movements of armed groups. However, any surveillance-enabled monitoring should adhere strictly to "Do No Harm" safeguards—protecting children's privacy, avoiding unnecessary exposure, and ensuring informed community participation. Interpreting AI-generated signals without local context risks false positives that could trigger panic, mistrust, or unintended harm. For example, an algorithm might misinterpret a large community gathering—such as a food distribution or religious event—as evidence of forced recruitment. If acted upon without verification, this could spread fear among families or erode trust between the community and humanitarian actors. Properly integrated, AI can sharpen situational awareness without undermining the trust that makes community-led protection work.

3.3. AI in Education and Psychosocial Support Tools

The erosion of formal schooling in conflict zones like North Darfur leaves thousands of children idle and at risk of exploitation. AI-powered educational platforms, particularly those that function offline and support voice-based learning in local dialects, can help fill this gap. Adaptive learning systems enable children to progress at their own pace, while AI-assisted chatbots can provide psychosocial check-ins where trained mental health professionals are unavailable.

However, these solutions should be designed with cultural sensitivity and gender-aware access strategies to prevent reinforcing digital inequalities. Another challenge is health literacy—the ability of children and caregivers to understand and use the information provided by AI chatbots or digital platforms. Without adequate literacy support, psychosocial tools may be misunderstood, misapplied, or even cause distress instead of relief.

Field interviews reveal that digital access is often skewed toward men, potentially excluding girls from benefiting equally. Embedding these tools within safe, community-controlled spaces—such as madrasa classrooms or child-friendly corners—can ensure inclusive participation and safeguard confidentiality.

By involving communities from the design stage, AI-based education and psychosocial tools become more than imported digital fixes; they evolve into trusted, locally adapted resources that strengthen resilience and protect children from recruitment pressures. Here, as in other domains, the goal is to augment human relationships, not replace them, ensuring that technology supports community structures can protect children from being recruited.

[29] UNHCR, *Understanding Community-Based Protection* (Geneva: UNHCR, 2024), https://emergency.unhcr.org/sites/default/files/2024-01/UNHCR%2C%20Understanding%20Community%20Based%20Protection.pdf.

3.4. The Role of Local Actors in Shaping AI Design

AI systems introduced in humanitarian settings carry deep implications—not only for service delivery, but for power, representation, and trust. In conflict-affected contexts like North Darfur, the involvement of local actors in shaping AI design is not a luxury—it is a necessity. Yet, as with many forms of innovation, there is a risk that new technologies will replicate the top-down dynamics of older humanitarian models, sidelining the very communities they aim to protect.

Customary authorities, women-led networks, and community protection groups already play a central role in identifying threats, responding to displacement, and guiding local prevention efforts. In North Darfur, informal checkpoints, adjusted routines, and community-selected focal points—such as the CBPNs—are practical illustrations of local systems actively managing risk. These actors possess critical contextual intelligence: they understand when, how, and why children become vulnerable to recruitment, and which coping strategies are trusted or resisted.

If AI tools are to be ethically deployed in such settings, they should be shaped by community knowledge and protection logic. This means involving local actors at multiple stages—from defining the problem, to selecting data inputs, to reviewing how predictions or alerts are interpreted. Without this, AI systems risk embedding biases, misreading intent, or misdirecting resources. In Darfur, where sensitivities around surveillance and power are acute, community governance over AI inputs and outputs is essential for safeguarding humanitarian space and social cohesion.

However, meaningful participation is not automatic. It requires time, dialogue, and investment in capacity strengthening, particularly for groups that are often excluded from digital spaces—such as women, youth and rural elders. Gender barriers in tech use, as reported during fieldwork observations (2023-2024), underline the urgency of inclusive design methodologies. Participatory models should also navigate security risks: communities may be reluctant to share data about armed actors, and ethical safeguards should prevent any misuse or forced disclosure.

Ultimately, the legitimacy of AI in child protection will depend on whether it is seen as an ally to local resilience, not a replacement. Humanitarian agencies should ensure that AI tools amplify—not override—the wisdom, agency, and decision-making of affected communities. In a context like North Darfur, where much of the protection architecture is informal and trust-based, this is not only an ethical imperative—it is a strategic one.

To sum up, the potential of AI to prevent child recruitment lies not in the technology alone, but in how it is designed, adapted, and governed. From offline early warning systems to hybrid monitoring models, from low-bandwidth educational tools to participatory design, each opportunity depends on marrying technological capacity with community ownership. This alignment can turn AI into a genuine force multiplier for child protection. Yet, as the next section will explore, these same opportunities carry inherent technical risks and structural challenges that demand equal attention.

4. Pitfalls: Ethical Risks and Structural Challenges

Building on the ethical foundations outlined in Section 2 and the opportunities identified in Section 3, this section addresses the other side of the equation: What ethical risks and structural challenges accompany the use of AI in efforts to prevent child recruitment? While AI offers significant promise in preventing child recruitment, its deployment in conflict-affected humanitarian contexts is fraught with pitfalls that are not merely technical but also moral, legal, and political.

Here, three interconnected challenges are examined: bias in datasets, value misalignment, and absence of transparency and community participation. They were chosen because they consistently emerge as the most significant cross-cutting risks in both emerging literature and humanitarian practice. Bias in datasets often leads to inequitable outcomes against marginalized groups[30]; value misalignment arises when AI systems fail to align with local cultural or ethical frameworks. These risks are especially pronounced in contexts like North Darfur, where digital divides, prolonged conflict, and fragile governance create high potential for harm.

4.1. Bias in Datasets: Representations, Generalizability, Harm

AI tools in humanitarian protection are never built on entirely neutral datasets. From data collection to model training, these systems carry the imprint of the values and structures of the societies that produced them. As Kreutzer et al. and UNICEF's AI for Children guidance note, even seemingly objective datasets can reflect structural inequalities rooted in colonial histories, racial hierarchies, or outdated assumptions about vulnerability[31][32].

One of the most consequential forms of bias emerges when models trained in one context are deployed in another without proper recalibration. For example, an algorithm developed elsewhere might misinterpret behaviors in North Darfur—such as school absenteeism or loitering—not as indicators of poverty or gendered labor expectations, but as recruitment risk signals. Likewise, algorithms trained on social media or non-governmental organization (NGO) reports may amplify the narratives of dominant actors while obscuring the lived realities of marginalized children. This can concentrate interventions on certain ethnic or geographic groups, perpetuating discrimination in ways reminiscent of "smart" surveillance systems disproportionately targeting racial minorities elsewhere.

Bias here is not just a statistical imperfection—it encodes human assumptions into technology, often invisibly. In humanitarian contexts, those hidden distortions can produce decisions with life-or-death consequences, particularly when they reinforce gender stereotypes or structural discrimination in identifying children "at risk" of recruitment.

4.2. Value Misalignment: Cultural Insensitivity and Perverse Outcomes

Even where data quality is high, AI systems can still cause harm if their objectives are not aligned with humanitarian and child-protection values. This "value alignment problem" reflects the reality that smarter systems are not automatically more ethical. A model optimised to reduce "risk" may overemphasize militarized indicators while neglecting local protection norms such as community mediation or traditional rites of passage[33].

In practice, this misalignment can erase the experiences of children whose recruitment is informal, sexualized, and undocumented—particularly girls engaged as porters, cooks, or "wives," as highlighted in UNICEF report[34]. When these experiences are excluded from formal definitions of recruitment, they risk being excluded from algorithmic detection as well, embedding gender bias into the system's core logic.

[30] Kreutzer et al., *"Ethical implications related to processing of personal data and artificial intelligence in humanitarian crises: a scoping review,"* pp. 9–10.

[31] UNICEF, *Policy Guidance on AI for Children, Version 2.0*, pp. 11–14.

[32] *Kreutzer et al., "Ethical implications related to processing of personal data and artificial intelligence in humanitarian crises: a scoping review,"* pp. 9-15.

[33] Berditchevskaia, Malliaraki & Peach, *Participatory AI for Humanitarian Innovation*, 2021, pp. 19–20

[34] UNICEF, *25 Years of Children and Armed Conflict: Taking Action to Protect Children in War* (New York: UNICEF, 2022), pp. 22–23, https://www.unicef.org/sites/default/files/2022-06/UNICEF-25-years-children-armed-conflict.pdf.

Value drift can worsen over time as machine learning models adapt to new data without adequate oversight. Initial safeguards—such as prioritizing women-led community signals or rejecting facial recognition—can erode, especially where technical teams and field actors operate in isolation. Without explicit "desirable bias" built in from the start, systems may misclassify acts of survival, such as seeking food or shelter, as indicators of recruitment intent.

Structural inequalities compound the problem. Unequal access to devices, connectivity costs, and entrenched gender gaps in digital literacy mean that a system designed for one group may inadvertently exclude others. The psychosocial impacts—heightened anxiety, loss of agency, or dependency—can be especially severe for children in unstable environments[35].

4.3. Lack of Transparency and Community Participation

Opacity[36] is another recurring failure. Many humanitarian AI tools function as "black boxes," providing little to no explanation of how inputs lead to outputs. When a child is flagged—or excluded—from protection services, neither they nor frontline staff may understand why, making wrongful decisions difficult to contest and undermining trust in humanitarian actors.

Geographic and cultural distance often deepens this opacity. Developers, frequently based in the Global North, may never witness the consequences of a false positive in South Sudan or a missed warning in the field. Therefore, the outsourcing ethical judgment to machines risks stripping decisions of the empathy, cultural understanding, and moral attention they require.

Participation is too often missing from the design process. AI tools are sometimes deployed without engaging local leaders, caregivers, or children themselves. The Nesta *Participatory AI* framework emphasizes that involvement should go beyond occasional feedback to genuine co-creation—embedding local definitions of risk, safety, and vulnerability directly into the system's architecture[37]. In North Darfur, where community protection often relies on oral customary law and informal networks, exclusion from design is more than a technical oversight—it is an ethical breach.

Accountability further complicates matters. When harm occurs, responsibility is diffuse: is it the field officer, the developer, the donor, or all of the above? Without clear appeal mechanisms and shared accountability, AI risks replicating the impunity it was meant to challenge.

Finally, there is the threat of surveillance creep. Tools introduced for protection may be repurposed by armed groups or authoritarian regimes, blurring the line between humanitarian monitoring and military intelligence. As Beduschi cautions, data collection itself can become a liability—especially when tied to GPS, biometrics, or behavioral profiling. In volatile conflict zones, such capabilities can fuel behavioral control, fear, and self-censorship, eroding trust and discouraging participation. Balancing the imperative to protect children with the equally vital duty to protect their privacy remains one of the most pressing challenges in humanitarian AI governance.

While these pitfalls are evident across global humanitarian contexts, they become particularly acute in Darfur, where governance is weak and digital infrastructures are fragmented. In North Darfur, such challenges intersect with political instability, mass displacement, and deeply rooted cultural protection norms in ways that will be examined in detail in the next section.

[35] Kreutzer et al., "Ethical implications related to processing of personal data and artificial intelligence in humanitarian crises: a scoping review," p. 11.

[36] Opacity refers to the lack of clarity or explainability in how AI systems operate.

[37] Aleks Berditchevskaia, Kathy Peach & Eirini Malliaraki, *Participatory AI for Humanitarian Innovation: A Briefing Paper* (London: Nesta, Sept. 2021), pp. 12–14 & 16, https://media.nesta.org.uk/documents/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf.

These structural challenges highlight that the ethical risks of AI are not abstract but deeply embedded in local realities. To illustrate this, the next section turns to North Darfur as a context where overlapping vulnerabilities and community-led practices test the feasibility of AI-based prevention tools.

5. North Darfur Case Focus: Local Realities and Ethical Dilemmas:

North Darfur presents a complex landscape of overlapping child protection risks, weakened formal systems, and emerging informal responses—all of which provide a grounded lens to examine the ethical feasibility of AI-based prevention tools. While this section is not presented as a full empirical case study, it illustrates how conflict, displacement, and digital exclusion complicate the ethical deployment of AI-based prevention tools. These realities test whether the principles discussed earlier—humanitarian, rights-based, and community-centred—can be meaningfully upheld when technology meets fragile protection systems.

Moreover, the exclusion of women from digital decision-making in North Darfur creates structural blind spots in humanitarian tech deployment. For example, tools relying on WhatsApp are accessible mainly to men and NGO staff, leaving women and other vulnerable groups unable to participate in risk reporting or benefit from early warning systems.

5.1. Protection Challenges and Vulnerabilities

The protection crisis in North Darfur is severe and multifaceted. Armed conflict has disrupted education, disintegrated family units, and eroded public services. Formal schooling systems have collapsed in many areas, leaving children idle and vulnerable to exploitation. Early and forced marriages, child labour, and recruitment into armed groups are on the rise. A particularly distressing trend is the emergence of child-headed households, with older children caring for younger siblings after losing their parents. Gender-based violence (GBV) remains widespread, particularly targeting women and girls in displaced communities, with minimal access to safe reporting channels or psychosocial support.

These dynamics underscore how structural collapse, rather than isolated incidents, renders children increasingly susceptible to recruitment. Any prevention entity—including those using digital tools—should account for the social and economic pressures driving these risks. It is within this context of institutional breakdown that community-driven protection responses, as discussed earlier, become the primary and often only functional safety net.

In response to these vulnerabilities, local communities have developed their own protective strategies that function in the absence of formal systems.

5.2. Community-Based Projection Mechanisms

In the absence of formal child protection programs, informal and community-led systems have emerged to provide basic safety and social cohesion. Madrasa classes occasionally serve as safe spaces for children, while local leaders and watch groups monitor the movement of armed actors. Communities adapt their daily routines—such as altering times for firewood collection or farming—to avoid danger zones. Informal checkpoints help regulate mobility and enhance security.

Extended family networks and host communities play a significant role in absorbing displaced persons and supporting vulnerable households. These practices are often organized by the CBPNs, locally selected and trained structures that form the backbone of community surveillance and reporting mechanisms.

Despite their unstructured nature, these systems offer culturally grounded, real-time responses to child protection threats. Their functioning rests on long-standing social norms and trust—bridging directly to the role of customary law in shaping both community behaviour and the potential reception of new technologies.

However, these community mechanisms do not exist in isolation—they are guided and constrained by customary norms and oral law, which continue to shape local decision-making and perceptions of legitimacy.

5.3. Customary Norms and the Role of Oral Law

One of the most significant yet underexamined elements of protection in North Darfur is the role of customary law. Passed down orally across generations, customary norms regulate community behavior and conflict resolution. While largely undocumented, they serve as the de facto legal and ethical framework in many communities. This poses a challenge for any digital tool designed without deep local consultation. Predictive models or automated alerts that fail to engage with these cultural norms risk undermining local legitimacy or triggering unintended consequences. For instance, customary law often assigns elders the responsibility of mediating disputes among youth to prevent escalation. If an AI system were to misclassify such gatherings as signs of recruitment activity, it could delegitimize the role of community mediators and erode trust in both local protection systems and humanitarian actors. Moreover, as communication increasingly depends on digital infrastructure, the reach and influence of customary norms are themselves being shaped, and sometimes limited, by patterns of digital access. Yet even customary systems are increasingly influenced by shifts in digital access, which introduces new forms of inequality and exclusion.

5.4. Digital Infrastructure and the Gendered Digital Divide

Digital access in North Darfur is sharply uneven. While mobile communication is largely nonexistent in some areas, satellite-based internet (e.g., via Starlink) has improved connectivity. Cyber cafes are the primary access point for internet use, particularly WhatsApp and Facebook. However, access remains skewed toward INGO staff, businesspeople, and men with financial means.

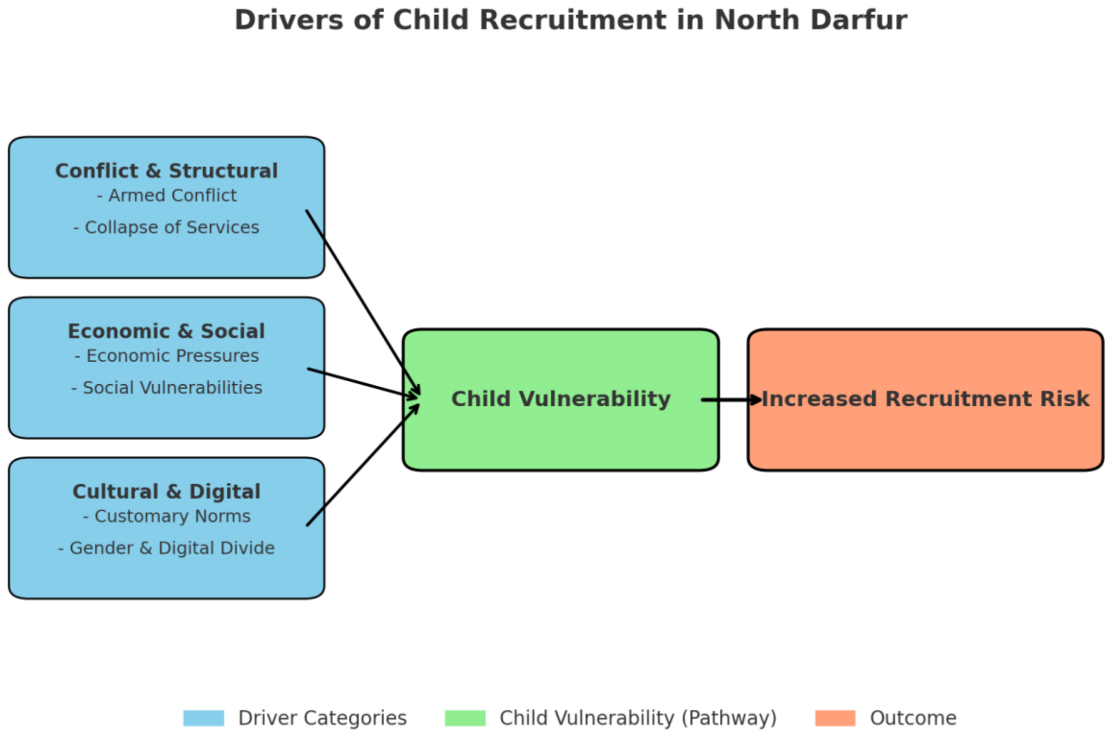
This digital divide reinforces existing inequalities. Women, lower-income families, and rural youth are often excluded from digital spaces[38]. Any tech-based intervention, especially those designed to identify or prevent child recruitment, must consider the risk of further marginalizing the very populations it intends to protect.

Across these dimensions—protection vulnerabilities, community-based mechanisms, customary norms, and the gendered digital divide—North Darfur reflects the tension between the promise of digital innovation and the risks of cultural, structural, and ethical misalignment. The existing resilience of community-led protection networks shows that external technologies should adapt to, rather than overwrite, local systems. This reality reinforces the argument that ethical AI should be built through participatory pathways, with safeguards tailored to local power dynamics, cultural legitimacy, and gender equity.

Taken together, these dynamics show that AI interventions cannot be divorced from the cultural, structural, and digital contexts in which they operate. Section 6, therefore, builds on these insights to propose a framework for embedding safeguards into AI design, deployment, and governance in humanitarian settings. This underscores the central question*: can AI tools meaningfully complement community resilience while safeguarding against deepening inequalities?*

[38] GSMA, Ipsos & Basis, *The Mobile Gender Gap Report 2022* (London: GSMA, 2022), p. 6-7, https://www.gsma.com/r/wp-content/uploads/2022/06/The-Mobile-Gender-Gap-Report-2022.pdf.

Section 6 directly addresses this question by proposing a framework for embedding safeguards into the design, deployment, and governance of AI tools in humanitarian contexts.



6. Ethical AI by Design: Principles, Risks, and Participatory Pathways

Technology alone cannot protect children in conflict zones—it should be designed to serve, not replace, the systems of care and protection that already exist. In places like North Darfur, where formal child protection structures have collapsed but community-led safety networks remain active[39], the challenge is not simply to "introduce AI" but to embed it into a model that respects rights, reflects local realities, and strengthens what is already working. This section outlines a framework for Ethical AI by Design—a community-led, rights-based model that integrates humanitarian principles, child protection standards, participatory governance, and robust legal safeguards into every stage of AI development.

6.1. Embedding HRBA and Child Protection Principles into AI Development

The first safeguard is to move beyond the assumption that AI systems are neutral. As emphasized earlier, all AI tools are value-laden, whether intentionally or not. Designing humanitarian AI, therefore, requires an explicit choice to embed desirable bias—that is, the deliberate prioritization of humanitarian principles, children's rights, and the "do no harm" imperative. For instance, predictive systems used in displacement camps should not only optimize efficiency but also actively minimize risks of stigmatization or surveillance abuse. By treating humanitarian law and child protection frameworks as default values, AI design can align with international norms while remaining context-sensitive.

[39] Safa Yagoub, *Beyond the Reintegration: The Role of Women in Preventing CAAFAG in North Darfur*, *Journal of Social and Political Sciences* 8, no. 3 (2025), pp. 177–191.

6.2. Participatory Governance and Algorithmic Transparency

Participatory governance ensures that AI systems are not only accountable but also reflective of community realities. In North Darfur, decision-making on protection measures is often dominated by men, while women—despite playing critical roles in care and informal protection—are frequently excluded from formal discussions. AI governance models should actively correct such imbalances by creating structured opportunities for women, youth, and marginalized groups to shape priorities, consent to data collection, and influence system design.

Transparency in how algorithms process data is equally important. In contexts where sharing information about military or governmental actors is culturally sensitive, communities should be able to see and understand how information will be used, and under what safeguards, before they agree to participate.

True participation is more than consultation; it requires shared ownership of AI design. Feminist and community-based approaches demand that local actors, including women and youth, are not merely invited to comment but are positioned as co-designers and decision-makers. This inclusion helps ensure that AI systems reflect lived realities rather than external assumptions.

6.3. Transparency, Accountability, and Legal Safeguards

A further dimension of ethical AI by design is transparency. Algorithms used in humanitarian settings should be explainable to both practitioners and affected communities, especially when outcomes influence life-saving interventions. Accountability mechanisms should therefore be dual: internal (through humanitarian agencies) and external (through independent review bodies). Embedding HRBA provides a concrete pathway to achieve this. It demands that humanitarian AI is non-discriminatory, participatory, and subject to clear lines of responsibility. Importantly, legal safeguards should align with the CRC, IHL, and existing humanitarian technology norms, ensuring that AI remains bound to the same principles as human actors[40].

6.4. From Safeguards to Action: Policy Recommendations

To move from principle to practice, humanitarian actors and policymakers should:

- Mandate humanitarian values in design – Embed child protection norms, humanitarian principles, and "do no harm" directly into algorithms as default parameters.
- Guarantee meaningful participation – Institutionalize consultation with children, caregivers, and local actors in all stages of AI development and deployment.
- Align with legal frameworks – Require compliance with CRC, IHL, and data protection standards as non-negotiable conditions for AI deployment. In practice, compliance with the CRC should ensure that AI systems uphold the child's best interests (Article 3), protect children from recruitment or exploitation (Article 38), and safeguard their right to privacy (Article 16). Alignment with IHL entails that AI tools should never contribute to the targeting of civilians or the militarization of child protection data and should instead reinforce protections afforded to children in armed conflict.
- Combine human and machine responsibility – Ensure human oversight remains central in decision-making, particularly for interventions affecting children's safety.
- Introduce sunset clauses and review mechanisms – All AI systems deployed in humanitarian crises should be temporary, subject to regular review, and dismantled once risks outweigh benefits.

[40] UNICEF & Ministry for Foreign Affairs of Finland, *Policy Guidance on AI for Children, Version 2.0* (New York: UNICEF, 2021), pp. 8–9, 11–12, https://www.unicef.org/innocenti/media/1341/file/UNICEF-Global-Insight-policy-guidance-AI-children-2.0-2021.pdf.

By embedding these safeguards, AI tools can move from abstract innovation to tangible, trusted mechanisms for protecting children in high-risk environments. This rights-based, community-led design is not only ethically sound but also operationally feasible—offering a pathway for technology to work with, rather than against, the social fabric that sustains child protection in conflict zones.

Safeguard	Key Lesson for Humanitarian Actors
Mandate humanitarian values in design	AI should embed child protection norms and "Do No Harm" principles by default, not as afterthoughts.
Guarantee meaningful participation	Children, caregivers, and local actors should be involved in co-design to ensure cultural legitimacy.
Align with legal frameworks	Compliance with CRC, IHL, and data protection standards is non-negotiable and requires practical checks.
Combine human and machine responsibility	AI can support analysis, but accountability should remain with human decision-makers.
Introduce sunset clauses and review mechanisms	AI tools in crisis settings should be temporary, regularly reviewed, and dismantled if risks outweigh benefits.

Key lessons for ensuring AI supports child protection in humanitarian contexts.

7. Conclusion

7.1. Summary of Findings

This paper has examined both the opportunities and risks of deploying AI in humanitarian contexts to prevent child recruitment in armed conflict. The analysis highlighted how AI tools can strengthen early warning, monitoring, and protective measures (Section 3), but also underscored the ethical risks of bias, cultural misalignment, and lack of transparency (Section 4). The North Darfur case study (Section 5) demonstrated the urgency of addressing these dilemmas in fragile contexts, where both risks and needs are acute. Section 6 then outlined a framework for embedding ethical AI by design, emphasizing participation, rights-based safeguards, and alignment with humanitarian principles.

7.2. Responsible Innovation and Community Agency

The path forward is responsible innovation: designing AI tools that intentionally embed humanitarian principles, child protection standards, and cultural sensitivity into their architecture and governance. This includes recognizing that AI cannot be neutral, and its datasets and algorithms reflect human values. The challenge is to channel this "desirable bias" toward humanitarian aims, ensuring tools actively promote inclusion, equity, and protection.

In North Darfur, this requires engaging with the very structures that have sustained child protection in the absence of formal programming: madrasa networks, community watch groups, extended family systems, and customary law. Women, youth, and other at-risk groups should be co-creators in the process, shaping how tools are designed, deployed, and adapted.

Responsible innovation also demands confronting the structural barriers identified in Section 4—dataset bias, cultural misalignment, and exclusion from decision-making—while embedding safeguards from the outset. This is not merely technical; it is a political and ethical commitment to uphold dignity, rights, and trust. Transparency, informed consent, and clear communication about how AI tools operate are essential for legitimacy and sustained use.

Finally, every deployment should be guided by built-in governance mechanisms, including sunset clauses and periodic lifecycle reviews, so that tools evolve with context or are retired once they no longer serve the community's best interests.

7.3. Call for Cautious Optimism

The North Darfur case offers a balanced lesson: digital inequities, protection risks, and cultural complexities set real limits on AI's applicability, but they do not erase its potential. As Section 5 showed, even in contexts with minimal connectivity, hybrid approaches—combining low-tech community monitoring with selective, secure digital tools—can enhance situational awareness and enable earlier, more targeted interventions.

This is a space for cautious optimism. AI, developed through a rights-based, community-led model, can help humanitarian actors act earlier, act smarter, and act with greater accountability. But optimism should be tempered by realism: without strong community ownership, continuous ethical oversight, and a commitment to inclusivity, technology risks amplifying harm rather than preventing it.

The challenge—and opportunity—for the humanitarian sector is to move beyond the allure of technology as a stand-alone solution and instead position AI as one component within a broader ecosystem of protection. The ultimate measure of success will not be the sophistication of the innovation itself, but the safety, resilience, and rights of the children it exists to protect.

8. Annexes (Tentative)

8.1. Table of Risks and Mitigation Strategies

Risk	Potential Impact	Mitigation Strategy
Bias in datasets	Exclusion of marginalized groups (e.g., girls recruited for domestic roles)	Diversity data sources; include qualitative community input; regular bias audits
Value misalignment	AI outputs contradict cultural norms or humanitarian principles	Apply feminist and community-centered design; embed humanitarian values in algorithms
Opacity (Black box) system	Wrongly exclusion/inclusion of children; lack of trust	Require explainable AI; train staff to interpret outputs; ensure community oversight
Digital inequality	Exclusion of women, rural elders, or low-literacy groups	Invest in offline/low-tech solutions; provide digital literacy support: gender-sensitive access design
Surveillance misuse	Data used by armed actors, risking child safety	Apply strict "Do No Harm" safeguards; anonymize data; enforce legal protections (CRC, IHL, GDPR)
Over-reliance on AI	Human accountability weakened; relational trust eroded	Maintain "human-in-the-loop" oversight; participatory decision-making; clear accountability lines
Lack of sunset clauses	Tools remain after crisis, creating long-term harm	Require time-bound deployments; regular reviews;

		dismantle systems once risks outweigh benefits

8.2. Excerpts From Field Reflections

Theme	Excerpt	Implications for AI and Child Protection
Gender barriers in digital access	Girls often have less access to phones and the internet compared to boys. Families worry about safety and cultural norms.	AI tools should account for unequal access; otherwise, girls' risk being excluded from protection systems.
Community Trust	Families trust local elders or women mediators more than outside organizations.	AI should complement, not replace, community-led protection systems.
Digital literacy gaps	Many rural elders are unfamiliar with mobile apps or online platforms.	Human-centered design and training are essential before AI-based tools can be effective.
Fear of surveillance	People worry that the information they share might reach armed groups.	Strong safeguards and data protection measures are critical to maintain trust and safety.

References

Arai-Takabashi, Y. (2019). *War Crimes relating to child soldiers and other children that are otherwise associated with armed groups in situations of non-international armed conflict. An incremental step toward a coherent legal framework?* QIL.

Beduschi, A. (2022). *Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks.* Cambridge: Cambridge University Press.

Berditchevskaia, A., Malliaraki, E., & Peach, K. (2021). Retrieved from Participatory AI for humanitarian innovation: https://www.humanitarianlibrary.org/sites/default/files/2023/10/Nesta_Participatory_AI_for_humanitarian_innovation_Final.pdf

Berditchevskaia, A., Malliaraki, E., & Peach, K. (2021). *Participatory AI for humanitarian innovation.* NESTA.

CaseBook: Child Soldiers. (2025). Retrieved from ICRC: casebook.icrc.org

Ferrer, X. (2021). *IEEE*. Retrieved from Bias and Discrimination in AI: A Cross-Disciplinary Perspective: https://technologyandsociety.org/bias-and-discrimination-in-ai-a-cross-disciplinary-perspective/?utm.

GSMA, Ipsos, GSMA, & Basis. (2022). *The Mobile Gender Gap.*

Guercio, L. (2025). *Artificial Intelligence and Future Perspectives of the International Humanitarian Law in Conflict Settings*. Retrieved from Taylor & Francis Group: https://www.taylorfrancis.com/chapters/edit/10.4324/9781003518495-26/artificial-intelligence-future-perspectives-international-humanitarian-law-conflict-settings-laura-guercio

ICRC. (2025). Retrieved from Our Fundamental Principles.

Klingefjord, O., Lowe, R., & Edelman, J. (2024). *What are human values, and how do we align AI to them?* New York: Cornell University.

KoboToobox. (2025). Retrieved from Intuitive and adaptable data tools to maximize your impact.

Kreutzer, T., Obrinski, J., Appel, L., An, A., Marrston, J., Boone, E., & Vinck, P. (2025). ethical implications related to processing of personal data and AI in humanitarian crises: a scoping review. *National Library of Medicine*, https://pmc.ncbi.nlm.nih.gov/articles/PMC11998222/?

Kreutzer, T., Orbinski, J., Appel, L., An, A., Marston, J., Boone, E., & Vinck, P. (2025). *Ethical implications related to processing of personal data and AI in humanitarian crises: a scoping review.* BMC Medical Ethics.

Legassicke, M., Johnson, D., & Gribbin, C. (2022). *Definitions of Child Recruitment and Use in Armed Conflict: Challenges for Early Warning.* Taylor & Francis.

Marelli, M. (2024). *Handbook on Data Protection in Humanitarian Action.* Cambridge: Cambridge University Press.

Martin, A., Sharma, G., de Souza, S. P., Taylor, L., & Eerd, B. v. (2022). *Digitisation and Sovereignty in Humanitarian Space: Technologies, Territories and Tensions.* Online: Taylor & Francis.

Mlambo, V. H., Mpanza, S., & Mlambo, N. D. (2019). *Armed conflict and the increasing use of child soldiers in the Central African Republic, Democratic Republic of Congo, and South Sudan: Implications for regional security.* Journal of Public Affairs.

Pizzi, M., Romanoff, M., & Engelhardt, T. (2021). *AI for humanitarian action: Human rights and ethics.* Cambridge: CCambridge University Press.

(2007). *The Paris Principles.* Paris.

UNHCR. (2025). Retrieved from Community-Based Protection (CBP).

UNICEF. (2022). *25 Years of Children and Armed Conflict: Taking Action to Protect Children in War.*

UNICEF. (2025). Retrieved from Policy guidance on AI for children: http://www.unicef.org/innocenti/reports/policy-guidance-ai-children

UNICEF, & Ministry for Foreign Affairs of Finland. (2021). *Policy Guidance on AI for Children.*