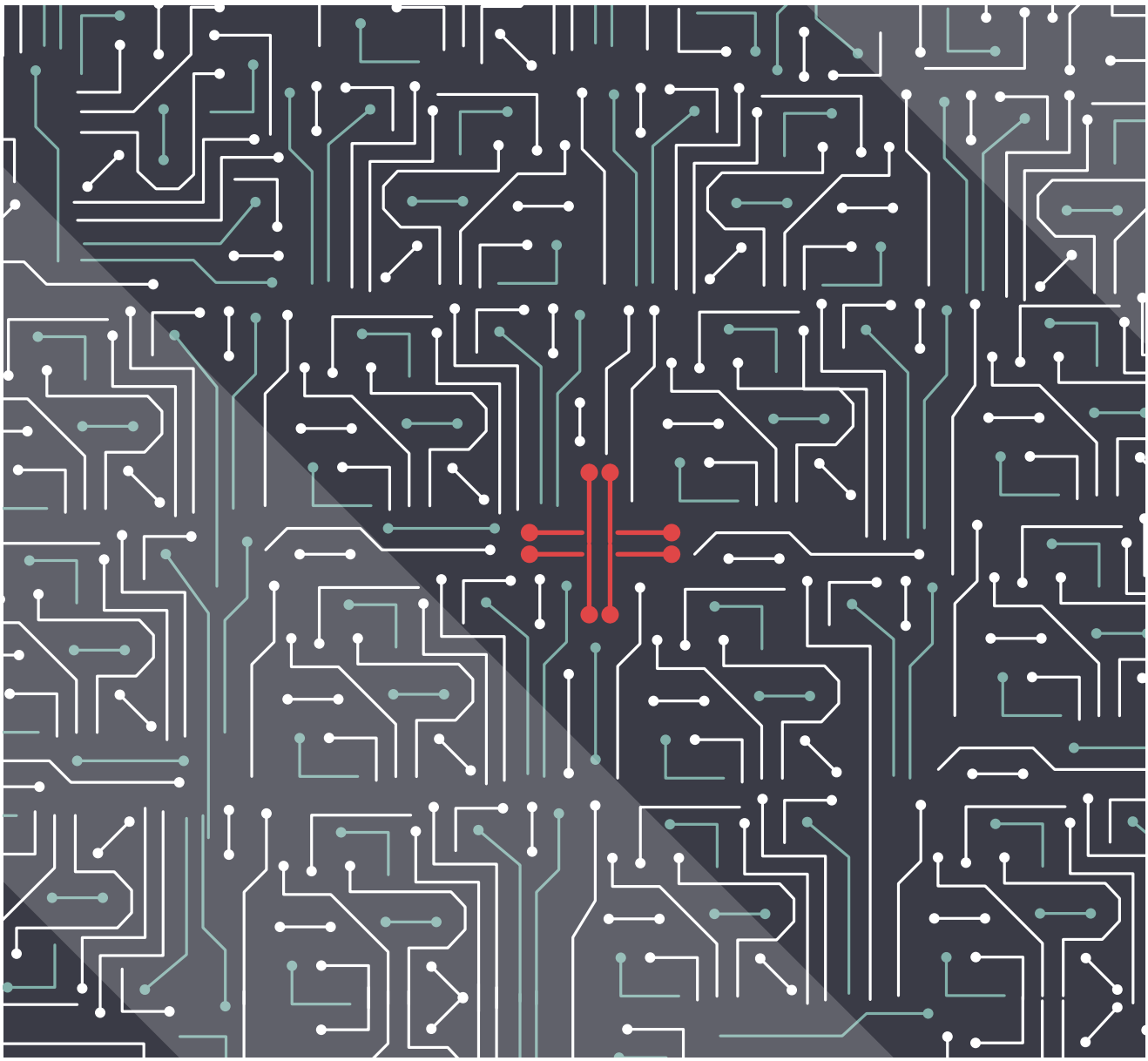
Humanitarian AI revisited

Number 89 May 2024

Network Paper

Seizing the potential and sidestepping the pitfalls

Sarah W. Spencer



HPN Humanitarian Practice Network

ODI

About the author
Sarah W. Spencer is an independent consultant and multi-domain expert working at the intersection of artificial intelligence (AI), conflict, security, and public policy. She helps governments, industry and civil society address the challenges posed by AI and capitalise ethically on its opportunities. She is currently on sabbatical from the British government. The views expressed in this paper are wholly her own and do not represent the official policies or positions of the UK government.

Acknowledgements
The author is indebted to those who kindly reviewed drafts and offered valuable insights, including Kerrie Holloway, Sara Hussain, Helen McElhinney, Jessica Rennoldson and others.

Disclaimer
The contents of this report were entirely human generated. No content created with artificial intelligence tools is included in this report.

Humanitarian Practice Network (HPN)
ODI
203 Blackfriars Road
London SE1 8NJ
United Kingdom
Tel: +44 (0)20 7922 0330
Fax: +44 (0)20 7922 0399
Email: hpn@odi.org.uk
Website: www.odihpn.org

About HPN
The Humanitarian Practice Network at ODI is an independent forum where field workers, managers and policy-makers in the humanitarian sector share information, analysis and experience. The views and opinions expressed in HPN's publications do not necessarily state or reflect those of the Humanitarian Policy Group or ODI.

Cover photo: Jessica Rennoldson for HPN

Network Paper
Number 89 May 2024

HPN
Humanitarian
Practice Network

Humanitarian AI revisited: seizing the potential and sidestepping the pitfalls

Contents

Chapter 1 Introduction

A great deal has changed since 'Humanitarian AI: the hype, the hope and the future' was published in November 2021. At that time, only a handful of humanitarian actors were exploring the potential of artificial intelligence (AI) and machine learning (ML) (see Box 1) and these experiments were largely led by agencies that had the requisite staff, data, cash, and/or relationships with technology firms to do so. In fact, identifying a sufficient number of individuals to interview in 2021 for that paper, who had both the interest and a basic understanding of both AI and humanitarian action, proved challenging.

Then, in November 2022, OpenAI released ChatGPT, arguably triggering the latest AI boom. Big Tech firms raced to release their own generative AI models. McKinsey estimated that generative AI 'could add trillions of US dollars in value to the global economy'. Comedians, writers, and media companies went on the picket lines and to the courts. Some of the world's leading AI engineers and pioneers became doomsayers. And it became seemingly impossible to read, listen to or watch something about AI without hearing the words 'inflection point'.

As Bill Gates suggested, it seems that 'the Age of AI has begun'. And humanitarian actors have taken note.

Box 1 Artificial intelligence and machine learning

While there is no universally accepted definition of AI, the term is often used to describe a machine or system that performs tasks that would ordinarily require human (or other biological) brainpower to accomplish, such as making sense of spoken language, learning behaviours or solving problems. There is a wide range of such systems, but broadly speaking they consist of computers running algorithms (a set of rules or a 'recipe' for a computer to follow), often drawing on a range of datasets. ML uses a range of methods to train computers to learn from existing data, where 'learning' amounts to making generalisations about existing data, detecting patterns or structures, and/or making predictions for new data. Experts offer conflicting views about the relationship between AI and ML. This paper takes no formal position in these debates, and throughout this paper AI and ML are referred to as interrelated systems.

The time, therefore, seems right to publish an update to the 2021 Network Paper on humanitarian AI. In 2021, many of the humanitarian experts interviewed for the paper expressed some reticence about the use of AI in humanitarian settings. Increasingly, these risk-averse attitudes now seem set to be replaced with more risk-neutral ones.

Dozens of aid agencies are now ostensibly developing AI applications and designing pilot projects to see how AI might help them improve the efficiency of their operations, boost their productivity, expand their

reach, increase the speed of their response, and improve the overall quality of the support they provide. Some are exploring higher-risk applications related to protection, aid delivery in Gaza, and children's facial recognition. A leader of one international non-governmental organisation (INGO) suggested that his agency was in a race to adopt AI and needed to 'catch up'.

This uptick in interest is set against a complex backdrop of rising need, stretched humanitarian financing and an increasingly brittle international system. More people are on the move, displaced by conflicts and climate-related crises, and political pushback against migrants and asylum seekers in Europe and the United States is growing. Narratives that offer panacea-like solutions grounded in AI to tackle the world's greatest challenges are rife – encouraged, in part, by tech firms and leaders with colossal geopolitical and economic influence.

This paper attempts to cut through some of the recent AI hype and offers suggestions on how AI can be safely and ethically adopted and deployed in support of humanitarian action.[1] It focuses solely on the **deployment or use of AI systems** rather than other parts of the AI life cycle, such as the **ethical design and development** of AI algorithms and systems, as this latter issue has been well covered across a range of industry and academic papers and sits outside the scope of this paper.[2]

It also provides insights into how humanitarian actors are using AI and algorithmic tools as well as the risks of doing so, many of which are similar to, if not the same as, those highlighted in the 2021 Network Paper. It focuses primarily on how AI is being integrated within existing humanitarian systems and adopted by large, multinational humanitarian agencies. Thus, the conclusions and recommendations in this paper may prove less relevant for smaller organisations or new market entrants, such as public–private ventures or social enterprises who offer so-called 'tech for good' services. This specific focus is not intended to confer a value judgement but rather reflects the aggregate fiscal and political power that a dozen or so of the largest aid agencies wield and thus the collective weight of their decisions related to AI.

Finally, the paper offers reflections on the role of international AI standards as well as recommendations on AI procurement and governance, including assurance tools to establish trustworthy AI – an AI system that generates reliably accurate results.

The paper is presented in five chapters:

Chapter 2 offers an overview of current and potential AI/ML applications across the humanitarian sector.

1 The paper is focused on AI more broadly and not only on generative AI.

2 See, for example: Brey and Dainow (2023), 'Ethics by design for artificial intelligence'; García-Cuesta (2020), 'Artificial intelligent ethics in the digital era: an engineering ethical framework proposal'; Muhlenbach (2022), 'Artificial intelligence and ethics: an approach to building ethical by design intelligent applications in El Morr (ed) *AI and society*; Rossini et al. (2020), 'Actionable ethics through neural learning'; and, Yampolskiy (ed.) (2018), *Artificial intelligence safety and security*.

Chapter 3 explores the risks and limitations of using AI/ML in the humanitarian sector, building on the arguments initially offered in the 2021 paper.

Chapter 4 offers a roadmap and recommendations for aid actors looking to safely and ethically integrate AI into their work.

Chapter 5 suggests additional considerations on how humanitarian actors across the sector should approach AI.

Chapter 6 concludes with thoughts on future trends related to humanitarian AI.

Hyperlinks to additional resources have been provided throughout the paper and in Appendix 2.

Chapter 2 AI applications with humanitarian impact

Humanitarian actors, donors and technology firms (amongst others) are exploring a range of ways to deploy AI systems in support of humanitarian action. While generative AI tools, like ChatGPT, have generated significant excitement of late, other AI applications could potentially yield important wins for humanitarian action. These include image recognition and classification, scientific and/or new product discovery, gene editing, writing computer code, object detection and tracking, facial recognition, expert systems, predictive analytics, robotics and uncrewed vehicles, task automation, and predictive maintenance.

This chapter includes examples of how AI and ML are being used to improve humanitarian impact. These are presented in three groups: applications that deliver immediate **impact for individuals, their communities, and the programmes** designed to support them; applications that improve the **internal operations and systems** of humanitarian agencies; and applications that are changing the **external environments in which humanitarian agencies operate**.

Though some of the applications below hold promise and may help to improve the impact of humanitarian action, all of the examples highlighted here come with risks that must be properly identified and understood before they are integrated into humanitarian aid programmes and systems. These include algorithmic failure, unintended socioeconomic harms, and cyber attacks. Whilst reference to some of these risks is included in this chapter, a more thorough analysis of the risks related to AI adoption is included in Chapter 3.

Individuals, communities and programmes

This section includes examples of AI applications that are designed to directly improve the wellbeing of individuals and/or their communities or otherwise deepen the impact of humanitarian interventions.

As in other industries, humanitarian agencies are exploring ways to use algorithms to gain new or novel insights into their existing internal data or publicly available, external data. AI/ML models are particularly adept at analysing geospatial imagery and some of these applications are being deployed to support early-warning systems and efforts related to climate adaptation. They have also been used to assess damage to building and infrastructure as a result of conflicts, earthquakes and other crises. These kinds of assessments are increasingly offered pro bono or made public by private sector software or geographic information system (GIS) firms, while other firms use AI/ML to transform data into localised maps that offer additional insights into community needs, attributes and preferences.

Agencies are also using AI to expand the reach of their needs assessments, for example, by using chatbots to engage with communities and ask questions about their needs through mobile messaging or other platforms. Chatbots and large language models (LLMs) are also being used to provide more personalised and targeted support, for example, in the design of bespoke curricula or education plans for teachers

and students alike, or providing a more personalised, customer-focused service. And other actors are exploring ways in which chatbots might provide information to affected communities about their rights, which services are available and where, and how to seek recourse when something goes wrong.

But poorly deployed chatbots can increase user frustration, decrease trust in the institutions deploying them, and, in the worst instances, violate laws related to data protection and privacy. Given the frustrations many of us experience when engaging with virtual assistants built into online shopping or banking platforms, it's easy to imagine scenarios where these frustrations increase exponentially for those seeking support for far more significant decisions related to personal safety, legal rights and wellbeing. Equally, as one expert suggested, replacing human operators with virtual assistants somehow ignores the 'human' element of humanitarian response, threatening the principle of humanity.

AI shows potential in predicting some types of need and, thus, delivering higher-impact responses that are more precisely targeted and/or delivered more quickly. This could be particularly useful for disaster risk reduction and anticipatory action linked to climate-related events, as well as for public health and epidemiological surveillance. For example, some models and applications can predict the location and impact of floods; if, where, and how an epidemic or pandemic will spread; how viral outbreaks might mutate; and levels of morbidity and mortality linked to these outbreaks.

AI systems have also been used to improve diagnostics in low-resource and/or remote settings by supporting screening for cervical cancer, tuberculosis, malaria, and antimicrobial resistance. Some actors have drawn on AI/ML to design vaccination campaigns and improve dissemination by combining population estimates with predictions in viral outbreaks and mutations to improve the delivery of vaccines.

Humans are 'collaborating with AI' to tackle unsolved problems and advance scientific discovery more broadly, which could arguably prove as game-changing for humanitarian action as, for example, ready-to-use therapeutic food was 30 years ago. This includes drug discovery – for example, the design or improvement of existing vaccines like the malaria vaccine – and identifying new forms of fuel or energy storage which could be less environmentally damaging and/or more efficient than existing ones. AI is also increasingly being used to improve agricultural yields, minimise waste, and reduce environmental harms related to farming, though these are not without their risks, as researchers have pointed out. For example, as with other applications, integrating and relying on AI to improve crop management and production increases the vulnerability of these systems to algorithmic failure and cyber attacks.

A few agencies are exploring other more complex, and arguably higher-risk, applications, such as supporting family tracing with AI-powered image recognition tools, using AI to improve the forensic identification of human remains, and drawing on AI to predict the movements of crisis-affected populations. These applications, however, are less frequent and have some important limitations, as discussed in Chapters 3 and 4.

Internal operations and systems

While the following examples of AI/ML applications may seem less glamourous than those noted above, they could yield efficiency or productivity gains which might help humanitarian agencies do more in the face of static or diminishing resources.

For example, in other industries, AI has been used to manage supply chains by supporting volume forecasts to move products closer to where they are more likely to be needed with greater efficiency, or by optimising transportation corridors and redesigning logistics networks. AI has also proven effective in managing infrastructure through predictive maintenance, though at this stage, the benefits of these types of application may only be realised by large agencies that own or operate significantly large amounts of infrastructure, such as aircraft and large fleets of vehicles, water treatment facilities, or power-generation systems.

A number of aid agencies are purportedly exploring how to use AI, particularly generative AI, to improve their institutional knowledge management. While humanitarian agencies generate significant amounts of knowledge and information, high rates of staff turnover can create gaps in institutional memory, both in country offices and headquarters. With the right inputs – including curated data and information, guardrails, and data governance strategies – virtual assistants can help staff explore and interpret institutional knowledge and expertise.

Generative AI tools can help reduce busy work when integrated into internal business systems and processes. For example, drawing on agency-specific material, AI can help to create new content, like press releases, fact sheets, summaries of requests for proposals, proposals, donor reports, job descriptions and interview questions. It can also summarise team chats or emails and offer suggestions on how you might respond to those emails or discussions.

AI and ML tools have also been used for some time in support of financial services. By analysing historical financial data to model future scenarios, AI can help spot outlier transactions or potential fraud, support accounts payable systems, and improve budget and payment management.

And some agencies are exploring how to use AI systems to improve their security assessments, by scraping the web and/or automating information collection and assessment related to sub-national security incidents. This mirrors similar trends in the defence and national security space where actors are exploring ways in which human–machine teaming can improve the analysis of large amounts of information.

External environments and structural dynamics

Whilst many humanitarian actors have increased their interest in AI, only a handful are tracking the changes it is having or might have on the environments in which they operate. State parties to conflict are increasingly integrating AI into their defence and national security infrastructure while non-state groups are allegedly seeking to exploit AI for their strategic advantage. This includes not only integrating

AI into artillery and weapons systems, but also: military decision-making; battlefield logistics and medical care; intelligence, surveillance, target acquisition, and reconnaissance (ISTAR); cyber-offensive and defensive capabilities; and detention.

Whilst this might seem only tangentially relevant for humanitarian actors, AI is arguably changing where and how wars are fought and planned. And yet, few (if any) humanitarian agencies have sight into how these AI systems are trained to recognise civilian objects and respect international humanitarian law (IHL). As parties to armed conflicts rarely disclose their military-technological capabilities, the increased use of AI could create, as one expert calls it, a 'double black box' where technical opacity is concealed in military secrecy, further hindering efforts to hold accountable those who violate IHL.

On the flip side, actors are exploring ways to use AI to understand and identify impediments to peace. In fact, some accounts suggest that while the Israeli government is using AI to 'accelerate a lethal production line of targets that officials have compared to a "factory"', others are using AI to help identify policy options for peace.

AI is also corroding the information ecosystem in conflicts. Studies have shown that 'false information spreads faster than true information'. While disinformation has long been used to confuse, discredit and undermine efforts by parties to armed conflicts, the increasing role of AI in disinformation and deepfakes poses significant risks in conflicts. AI is helping to create hyper-personalised and targeted messaging designed for specific audiences as well as more effective distribution campaigns. As the quality and dissemination of deepfakes and other forms of disinformation increase, experts suggest that it will become easier to dismiss authentic content as inauthentic, increasing pressure on any attempts to pause or end conflicts. It also creates a 'liar's dividend', allowing parties to conflict and other leaders to dispute the authenticity of their own genuine misbehaviour, spreading doubt in both fake and real communications alike, complicating diplomacy, and poisoning political discourse.

Chapter 3 Extant risks related to AI and humanitarian action

Some AI applications in the humanitarian space undoubtedly show promise, particularly those that make accurate predictions on the anticipated, geographic impact of hazards and natural disasters and share these predictions publicly. But, as ever, plenty of pitfalls remain that require careful consideration and the risks highlighted in the 2021 report still stand.[3] This chapter complements the risks highlighted in 2021 and offers analysis on additional, extant challenges related to AI and humanitarian action.

Humanitarians fail to keep crisis-affected individuals and communities at the centre of their work

As humanitarian agencies increase their exploration of AI and experimentation grows, the extent to which they engage and consult with those impacted by AI systems and algorithmic decision-making remains unclear. This includes not only consultations on how AI systems will impact the lives of populations that aid agencies are meant to serve, but also if and how any data related to humanitarian programmes, including personally identifiable information, can and will be used to power AI models deployed by humanitarian actors.

> 'Are refugees, migrants, and others who are asked to hand over their data fully aware of the many ways in which their data could be used? Do they feel empowered to say no? There's often a vast gap between policy, practice, and what people who use aid perceive.'
> *The New Humanitarian, 4 January 2023*

The 2021 report highlighted the risk that AI might supplant the voices and views of crisis-affected populations. This remains true, particularly in instances where AI systems are being deployed to predict or anticipate need. Moreover, in the near future, as the performance and reliability of AI systems improve, some aid agencies may increasingly rely on or value the outputs of these systems, prizing them over the views and preferences of affected populations. This could threaten community participation and the ability of crisis-affected populations to influence how resources are used, how programmes are designed, and who benefits.

Equally relevant is the risk that expanding and deepening partnerships between humanitarian agencies and AI providers undermines sector-wide commitments to localise humanitarian actions and shift power and resources to the Global South. Many of the world's most powerful AI systems are built and sold by large, multinational tech firms, predominantly based in North America and Europe. Collaborations between aid agencies and these providers, while potentially offering access to higher-performing systems, could unintentionally shift decision-making power away from the Global South. And, while there

3 The risks and challenges in the 2021 report include: effectiveness, bias and discrimination; localisation and community participation; AI accountability, transparency and human oversight; regulation, data protection and privacy; and safety and sustainability.

are small and medium-sized firms in the Global South building AI solutions with a specific eye toward delivering progress on the Sustainable Development Goals – particularly in tech hubs in East Africa, South Asia, and southeast Asia – the extent to which aid actors are exploring relationships with these firms remains unknown.

At the same time, the efficiencies offered by AI systems may also reduce the number of employees based in the Global South rather than the Global North, if jobs in the former are more susceptible to automation. Detailed cost–benefit analyses, mapped against specific humanitarian sectors and geographic areas, might help to mitigate any harmful impact on humanitarian labour.

Finally, there remains a significant risk that aid agencies are using or will use data related to crisis-affected populations, generated through the delivery of humanitarian programmes, to power AI systems without the explicit consent of these populations. Aid agencies' track record of seeking individual consent to collect and share sensitive data remains patchy at best. While data protection policies and the European Union (EU) General Data Protection Regulation (GDPR) have helped, some argue that such regulation doesn't apply to international public organisations, such as United Nations (UN) agencies. This same argument has been made in reference to the EU's AI Act.

This raises the question: can agencies who are not bound by, or fail to comply with, GDPR and the AI Act use the linked or linkable data they hold on millions of people without their consent to power an AI which might improve support for those populations? While some agencies are presumably discussing this challenge behind closed doors, a public debate on its legal and ethical trade-offs might advance a sector-wide position on data ownership, aid efficiencies, and the agency and consent of crisis-affected populations.

Deploying AI contributes to wider trends related to 'humanitarian surveillance'

Data is the fuel that powers AI. The drive to identify and adopt AI applications in humanitarian contexts has, arguably, compelled aid actors to collect, structure and clean more data than ever before. The humanitarian sector is still 'struggling to develop the calculus' to weigh the perceived benefits of digital systems against the costs to individual rights and privacy. Moreover, aid actors have yet to develop a collective position on the push from large institutional donors to collect linked or linkable information to demonstrate aid is not being wasted.

While some aid agencies argue that monitoring and collecting data are essential for delivering the right amount of aid to the right people at the right place and time, extensive data collection and deploying AI in armed conflicts can unintentionally contribute to humanitarian surveillance, particularly where these data and tools are shared with states and government agencies, as many international public organisations are required to do. Data assets and AI can be used to police vulnerable groups, track their movements across borders, control their mobility, and limit their opportunities. One aid worker reported that her agency had used biometric data to monitor whether the clients were accessing medical checks above their allotment.

Models fail to perform as expected

AI models can present incorrect statements as fact (also known as hallucinations), produce inaccurate results, or otherwise fail to perform as expected for a host of reasons. In cases where the data they rely on is insufficient, AI systems can fail catastrophically, sometimes referred to as brittle failures. AI is also vulnerable to adversarial attacks, such as data-poisoning attacks, where adversarial data is fed into AI and ML systems causing them to make mistakes. Adversarial AI can alter what AI learns from training data or how it solves problems, and such manipulation is often difficult to detect until after a mistake is made, which poses significant risks in cases where AI is supporting high-stakes decision-making. The accuracy of AI and ML models can also 'drift' when a model's performance or prediction power degrades over time due to changes in variables like data or key concepts related to the model.

These algorithmic errors or outright failures can be difficult to detect. For example, while many aid workers would presumably be able to spot outright bias and discrimination or factually inaccurate information generated by a system like ChatGPT, it's typically far more difficult to spot errors in other models, for example one designed to predict individual risk or summarise a 200-page request for proposals. Like many of the machines used by humans, AI systems need regular maintenance and performance testing to ensure they are operating as intended. This often requires employing third-party agencies who assure the reliability and performance of models (see Chapter 4).

AI increases cyber security risks

The risk of AI-powered cyber attacks is increasing. New research concludes that existing cyber defence infrastructures are already or will soon become unable to address the increasing speed and complex decision logic of AI-driven attacks. The United Kingdom's National Cyber Security Centre recently assessed that AI would 'almost certainly increase the volume and heighten the impact of cyber attacks over the next two years'. Whilst AI can strengthen the cyber security of aid agencies' digital systems, for example, by improving the detection and triage of, and response to, cyber attacks, once AI is introduced into a cyber security or other system, it increases the attack surface areas and creates new vectors for attack, making either the AI algorithms and/or data assets more vulnerable.

This should give agencies pause, given both the public and undisclosed number of cyber attacks on humanitarian actors as well as the fewer cyber-defensive resources they have at their disposal, as compared to private sector firms or government agencies.

Human–machine teaming adds unanticipated complexity and outcomes

Human–machine teaming could bring about faster decision-making and more robust situational awareness tools. But the ethical and legal implications of ever-increasing human–machine interactions in humanitarian and/or conflict settings remain unclear and under-researched.

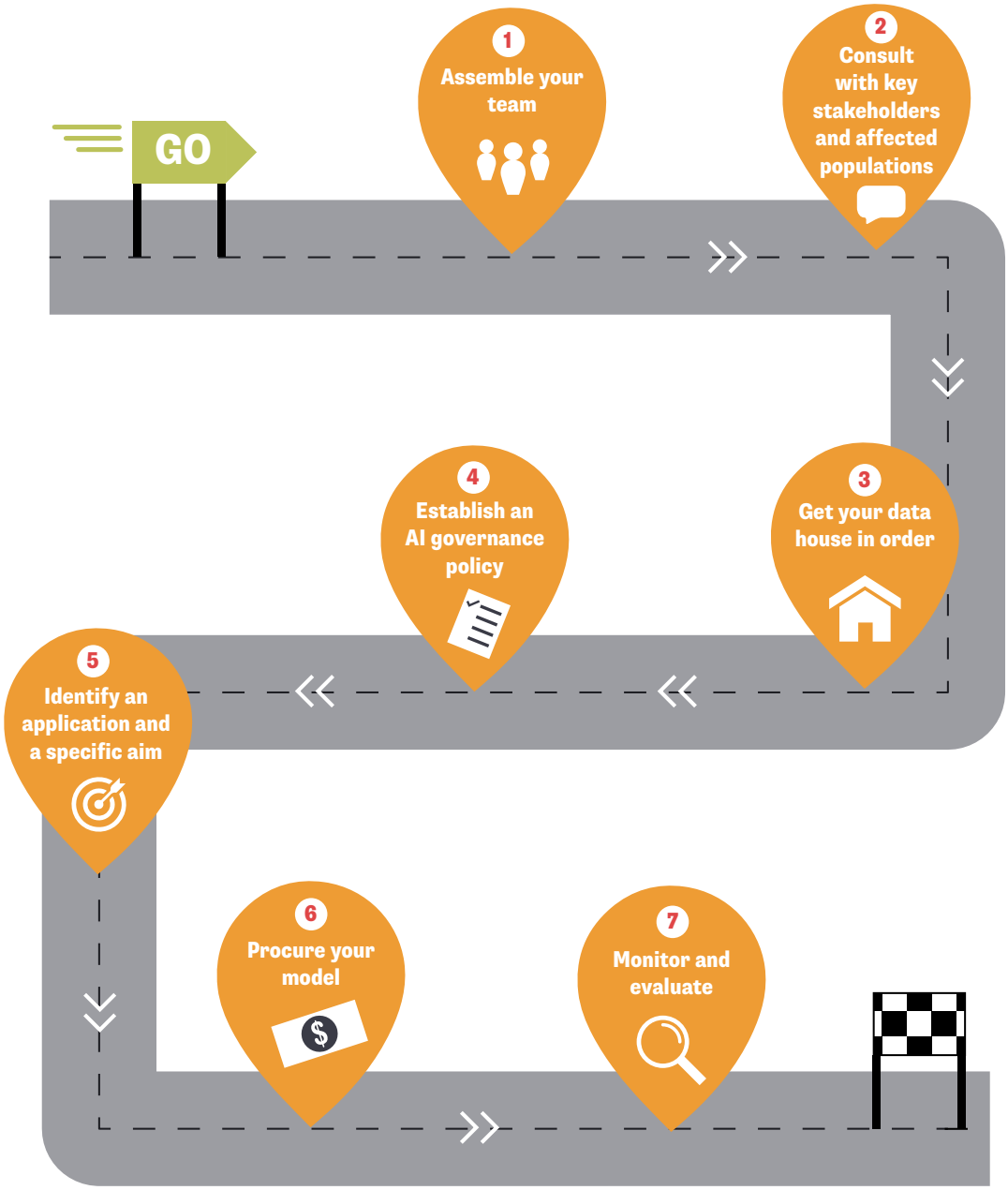
Human insights, emotions and bias related to AI and/or particular situations of armed conflict may skew how individuals observe, understand, and act upon AI-related outputs. And integrating an AI into existing humanitarian decision-making systems or automating certain parts of a task may not always drive greater efficiencies.

For example, when using the outputs of an AI to inform decisions about where and when to allocate humanitarian resources – say, with AI models designed to improve epidemiological surveillance – some people may be overly cautious in their assessment of the accuracy of an AI's outputs, imposing a higher cognitive load on humans who are trying to determine whether the AI is performing correctly. On the contrary, others may be less cautious or even reckless in their judgment, increasingly trusting the outputs of a model without properly interrogating its accuracy and reliability.

Chapter 4 A roadmap for the ethical uptake of AI in humanitarian action

To date, little attention has been given to how AI systems should be integrated and managed within humanitarian agencies' organisational structures and existing operations. Yet, many of the decisions AI is being called upon to support are quite literally life and death, raising the stakes considerably for humanitarian action. This chapter outlines some of the steps required to safely and ethically adopt AI systems in support of humanitarian outcomes (Figure 1). This list is not exhaustive; deploying high-risk technologies like AI requires particular care and caution.

Figure 1 A roadmap for the safe and ethical adoption of AI in humanitarian action



Assemble your team

Decisions around when and how to adopt and deploy AI systems are complex and require inputs from multiple teams with a variety of skillsets and expertise, including, amongst others, data protection officers, general counsel, IT teams, data scientists, programme and/or technical unit staff, community engagement experts, and staff who possess a deep understanding of the local context into which the models will be deployed.

Too often, the skills and expertise of computer scientists and AI/ML engineers are privileged over those of so-called 'domain experts', namely those with extensive humanitarian experience, technical expertise in relevant sectors – such as education, healthcare – and crisis-affected communities themselves.

In one example, often cited by academics in science and technology of how this bias can generate real-world harm, a group of ML engineers developing a model to support migraine diagnoses and treatment approached medical knowledge and expertise as something that could be extracted and almost downloaded. In their minds, migraine 'experts' were physicians and thus, these were the only experts with whom they consulted, overlooking not only the views and expertise of other medical staff like nurses, who spent more time interacting with patients, but also the patients themselves. It doesn't take much of a stretch to imagine a similar scenario playing out in the humanitarian sector, where the views of staff recruited from crisis-affected communities and/or recipients of support themselves are deemed less valuable than those of AI/ML engineers.

Consult with key stakeholders and affected populations

Hard-won gains have been made in recent decades to establish standards on accountability to affected populations. This includes sector-wide commitments to consult and partner with communities in the design of humanitarian interventions and to establish a right to recourse when things go wrong.

Despite these advances – including high-level commitments from agencies and the adoption of standards related to accountability – consultations with communities on the use of AI and algorithmic models remain worryingly infrequent.

The CDAC Network and FilmAid recently conducted community consultations with refugees living in Kakuma Camp in Kenya about AI and how it should be used in humanitarian contexts. Refugees and host communities with whom they spoke not only knew about AI but some were experimenting with LLMs to see how they could be used to write job applications and support small businesses. They also had views on how they thought AI should be used in Kakuma for both the administration of aid and to improve accountability and transparency.

Agencies seeking to integrate AI and algorithmic systems into aid operations must prioritise crisis-affected communities and humanitarians' accountability to them. This means engaging with communities before deploying AI systems that will impact the agency communities have over the

services they are provided. The costs and benefits to individuals and communities related to AI uptake must be factored into decisions about whether and how to deploy AI in crisis settings. And community participation processes must be embedded into and shape the design of programmes in ways that are consequential and not just rhetorical.

Get your data house in order

'Without data, there is no AI.' The uptick in AI hype may incentivise some agencies to amass more data or leverage existing data assets to power AI pilots. But AI cannot be used responsibly and safely deployed unless the data used to both train and run the AI is safely and ethically governed.

At a minimum, agencies should establish and adhere to a data governance strategy grounded in the principles central to data protection, particularly data minimisation, and comply with domestic and international regulations related to data protection and privacy. Increasingly, information and data protection commissioners are issuing guidance on how to deploy AI systems whilst remaining compliant with existing data protection requirements. And tools such as data protection impact assessments (DPIAs) – as well as AI-specific risk assessments – will continue to prove useful.

Establish an AI governance policy

While AI applications for humanitarian outcomes have increased in the last several years, collective discussions and agreement on how to govern its use have lagged. At the time of writing, only a handful of agencies appear to have established an internal, agency-specific document to guide their use of AI. In the absence of internal guidance and red lines, some teams and staff within humanitarian agencies have started experimenting with AI tools, at times without fully understanding the data privacy and security issues this might create.

The development of an AI policy should come before AI applications are piloted, as AI policies and internal red lines should help determine which AI applications are ethical and safe. At the minimum, an agency-specific AI policy should include:

- **The aims and objectives of the policy.** This could include, for example, promoting the ethical use of AI; ensuring compliance with relevant regulation, data protection, or privacy; and/or mitigating AI-related risks and harms.
- **Core principles on which the AI policy is grounded.** For humanitarian agencies, this should include, at a minimum, a fundamental commitment to humanitarian principles, ethics and existing standards.
- **An accountability framework.** This framework should be grounded in your core principles (above) and will be inherently linked to your risk-management approach (below). It should identify your core commitments to key stakeholders – such as crisis-affected communities, staff and donors – or what AI industry leaders often refer to as 'interested parties'. This includes how they will be consulted in the design and deployment of AI models and any right to recourse that interested parties may have if or when things go wrong.

- **A risk-management approach.** At a minimum, this should document identified risks, their potential impact and likelihood, and include risk treatment options. Risk and accountability should be allocated to those parties best able to manage it. For example, accountability for some issues may reside with specific humanitarian staff while others may lie with an AI supplier. AI impact and risk assessments can be used to regularly update the approach.
- **Agency-specific red lines.** This includes actions related to AI which are and are not allowable with regards to: the access to and use of data (building on your existing data governance strategy); the design of an AI model; procuring AI systems and supplier engagement; and specific applications of AI, including contexts in which models are deployed and populations who should not come into contact with AI systems, such as children.

While this may seem a daunting and time-intensive task, an internal AI policy can start as a list of dos and don'ts on which an agency-wide policy can be built. Additionally, an AI policy should be treated as a live document and regularly reviewed (at least several times a year, depending on the agency and its use of AI) and updated as necessary, given the rapid changes to AI models themselves as well as the wider AI industry and its regulation.

Given the widespread and longstanding use of humanitarian-specific standards, leaders in the sector might draw inspiration from existing international AI standards when drafting internal AI policies (see Chapter 5 below). AI standards are often a useful tool for multiple teams and departments within an institution. They can offer clarity and recommendations to manage the challenges related to AI governance and risk management. They can also provide practical guidance on safely designing and developing AI systems, including on how to manage the range of technological tools and components that are drawn on at various stages to design that system, often referred to as the tech stack.

Identify an application and a specific aim

Albert Einstein supposedly said that if given an hour to solve a problem, he would spend 55 minutes understanding the problem and 5 minutes solving it. Despite the apocryphal nature of the quote, its meaning remains acutely relevant when selecting a problem suitable for an AI solution.

Most humanitarian actors would agree that AI solutions should be problem-centred. Yet many AI applications and AI discussions related to humanitarian action still appear to be centred around AI. In other words, agencies are searching for potential problems that they could use AI to solve rather than identifying their most critical challenges and exploring a host of solutions.

It's time to put the horse before the cart. After all, when AI is your hammer, every problem will begin to look like a nail. But not every challenge in the humanitarian sector can or should be solved by AI. AI is generally considered a high-risk technology and there may be more privacy-preserving or effective solutions than AI. AI solutions can also be extremely costly, though this may not seem apparent at the outset, particularly if pro bono offers from tech firms are on the table (see 'Procure your model' section below).

If you conclude that an AI solution might prove useful, ensure the purpose of your application is suitably specific. One of AI's greatest claims to fame – though it remains hotly debated in some circles – is its ability to reduce the money, time, or effort required to achieve a specific output (efficiency). AI also promises to help individuals do more work over a given period of time (productivity).

Though linked, productivity and efficiency gains are not the same and if your AI solution aims to deliver productivity and/or efficiency gains, these should be made explicit at the outset and continuously assessed. For example, the outright costs of procuring an AI system – let alone the costs in staff time and salaries to integrate and maintain a model – can often outweigh any near-term benefits.

One public sector worker in the UK disclosed that a medium-sized technology firm had offered to build a bespoke AI model for local government to generate greater insights into social challenges related to housing, employment and other support for vulnerable populations. The costs associated in doing so (£250,000) could have arguably covered the wages of 8–10 additional social workers. And this amount excludes any additional costs related to data quality and management processes and AI maintenance. All the costs and potential benefits associated with deploying an AI model for humanitarian outcomes must be carefully considered to assess whether doing so results in a net positive for humanitarian agencies and the populations they serve.

Additionally, in the humanitarian sector, it's crucial to test your assumptions and work out who stands to win and lose when these gains are realised. For example, some AI-powered projects may, in fact, deliver efficiencies by reducing the number of staff based in the Global South while channelling resources to tech firms in the Global North. Equally, better data insights and analysis may not lead to a better or more benevolent response or, in fact, any response at all. Even if a model accurately predicts the impact and likelihood of a natural hazard on a given population, that may not, in and of itself, be sufficient to attract donor funding or garner the political impetus required to trigger a significant humanitarian response.

Procure your model

This is where the rubber hits the proverbial road for humanitarian agencies. While data protection and privacy and AI use cases can feel far more relevant to discussions related to the safe and ethical uptake of AI, procurement is equally important and, in some cases, more complex.

For a start, many agencies find it difficult to identify the right AI supplier and to determine which criteria are most important in vetting potential suppliers. Within the humanitarian sector, suppliers of AI systems appear to be selected for one of three reasons:

- Technology firms have pre-existing, **established agreements** with a humanitarian agency and AI services are discussed and agreed as an extension of that existing agreement.
- **Individual relationships** exist at the highest levels of agencies (often at the president or vice-president levels of management) and serve as the basis on which to establish AI partnerships that are mutually beneficial for both aid agencies and large technology firms.

- AI-related services and support for a small number of narrow use cases and pilot projects are procured through an **invitation to tender or a request-for-proposal** process.

Given its high cost, access to some AI systems is often provided to humanitarian agencies by tech firms at below-market pricing or at no cost. But these offers often bring with them unanticipated costs. For example, the cost of third-party stress-testing and model validation, as discussed above, may exceed the budgets of humanitarian agencies. Yet, they remain a critical component of assuring the deployment of so-called trustworthy AI. Equally, pro bono solutions offered to public or non-profit agencies are rarely bespoke and typically involve granting access to existing systems and products that aren't fit for purpose and deliver sub-optimal results.

Thorough due diligence can help avoid these pitfalls and ensure agencies are identifying all the benefits and costs associated with offers of support. To this end, agencies should employ the right experts to interrogate how the AI models were built, with which data, and how they were trained. This requires an examination of the full tech stack as well as how a firm has managed the entire life cycle of the AI system being deployed. It is equally critical to ensure that knowledge transfer and training are included in your partnership agreements with AI suppliers, including training for non-specialists. This is often overlooked in pro bono arrangements between tech suppliers and humanitarian agencies.

Finally, agencies should carefully consider from where they procure their AI. The amount of money invested into AI is vast and firms with vested interests in the technology have a strong incentive to see that AI delivers against a range of benevolent causes. Some may even be keen to co-opt the narratives and reputations of charitable endeavours to offset critiques in the press about AI risks.

Moreover, most of these agencies maintain relationships and procure services from the world's largest technology corporations, many of which, according to some, play an outsized role in the creation and execution of foreign and national security policy and operate as geopolitical actors. This is putting the principles of neutrality and impartiality under increasing pressure.

Identifying the interests and incentives of AI providers, particularly those offering below-market pricing or pro-bono services, remains critical. It's equally important to remember that Big Tech isn't the only gig in town. The number of AI suppliers and firms in the Global South are growing, with many receiving donor funding specifically to develop AI to advance the Sustainable Development Goals.

Monitor and evaluate

Whilst most humanitarian staffers are familiar with the purpose of and key concepts related to monitoring and evaluation, it's not immediately apparent that this experience has migrated across to their debates and discussions about AI. In fact, it is not clear how many agencies, if any, have committed the appropriate levels of investment into the tools and infrastructure required to both track and assess the impact of AI as well as to test and maintain a model's performance (see also Chapter 3, above).

Just as one wouldn't purchase a car or a computer and expect it to run perfectly without regular maintenance, we shouldn't expect AI systems to perform flawlessly without subjecting them to regular stress testing and auditing to validate their outputs and identify potential vulnerabilities or probability of error.

AI assurance techniques – activities that measure and evaluate the reliability and effectiveness of AI models – are a critical way to ensure the safe and ethical adoption of AI, particularly by non-technical institutions. The Organisation for Economic Co-operation and Development (OECD) has compiled a catalogue of tools and metrics for assuring AI is trustworthy and the UK government has recently issued a portfolio of assurance techniques. These compendia, amongst others, could prove useful touchstones for humanitarian actors designing AI-powered interventions.

Agencies exploring the uptake of AI should establish appropriate mechanisms that enable and allow scrutiny of an AI system throughout its entire lifecycle. This includes performance testing during procurement and contracting, as above, to ensure you're getting what you pay for, conducting regular algorithmic audits and AI impact assessments, and reviewing AI polices, integrating new industry standards and best practice as necessary. These assurance mechanisms should be linked to and grounded in the standards, principles, and governance mechanisms you've established as a firm.

A growing number of third-party firms can stress-test, verify and monitor your AI to identify its vulnerabilities and support its safe and ethical uptake. However, the cost of this support, as with procuring bespoke models, is significant. One provider estimated that for six weeks of support to an agency seeking to test and validate a model before procurement, the cost would be upwards of £135,000.

Yet these costs, along with the resources required to meaningfully consult with communities impacted by algorithmic use, are critical to ensuring the safe and responsible use of AI in crisis contexts.

Chapter 5 Additional sector-wide considerations

In addition to the roadmap above, considering the factors below would help enable a more ethical uptake of AI across the humanitarian sector.

Move slowly and test things

While AI might improve some elements of humanitarian action, the hundreds of millions of people in need of humanitarian assistance are some of the most vulnerable in the world. Start simply. Do not dive into complex AI waters with applications that involve service delivery, predicting population movements, and/or any personally identifiable information. There are safer and easier-to-use AI applications that can help agencies get their 'toes wet' while minimising harm and maximising safety to populations and their access to lifesaving support.

Stay laser-focused on humanitarian principles and ethics

Guidelines on responsible AI and corporate boilerplates on AI ethics abound. But the starting point for humanitarian AI should be our own humanitarian principles – humanity, neutrality, impartiality and independence – alongside key standards and commitments, such as those related to do no harm, accountability and localisation. Other commitments and principles related to AI specifically – such as transparency, explainability and accountability– can be added to these or further fleshed out with a humanitarian lens.

Commit to transparency and interagency co-operation

At the time of writing, only a handful of agencies appear to be working collectively to interrogate the harms and opportunities related to AI and humanitarian action. Many more are focusing on developing a single or several AI applications specific to their agency, ostensibly to gain a competitive edge in the pursuit of funding. But minimising harm and achieving the greatest impact for humanitarian AI requires collective action and dialogue on its potential harms. This is easier said than done when humanitarian funding is increasingly stretched and need is escalating. Thus, multiple approaches to cross-sector cooperation may be required.

For example, with new life breathed into it, the Data Science Ethics Group, or a similar forum, might offer a useful platform by which to share information about the AI applications that have worked and those that haven't, as well as AI suppliers and contexts into which AI systems are deployed. To improve information sharing, institutional donors and philanthropists could require their implementing partners to publicly disclose if and how they are using AI and algorithmic systems, to share AI audits and other reports that assess AI performance, and to demonstrate compliance with the EU's AI Act or other regulations, where applicable. This latter approach might be particularly relevant for European Civil Protection and Humanitarian Aid Operations (ECHO) and EU-based donors. Donors could also establish

procurement frameworks or other public platforms to help agencies procure solutions and services from pre-approved AI suppliers who comply with agreed terms and conditions, such as compliance with relevant standards and legislations and the humanitarian principles.

Develop a common standard for humanitarian AI

Preventing and mitigating the risks posed by humanitarian AI will require cross-sector, international cooperation and agreement on the governance of AI and on specific applications. Whilst requirements to comply with AI regulation, such as the EU's AI Act, will vary between agencies and the jurisdictions in which they operate, a common, sector-specific standard on the safe and responsible deployment of AI could help enable positive outcomes for communities directly impacted by AI and algorithmic decision-making.

Standards are developed by experts, grounded in globally-agreed evidence and best practice, represent common agreement on key technical terms, are measurable, are developed by consensus, and are regularly reviewed and updated. Humanitarian actors have relied on standards to strengthen crisis response, improve accountability, and to ensure crisis-affected communities are protected, receive assistance and enjoy the basic conditions for life with dignity.

A common standard on humanitarian AI would allow agencies to tailor AI to their specific needs whilst also enabling international cooperation and affording greater accountability to affected populations. It could draw inspiration from international AI standards developed by organisations – like the International Electrotechnical Commission, IEEE, the International Organization for Standardization, and NIST – whilst reflecting the humanitarian principles and dovetailing with existing humanitarian standards, like the Core Humanitarian Standard. Given their role in hosting the annual AI for Good Global Summit as well as the Global Standards Symposium, the UN's International Telecommunication Union (ITU) may have some role to play in supporting the creation of a common standard on humanitarian AI.

Additionally, international AI standards, like ISO/IEC 42001:2023 and ISO/IEC 23894:2023, may offer a particularly useful starting point for both aid agencies seeking certification in standards and wider discussions about a common humanitarian standard. Lauded as 'the world's first AI management system standard', ISO/IEC 42001 offers guidance on how to responsibly integrate an AI management system within the existing structures of an organisation while ISO/IEC 23894 offers guidance on assessing and managing risks specific to AI.

Treat humanitarian AI initiatives as a public good

Crisis-affected communities and the agencies working on their behalf should have real-time access to the outputs of humanitarian AI systems and reports on their technical reliability. This is particularly true in cases where life-and-death decisions are being made on the outputs of expert systems predicting the scale or location of humanitarian crises.

For example, it seems difficult to imagine a scenario where an INGO is asked to work to strategic and/or geographic priorities that are even partially informed by the outputs of an AI if that same INGO hasn't been afforded the opportunity to properly interrogate the model's governance framework and legal compliance, assessed the data used to train and run the model, or reviewed impact assessments or quality assurance reports on the technical reliability of the model.

Moreover, failing to make public the performance and outputs of humanitarian AI models not only risks strengthening the existing hierarchy that has arguably prevented the meaningful sharing of power between agencies in the Global North and the Global South, but might also increase the voice of private sector firms in humanitarian decision-making. This could shift discourse away from humanitarian principles and ethics and towards rhetoric related to efficacy and shareholder value.

Chapter 6 Conclusion

Dozens of agencies charged with delivering lifesaving services to millions of the world's most vulnerable populations are accelerating their exploration and experimentation with AI. This uptick in interest can partly be explained by the publicity surrounding generative AI tools, like ChatGPT, as well as the sector's steady adoption of innovation methodologies and its reliable interest in the 'shiny new object'. But it also stems from a desperate drive to find any and all workable solutions that could help meet escalating need set against a backdrop of stretched resources, waning adherence to IHL, and a rules-based international order in crisis.

Regrettably, the single-minded pursuit of AI use cases seems to be displacing vital debates and decisions related to AI ethics, governance, safety, accountability and humanitarian principles, as agencies seek to answer *if we can* adopt AI in humanitarian contexts rather than *how we should* adopt AI. Determining the latter requires a robust and pre-agreed approach to AI governance and safety grounded in a commitment to the humanitarian principles and equitable, power-sharing partnerships with communities.

The humanitarian sector, however, has a notorious allergy to anything but the mildest forms of governance. Even humanitarians' compliance with regulation like GDPR – arguably the most influential and consequential regulation on data privacy for a generation – has been patchy. But failing to put in place the necessary AI governance and safety measures, whilst also accelerating AI uptake, risks exacerbating extant, sector-wide trends related to surveillance humanitarianism, digital experimentation, and humanitarian extractivism. And these trends will only be exacerbated if we fail to challenge the 'semi-religious faith' in AI observed in other industries and exhibited by some of AI's staunchest enthusiasts.

As AI use cases increase across the humanitarian sector, in the Global North, the use of AI in the administration of public and social services is increasingly being challenged and support for the right to a human decision is growing. This is, in part, due to pressure and advocacy from civil society actors, journalists, and privacy activists, in places like the Netherlands, Poland, and the UK, who argue that claims about the positive benefits of efficiency and algorithmic optimisation belie the tendency of AI to 'to scale structural violence' and discrimination and conceal AI's necropolitical power – in other words, its power to dictate how people live and die.

The ability of crisis-affected communities to hold humanitarians to account in a similar fashion, if (or when) algorithmic outputs result in real harm to individuals or communities, will rest on the AI governance and accountability measures we deploy. It also requires a paradigm shift that prioritises human agency, autonomy and the principle of humanity above automation. Without sector-wide debates and agreement on a common approach to governing AI, it's hard to know when or if these measures will be adopted and whether the humanitarian principles will survive the Age of AI.

Appendix 1 Glossary

AI assurance 'The term "assurance" originally derived from accountancy, but has since been adapted to cover areas including cyber security and quality management. Assurance is the process of measuring, evaluating and communicating something about a system or process, documentation, a product or an organisation. In the case of AI, assurance measures, evaluates and communicates the trustworthiness of AI systems. [...] AI assurance processes can help to build confidence in AI systems by measuring and evaluating reliable, standardised and accessible evidence about the capabilities of these systems. It measures whether they will work as intended, hold limitations, and pose potential risks, as well as how those risks are being mitigated to ensure that ethical considerations are built in throughout the AI development lifecycle.' [Source: UK Government]

AI ethics 'AI ethics is a set of values, principles, and techniques that employ widely accepted standards to guide moral conduct in the development and use of AI technologies. These values, principles, and techniques are intended both to motivate morally acceptable practices and to prescribe the basic duties and obligations necessary to produce ethical, fair, and safe AI applications.' [Source: The Alan Turing Institute]

Algorithm 'An algorithm is a set of rules, a recipe, for a computer to follow.' An ML algorithm is only as good as its data because it uses data to establish the rules for how the algorithm functions, rather than a programmer establishing what the rules are. [Source: Oxford Sparks]

Artificial intelligence There is no one, universally accepted definition of artificial intelligence or AI. Broadly speaking, AI is the science and technology of creating intelligent systems. AI 'is often used to describe when a machine or system performs tasks that would ordinarily require human (or other biological) brainpower to accomplish, such as making sense of spoken language, learning behaviours or solving problems'. Put differently, 'an AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical (real) or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment'. There has been much publicity about the use of AI in decision-making, for example in the security and justice sectors. These AI are driven by ML tools, which have taught a computer to make decisions based on the data presented to it. AI systems are often enabled by ML and apply data-derived predictions to automate decisions. [Source: The Alan Turing Institute, OECD, and USAID]

Chatbot 'At the most basic level, a chatbot is a computer programme that simulates and processes human conversation (either written or spoken), allowing humans to interact with digital devices as if they were communicating with a real person. Chatbots can be as simple as rudimentary programmes that answer a simple query with a single-line response, or as sophisticated as digital assistants that learn and

evolve to deliver increasing levels of personalisation as they gather and process information. Driven by AI, automated rules, natural language processing and ML, chatbots process data to deliver responses to requests of all kinds.' [Source: Oracle]

Conversational AI 'Conversational AI refers to technologies, like chatbots or voice assistants, which users can talk to. They use large volumes of data, machine learning and natural language processing to imitate human interactions, recognising speech and text inputs and translating their meanings across various languages.' Humans typically interact with conversational AI through telephone calls, SMS or messaging platforms like WhatsApp or Facebook Messenger. Apple's Siri and Amazon's Alexa as well as a range of customer relations management tools are popular forms of conversational AI. [Source: IBM]

Data cleaning 'Data cleaning is the process of correcting and/or standardising data from a record set, table or database.' [Source: Centre for Humanitarian Data]

Data science 'Billions of gigabytes of data are generated globally every day. [...] Data science is the drive to turn this data into useful information, and to understand its impact on science, society, the economy and our way of life. The study of data science brings together researchers in computer science, mathematics, statistics, ML, engineering and the social sciences.' [Source: The Alan Turing Institute]

Data silo 'A data silo is a situation wherein only one group in an organisation can access a set or source of data. Data silos can result from several factors, including: cultural – competition or animosity between departments can cause employees to keep data from each other, rather than working together; structural – especially in large organisations, data silos can stem from a hierarchy separated by many layers of management and highly specialised staff; technological – applications might not be used or even designed to cross-reference or add to each other, or one department may simply not have access to a valuable app from another department because it was not purchased for their specific day-to-day tasks.' [Source: Plixer]

Evaluations 'Systematic assessments of an AI system's performance, capabilities, or safety features. These could include benchmarking tests, adversarial testing, or user feedback, amongst other methods.' [Source: UK Government]

Facial recognition Facial recognition is a 'software that maps, analyzes, and then confirms the identity of a face in a photograph or video'. [Source: The New York Times]

Forecast A 'prediction or estimate of future events and their expected impacts and consequences. It is the output of a predictive model.' [Source: Centre for Humanitarian Data]

GDPR 'The General Data Protection Regulation (GDPR) is the toughest privacy and security law in the world. Although it was drafted and passed by the European Union (EU), so long as they target or collect data related to people in the EU. The regulation was put into effect on May 25, 2018. The GDPR [levies]

harsh fines against those who violate its privacy and security standards, with penalties reaching into the tens of millions of euros.' It also has important clauses that link to the use of AI, in relevant jurisdictions, such as transparency, explainability, and fairness. [Source: GDPR.eu]

Guardrails Guardrails are 'pre-defined safety constraints or boundaries set up in an attempt to ensure an AI system operates within desired parameters and avoids unintended or harmful outcomes'. [Source: UK Government]

Hallucinations 'AI systems regularly produce plausible yet incorrect answers and state these answers with high confidence.' [Source: UK Government]

Humanitarian action/aid Humanitarian action is intended to 'save lives, alleviate suffering and maintain human dignity during and in the aftermath of man-made crises and disasters, as well as to prevent and strengthen preparedness for such situations'. [Source: Good Humanitarian Donorship]

Humanitarian principles Humanitarian actors are generally 'guided by four humanitarian principles: humanity, neutrality, impartiality and independence. These principles provide the foundations for humanitarian action. They are central to establishing and maintaining access to affected communities', whether in a disaster, an emergency or a situation of chronic conflict or instability. [Source: UN Office for the Coordination of Humanitarian Affairs]

Large language model (LLM) LLMs are a category of models trained on immense amounts of data which make them capable of recognising, understanding, and generating text and other content. LLMS provide the foundational capabilities needed to drive multiple use cases and applications, as well as resolve a multitude of tasks. [Source: UK Government and IBM]

Machine learning (ML) Machine learning (ML) brings together the fields of statistics and computer science to enable computers to learn how to do a given task, without being programmed to do so. ML uses a range of methods to train computers to learn from existing data, where 'learning' amounts to making generalisations about existing data, detecting patterns or structures, and making predictions for new data. This differs from how statistical analysis has traditionally been done, where a model is developed based on mathematical rules and then applied to data. ML approaches flip this process by finding patterns in data and returning a model that can make predictions for new data. [Source: Oxford Sparks and USAID]

Model drift 'Model drift, also called model decay, refers to the degradation of machine learning model performance over time. This means that the model suddenly or gradually starts to provide predictions with lower accuracy compared to its performance during the training period.' Model drift includes concept drift and data drift and can be detected with specialised tools and platforms to monitor model capabilities. [Source: AIMultiple]

Natural language processing Natural language processing (NLP) 'enables computers to read a text, hear and interpret speech, gauge sentiment and prioritise and connect to appropriate subjects and resources. The most familiar application is an automated call centre that sorts calls by category and directs the caller to recorded responses. NLP has become more widespread with voice assistants like Siri and Alexa. NLP uses machine learning to improve these voice assistants and deliver personalisation at scale.' [Source: GSMA]

Predictive analytics Predictive analytics involves the use of advanced analytic techniques, statistics and ML to analyse current and historical data in order to uncover real-time insights or anticipate an event or some characteristic of an event. Predictive analytics can support decision-making by examining data or content to answer the question 'What should be done?' or 'What can we do to make X happen?'. It is characterised by techniques such as graph analysis, simulation, complex event processing, neural networks, recommendation engines, heuristics and ML. Aid actors are using predictive analytics to identify trends or characteristics of future crises and events, including their probability, severity, magnitude and duration. Models have been developed to anticipate disease outbreaks and epidemics such as cholera, the movement of populations, changes to food security or extreme climatic events and disasters, such as floods. [Source: The Centre for Humanitarian Data and IBM]

Supervised learning 'Supervised learning, also known as supervised machine learning, is a subcategory of AI/ML. It is defined by its use of labelled datasets to train algorithms that classify data or predict outcomes accurately. [...] Supervised learning helps organisations solve [...] a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.' (*See also:* Unsupervised learning.) [Source: IBM]

Training data 'Training data is an extremely large dataset that is used to teach an ML model. [...] For supervised ML models, the training data is labelled. The data used to train unsupervised ML models is not labelled.' [Source: Techopedia]

Trustworthy AI Trustworthy AI refers to AI systems that embody an agreed set of principles or requirements that are typically values-based. For the OECD, trustworthy AI systems are those that respect human rights and privacy; are fair, transparent, explainable, robust, secure and safe; and are developed and used by actors who remain accountable for their actions. Others define trustworthy AI as systems that are legally compliant, ethically sound and technically robust. [Source: European Commission, Floridi et al., and the OECD]

Unsupervised learning 'Unsupervised learning, also known as unsupervised machine learning, uses ML algorithms to analyse and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Unsupervised learning's ability to reveal similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation and image recognition.' (*See also:* Supervised learning.) [Source: IBM]

Virtual assistant 'Data-driven and predictive (conversational) chatbots are often referred to as virtual or digital assistants. They are much more sophisticated, interactive and personalised than task-oriented chatbots. These chatbots are contextually aware and leverage natural language understanding, natural language processing and ML to learn as they go. They apply predictive intelligence and analytics to enable personalisation based on user profiles and past user behaviour. Digital assistants can learn a user's preferences over time, provide recommendations and even anticipate needs. In addition to monitoring data and intent, they can initiate conversations. Siri and Alexa are examples of consumer-oriented, data-driven, predictive virtual assistants.' [Source: Oracle]

Appendix 2 Helpful resources and online courses

Adversarial AI: fooling the algorithm in the age of autonomy. Fujitsu (2021), www.fujitsu.com/uk/imagesgig5/7729-001-Adversarial-Whitepaper-v1.0.pdf.

AI for humanitarians: a conversation on the hype, the hope, the future. Margffoy (2023), www.thenewhumanitarian.org/feature/2023/09/05/ai-humanitarians-conversation-hype-hope-future

Algorithms of war: the use of artificial intelligence in decision making in armed conflict. Stewart and Hinds (2023), https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict.

Armament, arms control and artificial intelligence: the Janus-faced nature of machine learning in the military realm. Reinhold and Schörnig (2022), https://doi.org/10.1007/978-3-031-11043-6.

Artificial intelligence and a pandemic: an analysis of the potential uses and drawbacks. Wiliams et al.(2021), https://doi.org/10.1007/s10916-021-01705-y.

Artificial intelligence and machine learning in armed conflict: a human-centred approach. ICRC (2020), https://doi.org/10.1017/S1816383120000454.

'Back to basics' with a digital twist: humanitarian principles and dilemmas in the digital age. Devidal (2023), https://blogs.icrc.org/law-and-policy/2023/02/02/back-to-basics-digital-twist-humanitarian-principles/.

Between surveillance and recognition: rethinking digital identity in aid. Weitzberg et al. (2021), https://doi.org/10.1177/20539517211006744.

Chatbots in humanitarian contexts. IFRC (2023), https://communityengagementhub.org/wp-content/uploads/sites/2/2023/06/20230623_CEA_Chatbots.pdf.

ChatGPT and large language models: what's the risk? NCSC (2023), www.ncsc.gov.uk/blog-post/chatgpt-and-large-language-models-whats-the-risk.

Deepfakes and international conflict. Byman et al. (2023), www.brookings.edu/articles/deepfakes-and-international-conflict/.

Detaining by algorithm. Deeks (2019), https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/.

Digital identity, biometrics and inclusion in humanitarian responses to refugee crises. Holloway et al. (2021), https://odi.org/en/publications/digital-identity-biometrics-and-inclusion-in-humanitarian-responses-to-refugee-crises/.

Disinformation endangering Red Cross work in Ukraine: ICRC. AFP (2022), www.france24.com/en/live-news/20220329-disinformation-endangering-red-cross-work-in-ukraine-icrc.

Displaced, profiled, protected? Humanitarian surveillance and new approaches to refugee protection, from *Postcoloniality and Forced Migration*. Ewert (2022), https://doi.org/10.56687/9781529218213-009.

Generative AI exists because of the transformer. Visual Storytelling Team and Madhumita Murgia (2023), https://ig.ft.com/generative-ai/.

Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks. Beduschi (2022), https://doi.org/10.1017/S1816383122000261.

How AI can actually be helpful in disaster response. Ryan-Mosley (2023), www.technologyreview.com/2023/02/20/1068824/ai-actually-helpful-disaster-response-turkey-syria-earthquake/.

Humanitarian AI: the hype, the hope and the future. Spencer (2021), https://odihpn.org/publication/humanitarian-artificial-intelligence-the-hype-the-hope-and-the-future/.

Humanitarian extractivism: The digital transformation of aid. Sandvik (2023), https://manchesteruniversitypress.co.uk/9781526173355/.

IBM Course: IBM, Introduction to Artificial Intelligence (AI). IBM, www.coursera.org/learn/introduction-to-ai.

IBM Course: Generative AI: Introduction and Applications. IBM, www.coursera.org/learn/generative-ai-introduction-and-applications?specialization=applied-artifical-intelligence-ibm-watson-ai.

In the age of AI, how do we scale digital opportunities and secure safer information landscapes for people caught in conflict? CDAC Network (2024), www.cdacnetwork.org/news/in-the-age-of-ai-how-do-we-scale-digital-opportunities-and-secure-safer-information-landscapes-for-people-caught-in-conflict.

Intro to LLMs and more. Google, https://developers.google.com/machine-learning/resources

Introduction to AI Assurance (e-training). CDEI and The Alan Turing Institute (2022), https://aistandardshub.org/training/introduction-to-ai-assurance/.

Israel offers a glimpse into the terrifying world of military AI. Tharoor (2024) www.washingtonpost.com/world/2024/04/05/israel-idf-lavender-ai-militarytarget/.

Lost in digital translation? The humanitarian principles in the digital age. Devidal (2024), www.cambridge.org/core/journals/international-review-of-the-red-cross/article/abs/lost-in-digital-translation-the-humanitarian-principles-in-the-digital-age/A4E233798E3AFA2516FEAA750E249AA3.

Of mosquitoes and models: tracking disease by satellite. Patel (2020), https://earthobservatory.nasa.gov/features/disease-vector.

Palantir demos AI to fight wars but says it will be totally ethical don't worry about it. Gault (2023), www.vice.com/en/article/qjvb4x/palantir-demos-ai-to-fight-wars-but-says-it-will-be-totally-ethical-dont-worry-about-it.

Romanes Lecture: 'Godfather of AI' speaks about the risks of artificial intelligence. Hinton (2024), www.ox.ac.uk/news/2024-02-20-romanes-lecture-godfather-ai-speaks-about-risks-artificial-intelligence.

The Cambridge handbook of responsible artificial intelligence: interdisciplinary perspectives. Mueller et al. (2022), https://doi.org/10.1017/97810092078 98.037.

The emerging threat of AI-driven cyber attacks: a review. Guembe et al. (2022), https://doi.org/10.1080/08839514.2022.2037254.

The future is now: artificial intelligence and anticipatory humanitarian action. Chen (2021), https://blogs.icrc.org/law-and-policy/2021/08/19/artificial-intelligence-anticipatory-humanitarian/.

'The Gospel': how Israel uses AI to select bombing targets in Gaza. Davies et al. (2023), www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

The role of artificial intelligence in achieving the Sustainable Development Goals. Vinuesa et al. (2020), https://doi.org/10.1038/s41467-019-14108-y.

The role of artificial intelligence in disinformation. Bontridder and Poullet (2021), https://doi.org/10.1017/dap.2021.20.

Valent Projects – news & insights. A digital agency working on mis- and disinformation, https://valent-projects.com/investigations-and-media/.

What are large language models? IBM (2024), www.ibm.com/topics/large-language-models.

Why AI is incredibly smart and shockingly stupid. Choi (2023), www.ted.com/talks/yejin_choi_why_ai_is_incredibly_smart_and_shockingly_stupid?language=en.