Standards and Assurance Framework for Ethical AI

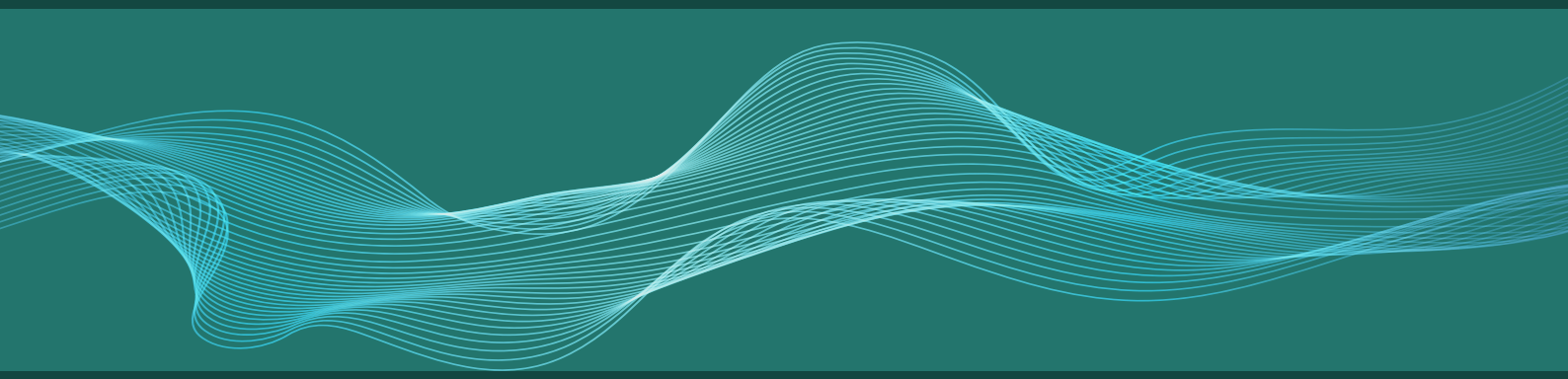
SAFE AI

A Governance Framework for Humanitarians using AI

How to turn SAFE AI principles into practical action

May 2026

Version 1.1



Helen McElhinney, Dr Anjali Mazumder, Dr Michael Tjalve,
Suzy Madigan and Sarah Spencer

CDAC NETWORK
communicating with disaster-affected communities

The Alan Turing Institute

Humanitarian AI Advisory

Foreword

Artificial intelligence offers genuine opportunities to strengthen humanitarian action – improving how needs are understood, how resources are targeted, how information reaches people in crisis. Realising that potential requires getting governance right before those decisions are locked in.

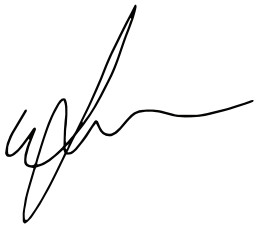
The UK government has led sustained international action on safe, ethical, and inclusive AI. From the Bletchley Declaration in 2023, through Seoul, Paris, and the New Delhi AI Impact Summit in 2026, the UK has anchored a global recognition that responsible governance must accompany deployment, and that the benefits of AI must reach beyond advanced economies. Through the Department for Science, Innovation and Technology AI Assurance Roadmap and the UK's investment in AI assurance infrastructure at home, we have demonstrated that responsible development and deployment of AI is not a constraint on innovation – it is what makes innovation trustworthy and scalable.

SAFE AI is the direct expression of those commitments in one of the highest-risk contexts in which AI is being deployed: humanitarian response. Funded by FCDO and developed by CDAC Network, The Alan Turing Institute, and Humanitarian AI Advisory, SAFE AI applies the logic of UK AI assurance leadership to contexts where the cost of failure is borne by people with the least power to challenge it.

We are proud to support SAFE AI. As the UK shifts its development approach to move from being a donor to an investor, and from service delivery to system support, governance infrastructure is precisely the kind of enabling architecture that makes other investments viable. AI is already influencing humanitarian targeting, eligibility decisions, fraud detection, and public communications in fragile contexts. Without a shared accountability infrastructure, those deployments carry risks for communities, for implementing agencies, and for the donors and governments backing them.

The SAFE AI Framework lays the foundation to reduce that risk. It provides practical, proportionate tools that help organisations move forward with confidence. It embeds community accountability as a governance requirement to be demonstrated. It creates the conditions for a shared standard to which every donor and implementing agency can align – reducing duplication, strengthening oversight, and protecting the long-term credibility of AI-enabled programming in the sector.

We call on fellow donors, UN agencies, and implementing organisations to engage with this framework, to align their own AI governance expectations with it, and to invest in the infrastructure that will allow it to evolve as a genuinely sector-owned standard. SAFE AI establishes the foundation. Independent assurance is what will make it enforceable. The value of a shared standard grows with adoption. This is the moment to begin.



Nathanael Bevan
Director for Global Tech
UK Foreign, Commonweath & Development Office

Will Hines
Humanitarian Director
UK Foreign, Commonweath & Development Office

Disclaimer

SAFE AI is a governance and assurance framework designed to support responsible and proportionate use of artificial intelligence (AI) in humanitarian contexts. It provides structured guidance, tools and illustrative examples to inform organisational decision-making.

SAFE AI does not constitute legal advice and should not be relied upon as such. Organisations remain responsible for ensuring compliance with applicable laws and regulations in the jurisdictions in which they operate. Independent legal advice should be sought where appropriate.

SAFE AI does not create binding obligations, confer certification status, or replace internal governance, risk management, or compliance processes. It is intended to support informed judgement and structured oversight, not to substitute for professional, legal, or regulatory assessment.

CDAC NETWORK
communicating with disaster-affected communities

UK International Development
Partnership | Progress | Prosperity

First published May 2026

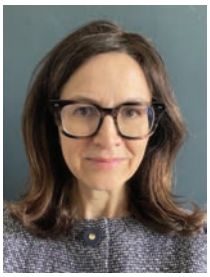
Registered charity number: 1178168
Company number: 10571501

This material has been funded by UK International Development from the UK government. However, the views expressed do not necessarily reflect the UK government's official policies.

Contents

About the authors

Helen McElhinney is Founding Architect of SAFE AI and Executive Director of CDAC Network, the global alliance for trustworthy information in crises. With twenty years in international affairs, she has worked for the ICRC and in UK government on crisis response, protection and security policy. She sits on the IASEAI Council and chairs its Working Group on the Right to Know. She holds a master's in International Relations from Cambridge.



Sarah Spencer is a co-founder of SAFE AI and was previously an independent AI researcher and consultant. She has spent a decade working on AI and its connection with global and national security. Sarah currently serves as Deputy Head of the International Tech Department in the UK's FCDO, is a Fellow at the Royal Society of Arts, and was in 2025's list 100 Brilliant Women in AI Ethics™. Sarah has a Master's degree from Harvard University in Public Policy and a Master's from Cambridge University in AI Ethics.

Dr Anjali Mazumder established the AI and Human Rights Theme and is Director of Accountability, Inclusion and Rights at The Alan Turing Institute, the UK's national institute for data science and AI. She is visiting Faculty at University of Cambridge, and Director of Artworthy Systems. She is an elected trustee of the Royal Statistical Society and holds a doctorate in Statistics from the University of Oxford.

Dr Michael Tjalve is Founder and Director of Humanitarian AI Advisory. He is Assistant Professor at the University of Washington and Co-Founder of RootsAI Foundation. He previously served as Chief AI Architect at Microsoft Philanthropies and as Senior Advisor to the UN on AI in humanitarian action. He holds a doctorate in Artificial Intelligence from University College London.

Suzy Madigan is Founder of The Machine Race, a Responsible AI consultancy. Named in the global 100 Brilliant Women in AI Ethics™ 2025, she was previously Responsible AI Lead at CARE International, where she led research with Accenture on inclusive AI collaboration between tech companies, governments and Global South civil society. She draws on 15 years' frontline crisis response for the UN, NGOs and government.

Introduction

What is the SAFE AI Framework?

AI has the ability to accelerate innovation and change humanitarian action for the better. Yet there is a real risk that an underfunded, overstretched humanitarian sector will accelerate towards unsafe use of AI to reduce costs, with serious unintended consequences for vulnerable populations.

As documented in *The Governance Gap in Humanitarian AI* (McElhinney, Mazumder and Tjalve, 2026), humanitarian organisations are increasingly deploying AI in contexts characterised by vulnerability, power asymmetry, and limited avenues for redress, without a sector-shared operational framework that translates global governance norms into safeguards appropriate for these environments.

This framework is a direct response to that governance gap.

In this critical space, CDAC Network, The Alan Turing Institute, and Humanitarian AI Advisory have created the Standards and Assurance Framework for Ethical Artificial Intelligence (SAFE AI) – funded by the UK Foreign, Commonwealth & Development Office (FCDO). SAFE AI is not a substitute for existing organisational duties (e.g., safeguarding, data protection, security risk management). It is intended to strengthen and connect those duties to AI-specific risks. It sets out the standards and tools humanitarian organisations need to deploy AI with confidence and accountability, with humanitarian principles as the governance foundation and community in the loop embedded as a lifecycle requirement.

SAFE AI also operationalises an emerging global governance norm across the AI lifecycle: that people whose lives are shaped by AI-influenced decisions have a right to know when automation is in play, a right to understand how it is shaping decisions about them, and a right to contest those decisions. This right is held most acutely by people in the Global Majority contexts where humanitarian AI is deployed and where the governance infrastructure to enforce it is weakest. This right operates at both individual and collective level. Where multiple organisations deploy AI in the same context, communities have a right to know whether the system being used on them by one organisation is held to the same standard as the system being used by another.

What SAFE AI is not

- SAFE AI is not a data governance framework. It governs what AI systems do with data, how they shape decisions about people and whether those decisions can be explained and challenged.
- SAFE AI is not an AI literacy course. It provides targeted guidance and templates to support safe implementation decisions.
- SAFE AI is not an ethics framework or a self-assessment process. It is the operational layer for humanitarian AI assurance.
- SAFE AI is not a one-size-fits-all compliance checklist. It is designed to be risk-based and proportionate (based on your SAFE AI risk tier, see below).
- SAFE AI is not a menu of optional tools. Its components are designed to work as a coherent assurance architecture and should be applied through the SAFE AI journey and tiering logic set out in this document.

SAFE AI is designed to dock into existing organisational governance rather than sit alongside it. The framework provides AI-specific instruments, including the tier logic, the SAFE AI Impact Assessment and Transparency Card, that connect AI risk to the duties, documentation, and decision structures humanitarian organisations already have.

SAFE AI's principles and approach are grounded in CDAC Network's expertise in humanitarian communication, Accountability to Affected People (AAP), and information integrity in crisis settings. SAFE AI is also rooted in CDAC Network's mission to shift decision-making power towards people affected by crises and the local and national actors closest to them. The framework is designed to make that localisation and power shift practical in the AI lifecycle. CDAC stewards the framework on behalf of the sector, ensuring it remains anchored in humanitarian principles, participation, and accountability rather than abstract technical performance alone.

SAFE AI is designed to help organisations identify, manage, and document AI-related risks in humanitarian action. Our tiering system will help you categorise the risk and governance level to your AI use case. Some might be concerned with the safe use of personal or commercial AI tools by staff, whereas others could be designing organisation-wide AI systems closely interacting with communities affected by crisis. Your 'tier' will guide you to relevant sections of the SAFE AI Framework.

At cdacnetwork.org/safe-ai, you will find, practical tools and guidance to check if AI systems are:

- Fair, trustworthy, and in line with humanitarian values and principles.
- Legally and organisationally compliant.
- Accountable to communities affected by crisis, ensuring they can participate and have a real say in how AI is used.

The SAFE AI Transparency Card is the central living document and a core output of the SAFE AI Framework. This is where you record decisions, risks, and safeguards as you progress through the SAFE AI journey. The Card is the practical instrument through which the right-to-know principle is operationalised, making AI deployment legible to communities, colleagues, and donors.

By the end of this process, you will be equipped to:

- Deploy AI in humanitarian contexts with a structured, defensible governance approach.
- Engage constructively with AI proposals from suppliers, partners, and colleagues by asking the right questions and understanding the answers.
- Explain and defend decisions about AI, including decisions to proceed, pause, redesign, scale or refuse deployment.

Who is the SAFE AI Framework for?

The SAFE AI Framework is designed to support humanitarian organisations on their AI journey, and navigate the risks raised. Organisations procuring AI systems or designing AI projects will benefit from each tool within this framework, which is intended as a practical approach to using, designing, procuring, deploying, or scaling AI-enabled systems in humanitarian action.

It helps humanitarian actors identify, reduce, and document socio-technical risks across the AI journey, enabling proportionate innovation aligned with humanitarian principles, the Core Humanitarian Standard, and IASC – the Interagency Standing Committee Accountability to Affected People commitments.

On a day-to-day basis, the SAFE AI Framework is designed to support senior operational and governance decision-makers within humanitarian organisations. The outputs of SAFE AI processes are intended to provide a defensible basis for decisions on whether an AI system should proceed, be modified, be paused, or be decommissioned.

SAFE AI is for organisations at different stages of AI governance maturity

Humanitarian organisations sit at different stages of AI governance maturity, and SAFE AI is designed to be useful at each.

- For **organisations with mature AI governance frameworks already in place** – large UN agencies, major INGOs, and others that have invested in dedicated AI ethics structures, policies, and review processes – SAFE AI's primary value is comparability. Internal AI governance, however sophisticated, cannot produce evidence in a form consistent with what other organisations are producing. These organisations should run through SAFE AI once at the outset to test their existing AI governance for blind spots and to identify any AI-specific gaps the framework surfaces. Thereafter, the SAFE AI Transparency Card is the mechanism through which deployment evidence becomes comparable across the sector. SAFE AI is the additional sector-facing record produced through existing governance processes.

- For **organisations with established general governance but without AI-specific frameworks** – most national NGOs, many INGOs, and partners operating at scale in humanitarian response – SAFE AI extends existing governance to cover AI risk. An organisation's data protection policy, risk management framework, procurement function, and AAP infrastructure already do important work. SAFE AI provides the AI-specific lens applied through these functions, naming the AI-specific contractual conditions, risk types, and assessment questions that traditional frameworks may not yet cover.

- For **organisations at the start of building their AI governance** – local and national organisations, smaller INGOs, and others encountering AI deployment without prior governance infrastructure – SAFE AI provides the operational architecture from the ground up. The journey, the tools, and the cumulative governance records are all there to be used.

The framework offers value commensurate to the organisation's starting point. SAFE AI is designed to lift the floor without lowering the ceiling.

SAFE AI improves through use. Organisations applying the framework, regardless of governance maturity, are encouraged to feed back where the standard could be sharper, where gaps emerge in practice, or where adaptations have proved useful. SAFE AI is a living standard. The Transparency Cards produced under Phase 1 and the experience of organisations applying the framework form the evidence base for v1.1 and Phase 2 to follow.

The *Governance Gap in Humanitarian AI,* published in March 2026, also identifies pooled and pass-through funding mechanisms as a high risk gap: these channels are closest to affected communities and furthest from the governance requirements that bilateral funding agreements sometimes carry. Encouraging donor adoption of SAFE AI as an expected standard is the mechanism for closing that gap – see Closing the governance gap, below.

SAFE AI is designed to be compatible with emerging United Nations (UN) and Inter-Agency Standing Committee (IASC) approaches to federated data governance, inclusive standards-setting, and privacy-by-design architectures.

SAFE AI is built on the assurance pattern established in other regulated sectors: externally-set standards, evidence produced in audit-ready form, and independent verification by a body separate from the organisation being assessed. It is the pattern that gives aviation, finance, food safety, and pharmaceuticals their credibility, and it is the pattern AI assurance is moving toward globally.

How SAFE AI differs from self-assessment

SAFE AI is being developed in two phases. Phase 1 is the governance framework you are reading now, in which humanitarian organisations apply SAFE AI to their own deployments. The standards are externally set and structured for independent audit. In Phase 2 we aim to establish that independent audit function.

In a self-assessment model, an organisation defines what good looks like and produces the evidence its own definition requires. The exercise can be rigorous, but the standard against which it is measured is internal. SAFE AI is built differently. Evidence is produced in a form designed to be inspected by someone other than the organisation that produced it, and against the same criteria as other organisations doing similar work. Having sector-level standards mean deployments can be compared across organisations, learned from across the sector, and held to a consistent measure by donors and crucially by the affected communities AI systems are supposed to serve in humanitarian efforts.

Phase 1 puts those audit-ready artefacts in place. Phase 2, planned for investment, establishes the independent audit function that turns the architecture into operational assurance. The framework you are reading now is the foundation for that assurance, not assurance itself.

I'm new to AI – is this for me?

- You do not need to be an AI specialist to use the SAFE AI Framework. An overview of what AI is and how different types of AI introduce risks in humanitarian contexts can be found on the SAFE AI website.
- Throughout, we signpost optional, supportive training and learning resources to strengthen AI literacy. Some technical terms are used in this document and the framework's tools, but simple definitions are provided in the SAFE AI Glossary on the SAFE AI website.

Just as AI is a tool that can be used for a multitude of tasks, so implementing AI projects also requires a combination of skillsets from within an organisation. Applying the SAFE AI Framework requires a cross-departmental approach, preferably within a SAFE AI Project Team, uniting colleagues who bring expertise and experience in technology, procurement, programmes, compliance, safeguarding, and community engagement.

As you go through your own AI implementation journey, SAFE AI offers tools and guidance for different members of your multi-disciplinary team at different points.

What SAFE AI expects of humanitarian organisations

- Assign clear internal accountability for AI-related risks, rather than delegating responsibility to suppliers or technical partners.
- Use SAFE AI early, before significant design or procurement decisions are locked in.
- Secure right-to-audit clauses in AI procurement contracts, proportionate to the SAFE AI tier, so that the deployed system can be audited throughout its operational life and after termination.
- Use SAFE AI's outputs to inform real decisions, including whether to proceed, redesign, pause, or stop an AI use case.
- Revisit SAFE AI assessments when material changes occur to the system or context.
- Make AI use legible to affected populations. Where AI shapes decisions about people, those people should know that automation is in play and have routes to challenge what it decides.

SAFE AI and humanitarian principles

Humanitarian principles provide the ethical and operational foundation for SAFE AI. However, applying these principles to AI systems requires interpretation, as AI can introduce risks that are relational, indirect, distributed, or difficult to observe or verify, even when intentions are humanitarian. Given the speed of evolution of technology, this requires regular consideration.

SAFE AI applies the principles as practical decision criteria to identify when an AI system could undermine trust, access, safety, or humanitarian legitimacy. SAFE AI expects teams to explicitly assess whether an AI system could undermine the humanitarian commitment to do no harm, operationalised through the principles of:

- **Humanity**: The principal motivation behind humanitarian action is to save lives and alleviate suffering while upholding and restoring personal dignity.
  - AI systems can cause harm even when they improve efficiency. SAFE AI focuses on the cost of error, exposure to harmful outputs (including misinformation), and the minimum human oversight required to protect dignity and wellbeing. Cost of error is explored further in Step 2.1, Risk identification and mitigation, below.
- **Impartiality**: Humanitarian action should be based solely on need, with priority given to the most urgent cases irrespective of factors such as race, nationality, gender, religious belief, political opinion, or class.
  - AI can introduce exclusion through language barriers, data gaps, connectivity, or proxy variables. SAFE AI requires teams to consider whether AI use could create unequal access to assistance or information, and whether such effects can be detected and corrected.
- **Neutrality**: The neutrality of humanitarian action is upheld when humanitarian actors refrain from taking sides in hostilities or engaging in political, racial, religious, or ideological controversies.
  - Particularly in conflict-affected settings, AI systems may create perceived or real

alignment with parties to conflict e.g. through data sources, infrastructure, or partnerships. SAFE AI requires such trade-offs to be explicitly recognised, justified, and mitigated.

- **Independence**: This requires humanitarian actors to be autonomous. They are not to be subject to control, subordination, or influence by political, economic, military or other non-humanitarian objectives.

  - Reliance on technology suppliers, platforms, or data providers can shift control away from humanitarian organisations. SAFE AI emphasises maintaining internal decision authority, including the ability to pause, modify, or decommission AI systems.

The SAFE AI Impact Assessment and SAFE AI Transparency Card will help you explicitly document if your AI application is aligning with humanitarian principles. If your context or risks change, be sure to revisit your SAFE AI Transparency Card, with SAFE AI Technical Assurance, and ongoing monitoring.

Your humanitarian principles check will:

- Identify non-negotiable red lines.
- Document mitigation and residual risk.
- Support explicit decisions on whether an AI system should proceed, be redesigned, paused, or stopped on principled grounds.

Keeping affected communities in the loop

The tools in the SAFE AI Framework help humanitarian organisations to design and use AI systems in ways that are proportionate, meaningful, and accountable.

The framework embeds participation across the AI lifecycle – what we call keeping communities in the loop – and is key to ensuring your use of AI respects humanitarian principles.

The SAFE AI Framework acknowledges that not all uses of AI will require the same level of community participation – but recognises that participation is part of risk mitigation and quality control as communities themselves are 'domain experts'. Where AI systems shape people's access to information, assistance, or protection, those people hold the right to know how the systems are being used and to shape that use.

AI systems also carry the bias of their training data into every deployment. The communities affected need to be in the governance loop to see and challenge what those systems produce. That is why community-in-the-loop is a technical requirement of the architecture.

The SAFE AI co-design guidance supports Accountability to Affected People (AAP) principles, and enables your AI implementation to be more equitable and effective.

The SAFE AI User Guide

Mapping out your SAFE AI journey

This section is intended as practical a tool to help you navigate the SAFE AI Framework. It shows you how to:

- Understand the stages you should pass through in your SAFE AI journey, and the steps you should take in each stage.
- Select which SAFE AI tools apply to your specific AI use case, based on its risk level.
- Identify the level of governance you need to adopt.
- Decide when to revisit your past assessments.
- Hold to SAFE AI standards even in a crisis.

SAFE AI aligns its tools to four stages:

Stage 1: Problem definition and concept
Stage 2: Design
Stage 3: Development
Stage 4: Deployment, monitoring and evaluation

Each stage contains several steps, in which you will use one or more of the SAFE AI tools and processes we've developed. Some are straightforward enough to include in the main body of this document; others are more in depth and have been collected in the SAFE AI Tools and Guidance section of the SAFE AI website. Which tools you need will depend on which risk tier your AI use case falls into.

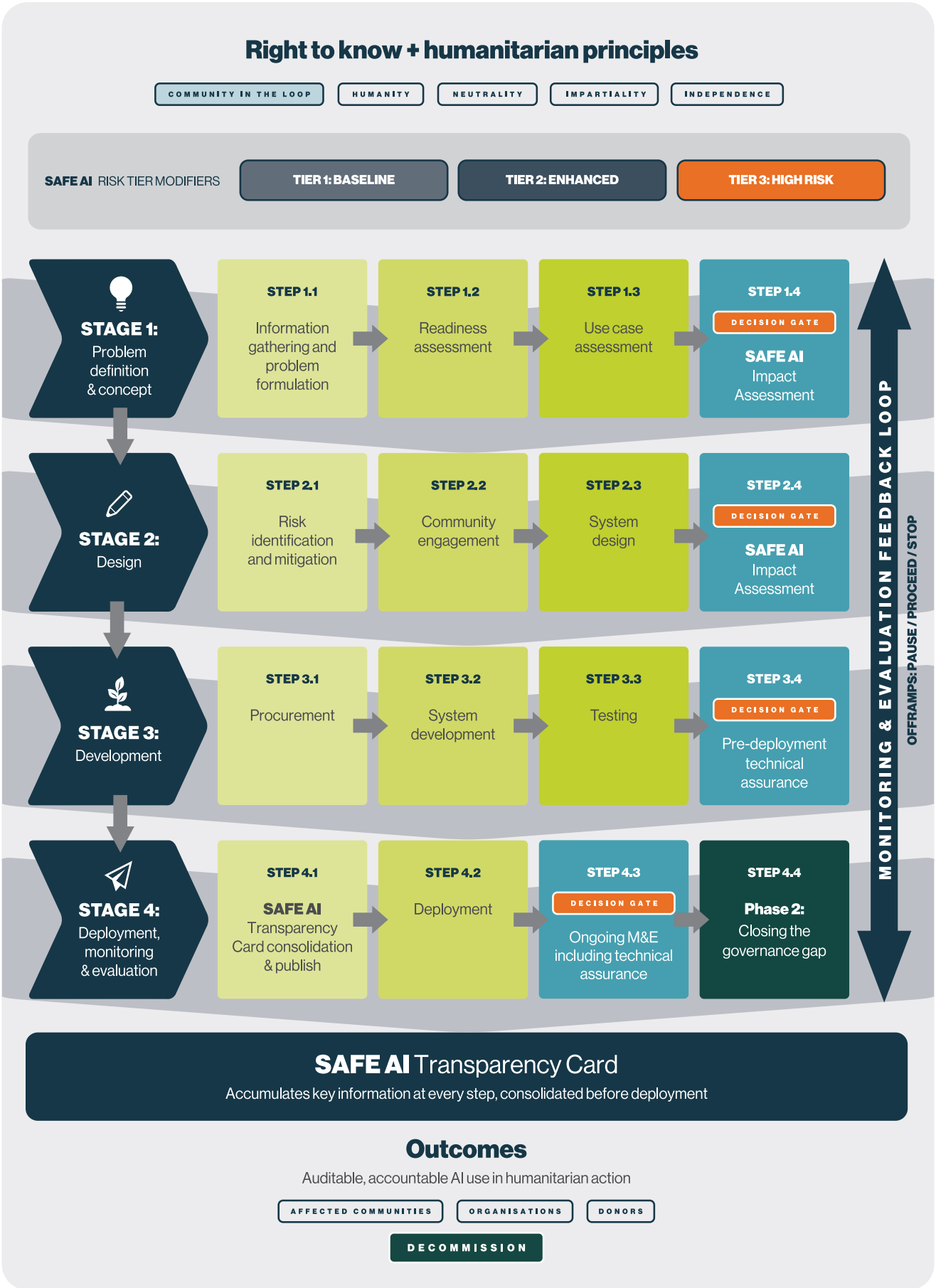
As we introduce each tool, you will find a quick overview of what it is, who it is for, and when and how to use it. The journey and its tools have all been developed with established humanitarian principles and community in the loop thinking as their foundation.

The outputs of these stages, steps and tools will inform your SAFE AI Transparency Card - the central record of how your AI use case is designed, governed, deployed and monitored, and of the decisions, risks and safeguards identified along the way. You should update it continuously as you progress.

Your SAFE AI Transparency Card is the main output of the framework and the place where the different parts of your assessment come together. As you complete it over the course of your SAFE AI journey, it creates a clear, shareable account of what you decided, why, who was involved, and how risks will be managed – making it easier to explain and justify your approach to colleagues, partners, affected communities and donors, and to revisit your reasoning if needed.

We strongly recommend working through each stage and step in order. Skipping one may mean missing considerations that could affect your use case's risk profile or implementation effectiveness. That said, some activities will run in parallel, some are iterative, and some will require revisiting earlier steps – a risk mitigation strategy, for example, may mean going back to redesign the system.

The SAFE AI Framework



There's an emergency! How long will all this take?

Designing, developing, and deploying AI systems takes time. However, there are tools and guidance on the SAFE AI website to give you 'quick wins' on the more responsible use of AI, such as how to develop an AI policy, and 'dos and don'ts' for staff. The framework is particularly suited for crises where humanitarian operations are already up and running. It can also support organisations plan for future rapid onset crises where AI approaches might be proposed, particularly in early warning and anticipatory action for example.

Finding the right way forward

As with much in the humanitarian sector, real world AI implementations do not tend to follow a linear route. Organisations rarely begin AI work at the same point. Some will encounter SAFE AI while exploring organisation-wide readiness or establishing the minimum conditions to use AI safely. Others will engage when a specific project, supplier, or tool is already under consideration.

At the same time, while the SAFE AI journey may look linear, activities can happen in parallel, and while SAFE AI applies to all humanitarian uses of AI, you don't need to use every SAFE AI approach for every use case.

How new AI models affect SAFE AI

When using AI within any organisation, it's important to remember how quickly the technology is evolving. This is a rapidly emerging technology. AI tools and capabilities will appear that change the risk (and opportunity) landscape overnight. Teams may – and should – revisit earlier steps of the AI framework to ensure their choices still stand as their needs, circumstances, risk profile, or tech options change.

Choosing the right level of SAFE AI governance for you

SAFE AI recognises that not every use of AI carries the same level of risk – and that humanitarian organisations have very different levels of time, capacity and technical support available. A team using AI to summarise internal notes does not need the same level of governance, evidence, and review as a team deploying a community-facing tool that may affect people's access to information, assistance, or protection.

Through the SAFE AI tiers, governance effort scales to actual risk, making adoption realistic for under-resourced organisations rather than an additional burden. SAFE AI therefore uses three tiers to help you decide how much of the framework to apply. The broad stages of the SAFE AI journey remain the same, but your tier determines which tools and activities are relevant, how much documentation and consultation is needed, and when stronger review or technical assurance may be required.

Consider both the inherent risk of what the technology does and how and where it is used. If, for example, AI is used to translate a report into a language you can read, that will generally be in the lower-risk tier, whereas if you use real-time, on-the-ground spoken language translation in medical triage, an AI error can lead to fatalities.

You are likely to move between tiers as your understanding of the use case develops. Where there is uncertainty, apply the higher tier to minimise risk.

At a glance: SAFE AI use case risk tiers

SAFE AI uses tiering to ensure governance is proportionate to risk and context, and to guide you to the appropriate SAFE AI tools to use for your needs and situation.

Local and national organisations often hold the most granular community-level data and often have the least governance infrastructure to manage AI.

SAFE AI is designed to be accessible and proportionate for organisations of all sizes.

The tiering logic scales requirements to risk, not to organisational capacity

Whatever your AI initiative, you will need some level of safeguards. As you begin on your SAFE AI journey, apply a provisional tier to your AI use case. This may change as you progress through the framework, but should reflect:

- The degree of influence the system has over decisions affecting people.
- Whether it is community-facing or internal.
- The potential impact of, or harmful response to, generated content.
- The potential cost of error (especially for autonomous systems) or data leakage.

The sensitivity of the operating context (e.g., conflict, displacement, adversarial information environment, protection risks, lack of humanitarian access)

Tier	What kind of use case	Typical journey time	SAFE AI Governance and tools required
TIER 1: BASELINE	Outputs are advisory and easily reversible. Does not directly influence decisions affecting people. Primarily internal use.	1 to 2 days	Governance: Minimum organisational safeguards in place SAFE AI tools: Impact Assessment (Section 1) and SAFE AI Transparency Card (Lean)
TIER 2: ENHANCED	Influences prioritisation, targeting or operational decisions. Human oversight is retained, but AI shapes choices. May use community data.	1 to 2 weeks	Governance: Targeted 'community in the loop' co-design guidance applies SAFE AI tools: Impact Assessment (Full), SAFE AI Transparency Card (Full) and Technical Assurance
TIER 3: HIGH-RISK	Directly affects people's access to assistance, protection or information. Community-facing or automated at scale. High cost of error.	Timeline will vary / ongoing	Governance: Ongoing monitoring and independent specialist support likely required SAFE AI tools: Impact Assessment (Full), SAFE AI Transparency Card, Technical Assurance, full 'community in the loop' co-design guidance becomes mandatory

For more detail on the characteristics of each SAFE AI tier, see the Tools and Guidance section of the SAFE AI website.

When to reassess your SAFE AI tier

Where there is uncertainty, apply the higher tier by default.

Also reassess the tier – and any previous SAFE AI tool outputs – if:

- The AI model or system is updated.
- Data sources, providers, or governance arrangements change
- Suppliers or supply chain dependencies shift materially
- Outputs are scaled, automated, or reused beyond the original scope
- The deployment context becomes more sensitive
- You observe performance degradation, new failure modes, complaints, safeguarding concerns, or protection risks.

In practice: Tiers and the SAFE AI journey

Regardless of where you enter SAFE AI's four-stage journey, organisations should follow the framework's logic to minimise risk.

This may mean that even for a Stage 3 Deployment initiative, you need to review earlier-stage tools or resources to understand your likely risk tier.

As you progress, new discoveries may also require you to reassess your provisional tier and revisit earlier steps to move forward safely.

Real-world AI deployments rarely follow the sequence described in the SAFE AI framework without interruption. A security breach may surface mid-design; a mass displacement event may silently invalidate a deployed model's training data; a cybersecurity incident can easily become a data protection crisis, a community trust breakdown, and an operational emergency simultaneously.

SAFE AI's governance architecture – the risk tiers and mitigation strategies, the decision gates, the accountability requirements, the 'community in the loop' rights to consultation and veto, the sign-off and override protocols – is designed to function in these conditions, not only in controlled ones.

The iterative structure of this framework is not a theoretical preference. It reflects what responsible governance actually requires when the unexpected occurs.

SAFE AI and incident management

When an AI system causes or contributes to harm, the response requires a defined cross-functional process, not ad hoc decisions.

In a crisis, SAFE AI tools will need to be revisited – to provide guidance on how you can respond more effectively, not to slow down your reaction.

Before deployment – and to ensure they're ready for the next incident before it happens – organisations should establish an incident response protocol within their organisational AI policy that specifies:

- Who is responsible for identifying and escalating incidents.
- What constitutes a reportable incident.

What triggers a system pause or withdrawal.

- How affected populations will be notified and supported.
- How the incident is documented for accountability and learning purposes.
- To relevant data protection and AI regulatory authorities in the jurisdictions of the organisation and of the affected population, and to law enforcement where an incident involves suspected criminal conduct or safeguarding concerns.

This protocol should be in place before any Tier 2 or Tier 3 AI system is deployed.

At a glance: SAFE AI journey stages, steps, tools, and outputs

Stage 1: Problem definition and concept

Step	What happens	Main SAFE AI tool(s)	Key output(s)	Where you document this in your SAFE AI Transparency Card
1.1 Information gathering and problem formulation	▪ Define the humanitarian problem ▪ Document if AI is the appropriate response	▪ SAFE AI Use Case Assessment Checklist ▪ Co-design guidance (part 1)	▪ Plain-language use case description ▪ Documented rationale for using AI	**Section 1:** Purpose, scope, benefit **Section 2:** Context and initial humanitarian alignment
1.2 Organisational readiness	▪ Assess governance, capacity, and accountability structures	▪ SAFE AI Readiness Checklist (parts 1–2) ▪ Tiers guide	▪ Go/no-go on readiness ▪ Governance gaps identified ▪ Tier assigned ▪ Accountable leads named	**Section 1:** Organisational context, accountable leads **Section 5:** Oversight roles including a single responsible owner (SRO)
1.3 Use case assessment	▪ Confirm readiness to proceed: data, budget, legal, community engagement	▪ SAFE AI Readiness Checklist (part 3) ▪ Impact Assessment (part 1) ▪ Co-design guidance	▪ Readiness confirmed ▪ Initial risk screening completed	**Section 1:** Data sources, dependencies **Section 2:** Initial risk and impact summary
1.4 SAFE AI Impact Assessment	▪ Formal assessment of the proposed AI use case, risks, governance conditions, and organisational readiness before moving into design	▪ SAFE AI Impact Assessment	▪ Completed impact assessment ▪ Documented go / no-go decision ▪ Identified conditions, mitigations, or redesign requirements	**Section 1:** System snapshot and rationale **Section 2:** Initial risks and impact **Section 5:** Governance and accountability

Stage 2: Design

Step	What happens	Main SAFE AI tool(s)	Key output(s)	Where you document this in your SAFE AI Transparency Card
2.1 Risk and mitigation	■ Identify risks, assess cost of error, set mitigations and red lines	■ SAFE AI Risk Mitigation Strategies ■ Impact Assessment (Section 2, Tiers 2–3)	■ Completed impact assessment ■ Red-line decisions documented	**Section 2:** Risk, impact, and mitigation summary
2.2 Community engagement	■ Involve affected communities in shaping system design	■ SAFE AI Co-design guidance (Stage 2/3)	■ Documented engagement record and insights to inform stage 3 ■ Clarity on how community will feedback	**Section 2:** Community engagement **Section 5:** Oversight, redress pathways
2.3 System design	■ Choose AI approach, architecture, hosting, data governance	■ SAFE AI Architecture and Tech Stack Decision Guide	■ Architecture decisions with key trade-offs identified and documented	**Section 1:** System stack, suppliers **Section 3:** Data sources, model design
2.4 SAFE AI Impact Assessment	■ Formal review of the proposed system design, identified risks, safeguards, community engagement, and governance mechanisms before development begins	■ SAFE AI Impact Assessment	■ Updated impact assessment ■ Documented design-stage decision ■ Confirmed mitigations, safeguards, and escalation requirements	**Section 2:** Risks and mitigation **Section 3:** Data and model summary **Section 5:** Governance and oversight **Section 6:** Decision record and versioning

Stage 3: Development

Step	What happens	Main SAFE AI tool(s)	Key output(s)	Where you document this in your SAFE AI Transparency Card
3.1 Procurement	■ Secure governance obligations contractually ■ Evaluate suppliers	■ SAFE AI Procurement Guide	■ Contract with governance obligations embedded including those from the community	**Section 1:** Supplier dependencies. **Section 4:** Supplier monitoring. **Section 5:** Supplier accountability

Step	What happens	Main SAFE AI tool(s)	Key output(s)	Where you document this in your SAFE AI Transparency Card
3.2 System development	■ Develop or configure ■ Document baseline	■ Standard software development best practice	■ Working system with documented baseline	**Section 3:** Model version, data, configuration **Section 4:** Timeline
3.3 Testing	■ Test against KPIs ■ Community validation where applicable	■ Standard software development best practice ■ Co-design guidance (validation)	■ Test results ■ Proceed / redesign / halt decision	**Section 3:** Test results **Section 2:** Updated risk summary
3.4 Technical assurance (pre-deployment)	■ Formal verification and sign-off before deployment	■ SAFE AI Technical Assurance	■ Deployment sign-off ■ SRO named ■ Performance baseline documented	**Section 4:** Pre-deployment verification **Section 5:** Roles confirmed **Section 6:** Sign-off record

Stage 4: Deployment, monitoring and evaluation

Step	What happens	Main SAFE AI tool(s)	Key output(s)	Where you document this in your SAFE AI Transparency Card
4.1 SAFE AI Transparency Card consolidation and publish	■ Finalise, approve and publish your Transparency Card ■ Confirm community access	■ SAFE AI Transparency Card	■ Published SAFE AI Transparency Card ■ Monitoring baseline confirmed	**All sections**
4.2 Deployment	■ System live ■ Community notification ■ Monitoring activated	■ SAFE AI Transparency Card ■ Standard software development best practice	■ Card updated ■ Monitoring activated ■ Community informed	**Section 4:** Deployment date **Section 5:** Feedback mechanisms, redress
4.3 Ongoing monitoring and evaluation including technical assurance	■ Ongoing performance monitoring ■ Community feedback ■ Incident response	■ SAFE AI Transparency Card ■ Technical Assurance (ongoing)	■ Continued performance evidence or reassessment triggers	**Section 4:** Monitoring outputs. **Section 5:** Feedback, redress. **Section 6:** Version history

Protecting the people you serve – and your organisation

Every decision you make on your SAFE AI journey – whether to proceed, pause, redesign, or stop – has consequences for the people your organisation serves.

The accountability challenge in humanitarian AI is structural, as we found in our March 2026 report *The Governance Gap in Humanitarian AI*. The populations most affected by AI-influenced decisions – those whose access to assistance, protection, or information is shaped by automated systems – are typically the furthest from the institutions deploying those systems and have the fewest mechanisms to question or correct them.

SAFE AI's documentation and redress requirements are a direct response to this asymmetry. The SAFE AI tools put a lot of emphasis on documentation to make those decisions visible, traceable, and defensible. Work through each step of the journey, use the relevant tools, record the relevant output in your SAFE AI Transparency Card, and you have a clear account of what you decided, why, and on what basis. You can then share this with communities, colleagues, partners, donors, or regulators, depending on what's required. This also provides a means of retaining knowledge during staff absences, RnRs or turnover.

Increasingly, donors are also keen to understand not just what humanitarian organisations are doing, but how they are doing it. SAFE AI can and should be used to support ongoing dialogue and judgement across the funding cycle. It can assist donor decision-making and risk management across policy, programme, and oversight functions.

What about individual staff using commercial AI tools?

The most common use of AI within any organisation is individuals using private AI accounts, such as ChatGPT, Claude, or Gemini. These tools can support people in their work and are particularly helpful for staff navigating a complex humanitarian system that often demands reports and funding applications in e.g. English or French.

Organisations are not limited to commercial providers at this layer: locally-hosted, open-weight, and frugal AI tools are also available for staff productivity uses, and may be more appropriate where data sensitivity, sovereignty, or sustainability is a concern.

However, commercial AI tools also present real risks in humanitarian settings, particularly if people upload sensitive data, trust outputs without verification, or when decisions taken are influenced by misleading information in AI outputs.

For organisations considering policies on individuals' use of private AI accounts and how to mitigate risks associated with generative AI, SAFE AI tools and guidance such as the SAFE AI Onboarding Readiness Checklist, SAFE AI Risk and Mitigation Strategies, and SAFE AI Impact Assessment can all be useful.

Connecting AI governance to existing organisational structures

Most humanitarian organisations already have the governance structures responsible AI deployment needs. SAFE AI is designed to plug into these as much as possible.

- **Operational AI work**, the work of building and deploying the system, sits with a multidisciplinary project team drawing on existing programme, technical, data protection, safeguarding, and community engagement functions.
- **Organisational oversight** of that work, including sign-off at Decision Gates and authority to pause or stop deployment, is governance work that most organisations already have functions for: risk committees, digital or data governance committees,

ethics review boards, programme quality assurance teams, or senior leadership groups. The function does not need to be AI-specific, although you may want to establish one depending on tier. (While we don't prescribe, for the purposes of SAFE AI tools, this is referred to as the 'oversight body'.)

- **Institutional accountability** sits with the Board, executive committee, or equivalent governing body, as it does for any consequential organisational decision.

The principle that matters is separation: the project team that builds the system should not also be the body that signs off on its deployment. SAFE AI extends existing structures to cover AI risk. Where no appropriate function exists, organisations should identify the closest one and extend it, rather than build new infrastructure.

Embedding trustworthiness in the SAFE AI Journey

Trustworthiness is foundational to responsible AI, encompassing reliability, transparency, accountability, fairness, and safety.

As AI systems increasingly shape consequential decisions, a growing ecosystem of governance frameworks has emerged to operationalise these principles.

- The EU's *AI Act* (2024) establishes risk-based obligations for trustworthy AI, while the OECD's *Principles on AI* (2019, updated 2024) emphasise human-centred values and transparency across member states.
- The *NIST AI Risk Management Framework* (2023) offers a structured approach for organisations to identify, assess, and mitigate AI risks.
- In the humanitarian sector, the ICRC's guidance on digital technologies and the UN Office for the Coordination of Humanitarian Affairs (OCHA) have underscored that AI deployed in crisis settings must meet heightened standards of accountability and do-no-harm.
- The Humanitarian OpenStreetMap Team and organisations like UNHCR have similarly stressed data dignity and community trust.
- UNESCO's *Recommendation on the Ethics of AI* (2021) provides a globally endorsed normative framework stressing inclusivity and environmental sustainability.
- IEEE's *Ethically Aligned Design* further bridges industry, civil society, and technical communities.

Together, these frameworks signal a global consensus: trustworthy AI is not optional but essential for equitable, safe, and human-centred technology.

SAFE AI and the NIST AI Risk Management Framework

The seven trustworthiness characteristics defined in the NIST AI Risk Management Framework – valid and reliable, safe, secure and resilient, fairness with harmful bias mitigation, privacy-enhanced, explainable and interpretable, and transparent and accountable – are embedded as active, recurring commitments that run through every stage and every step of SAFE AI.

This integration is deliberate and structural.

Each stage of the journey builds on the trustworthiness work of the stage before it, from the first questions about organisational readiness to the final stages of ongoing monitoring and evaluation and independent technical audit of the AI system:

- Stage 1 establishes the governance foundations and problem conditions under which trustworthy AI is possible.
- Stage 2 translates those foundations into responsible design choices, with community co-design functioning as a governance mechanism rather than a consultation exercise.
- Stage 3 verifies that design commitments are real in procurement contracts, in built systems, in test results, and in formal sign-off.
- Stage 4 ensures that trustworthiness is sustained operationally over time, not merely declared at launch.

The feedback loops built into the SAFE AI Framework and its decision gates that prevent progression until conditions are met mean that trustworthiness is not a destination reached at the end of the journey, it is the discipline applied at every step.

The SAFE AI Transparency Card provides an auditable record through which all seven characteristics can be verified by affected communities, colleagues, donors, and regulators alike. Continuous monitoring ensures and requires ongoing assurance and transparency card updates.

The SAFE AI Transparency Card

The central record of your SAFE AI journey

You should progressively fill out your SAFE AI Transparency Card with the outputs and insights you will generate from the other SAFE AI Tools and Guidance along your journey. It comes in two versions, aligned to your tier:

The Lean version is designed for Tier 1 use cases: a lighter, faster record focused on the essentials. The Full version covers the complete scope of decisions, risks, safeguards, and community engagement, and is required for Tier 2 and Tier 3 use cases

If your tier changes as your understanding of the use case develops, the Lean Card feeds directly into the Full Card – so nothing is lost. (See When to reassess your SAFE AI tier for situations in which revisiting existing SAFE AI outputs is vital.)

SAFE AI Transparency Card overview

Section 1: AI system snapshot, purpose, and intended use

Section 2: Humanitarian context, risks, participation, and safeguards

Section 3: Data sources, model design, testing, and evaluation

Section 4: Deployment, monitoring, and lifecycle management

Section 5: Community engagement, governance, accountability, and redress

Section 6: References, version history, and documented decisions

Completing the fields within the card is a shared responsibility. It is not a task for a single person, and it is not completed in a single sitting. Different sections will be owned by different members of your team – technical leads, programme staff, compliance, community engagement – and filled out at the relevant stage of your journey. The SAFE AI Transparency Card's value comes from that breadth: It reflects the full picture of your AI governance, not just one team's view of it.

The SAFE AI Impact Assessment and the SAFE AI Transparency Card do different work. The Impact Assessment is the structured analytical exercise conducted at each Decision Gate to test whether the conditions for proceeding are met. The Transparency Card is the single, accumulating governance record that holds the outputs of the Impact Assessment alongside the design decisions, risk assessments, community engagement records, and assurance findings produced across the journey. The Impact Assessment is the work; the Transparency Card is the record.

The practical benefit of a well-maintained SAFE AI Transparency Card is significant. If a donor, partner, or affected community asks how your AI system works, who is accountable, or how risks are being managed, it gives you a clear, documented answer. Done well, it makes those conversations easier – and its information helps you protect your organisation if questions are raised later.

Full guidance on completing both versions of the SAFE AI Transparency Card can be found in the SAFE AI Tools and Guidance section of the SAFE AI website.

Stage 1:
Problem definition and concept

What this stage is for

- Defining the humanitarian problem, assess whether AI is an appropriate response, and confirm organisational readiness to proceed

What it will help you produce

- A clear use case
- Documented rationale for using AI
- An initial risk and impact view
- A provisional SAFE AI tier
- A record of organisational and data readiness issues to address
- A go / no-go decision on whether you should proceed with your proposed AI use case
- A senior responsible owner (SRO) and accountable leads are identified

Who needs to be involved

- Programme, technical, compliance and community engagement leads, with clear senior leadership accountability

How communities are kept in the loop

- Engage affected communities early to shape the problem and assess whether AI is appropriate (see Co-design guidance)

SAFE AI Transpency Card sections filled in

- Section 1 (context, purpose, ownership)
- Section 2 (initial risks and impact)
- Section 5 (oversight roles)

Step 1.1: Information gathering and problem formulation

Know what problem you're trying to fix

What is this?

This step is designed to help you define your humanitarian problem and to decide whether an AI system or tool might help you address it.

Who should be involved?

Affected communities as co-definers of what the problem is. Local partners and field staff with operational knowledge of the problem and the context. Senior leaders responsible for the decision to proceed.

As with any successful humanitarian programme, SAFE AI solutions should be shaped to fit identified problems, not the other way around, so problem identification is the most consequential stage of the SAFE AI Journey. Decisions made here shape everything that follows. A clearly defined problem creates the conditions for the rest of the framework to do its work. A poorly defined problem will be processed by the framework into an assured answer to the wrong question.

Where you are considering AI to address a problem, start with the needs assessment work that has already identified it, and state the problem in plain language, before any reference to AI. If the problem cannot be stated clearly without invoking AI as part of the definition or cannot be located in existing needs assessment work, the problem is not yet ready for this framework.

The humanitarian sector already has collective processes for identifying need: the Humanitarian Needs Overview, Multi-Sector Needs Assessments, cluster-specific assessments, and organisational context, conflict, gender, and protection analyses. Problem identification for SAFE AI should connect to these processes.

Keeping communities in the loop

Good problem identification in humanitarian response means keeping communities involved at every stage of your SAFE AI journey. They are the people best placed to identify the issues they face, and their views are most useful in this first step, where they can shape whether and how a system is built, rather than at testing, where they can only validate decisions already made. Community engagement at problem identification should build on the accountability to affected people (AAP) and community engagement work organisations already do, including in needs assessment design, post-distribution monitoring, and feedback and complaints mechanisms.

For community-facing AI systems, engagement at problem identification is essential. For back-office or productivity systems, communities may not need to be part of problem identification, but the question of whether the system could affect them indirectly should be asked here, not deferred. Where these processes exist, they are the right starting point. Where they do not, the gap should be addressed before a SAFE AI Journey proceeds, not after.

The risks of doing community engagement badly at this stage are real and well documented. Participation washing, or consultation methods that give the illusion of meaningful engagement without genuinely transferring power or acting on community input (Madianou, 2019), is the dominant failure mode in current practice. Findings from SAFE AI co-design fieldwork in Kakuma Refugee Camp and Kalobeyei Settlement confirm

that the risk is not driven by communities lacking AI awareness; it is driven by underlying power asymmetries between aid recipients and providers, and between technical expertise and lived experience. Engagement that does not address these power dynamics reinforces them. Community engagement at problem identification is meaningful only where it is resourced, structured to surface dissent, and capable of producing the legitimate outcome of **not** proceeding with AI. Detailed guidance is set out on the SAFE AI website.

Is AI really the solution?

Once the problem is clearly defined, ask whether AI is the right tool for it. Many humanitarian problems do not require AI. Some are made worse by it. A non-AI solution that is faster, cheaper, more accountable, or more aligned with humanitarian principles is often the right answer, and concluding so is itself a SAFE AI outcome. Where AI is the right tool, document why in your SAFE AI Transparency Card.

Different people, with different backgrounds, may shift your understanding of the problem in ways no single perspective can. The members of your SAFE AI team, alongside the communities you serve, should discuss and document why you believe AI is the right tool or that it is not. Capturing that reasoning is the foundation of every later decision the framework will ask you to make.

Be alert to adoption pressure. Most poorly identified humanitarian AI problems share a common origin: the technology came first. Donor signals, leadership ambition, supplier pitches, or peer-organisation announcements created pressure to "do something with AI", and a problem was retrofitted to fit. Adoption pressure is not a problem statement. SAFE AI teams should name this dynamic where it is present and treat it as a warning signal, not a starting point.

Where AI may be appropriate, the range of options includes locally-built, frugal, and open-weight systems alongside foundation models from commercial providers. The choice between them is an architecture decision (see the SAFE AI Architecture and Tech Stack Decision Guide on the SAFE AI website) and should not be assumed at problem identification stage.

Responsible refusal: When humanitarian organisations should NOT use AI

In some situations, using AI could introduce unacceptable levels of risk. While risk identification and mitigation is covered in detail in another part of the SAFE AI Framework (see Step 2.1: Risk identification and mitigation), choosing not to go ahead with AI is a valid and responsible outcome. If any of the following conditions apply and cannot be mitigated, organisations should pause, redesign, or stop.

- **No meaningful path to redress:** There is no clear, accessible way for affected people to question, challenge, or seek remedy if the AI system produces harm, error, or exclusion.

- **Accountability cannot be clearly assigned:** It is unclear who is responsible for decisions influenced by the AI system, who responds when something goes wrong, or who has authority to pause or stop the system.

- **Data reuse exceeds reasonable expectations:** The AI system relies on historical or secondary data in ways that stretch consent beyond what people would expect today, without additional participatory safeguards.
- **Disproportionate impact on people who cannot opt out:** The AI system would significantly affect people with limited ability to refuse, exit, or meaningfully consent (e.g., due to dependency on aid, legal status, language, or connectivity).

- **Automation replaces essential human judgement:** The AI system would substitute for human decision-making in ways that materially reduce contextual understanding, discretion, or relational engagement.

- **Pressure-driven adoption:** The primary rationale for using AI is urgency, cost pressure, or donor expectation, rather than a clear, evidenced improvement in outcomes for affected populations.

Additionally, see the humanitarian principles check in the SAFE AI Transparency Card.

Step 1.2: Readiness assessment

Onboarding AI in humanitarian organisations

What is this?

A review of key things humanitarian organisations will need to do to set themselves up for success with SAFE AI principles in practice.

It highlights the conditions needed to use AI ethically, responsibly, and inclusively in your organisation, while adhering to humanitarian principles.

Who should be involved?

Country Directors, and Heads of Mission, leaders of AI initiatives introducing AI into back-office functions and/or external operations.

This includes senior leaders, IT heads, programme managers, or any staff initiating AI projects.

Getting started with responsible AI in humanitarian settings can feel daunting. This step introduces the SAFE AI Onboarding Readiness Checklist to help set up your organisation and your use cases for success.

You don't need to have all elements here fully in place to progress on your SAFE AI journey, but you should have a clear plan to address any gaps before moving forward.

By working through each step in this Problem definition and concept stage of the SAFE AI Framework, you should also be able to identify if your intended use case is likely to fall into a higher SAFE AI tier, which will affect which other parts of the framework you need to use.

Getting your organisation AI ready

Practitioner perspective:

"Meaningful AI deployment is largely beyond the reach of many humanitarian organisations. It is mostly limited to larger, well-resourced actors with existing infrastructure and specialist capacity."

As the AI landscape is moving so fast, and the need for assistance in crisis situations can feel overwhelming, this step focuses on establishing minimum governance, clarity, and safeguards to reduce risk and confusion.

It is designed to find ways to help you realise the benefits of AI for your organisation, rather than waiting for perfect readiness or time and resources you don't have.

If you don't have one, create an AI Policy

Clear guidance helps organisations move from informal or inconsistent practice to shared expectations.

It can reduce ambiguity for staff, manage risks arising from existing AI use, and provide clear pathways for escalation where AI use may introduce higher humanitarian, legal, or reputational risks.

One size does not fit all. Your goal should be to develop an AI policy tailored to your organisation's mission and operational contexts, which defines your objectives and ethical guidelines.

This is why you should treat your AI policy as a living document. New AI developments and shifting humanitarian contexts will require your policy to adapt – build in a process for regular review.

The following questions reflect some of the key issues humanitarian organisations need to consider before progressing with their SAFE AI journey.

- Do staff have clear guidance on how and how not to use AI in their day-to-day work?
- Are there obvious gaps between formal guidance and actual practice?
- Is there a clear, trusted route for escalating AI use that may introduce humanitarian risk?
- Do staff understand when AI use moves beyond low-risk, internal activity into a higher SAFE AI risk tier?
- Does the organisation have clear, practical mechanisms in place to reduce the risk of staff or other users inputting sensitive or confidential data into AI systems, including commercial AI tools?
- Are there clear sign-off and escalation procedures, with identifiable accountable leadership, for any proposed procurement, development, or expanded use of AI systems?

If you can't answer yes to these questions, your organisation may not be SAFE AI ready.

The more detailed questions in the full SAFE AI Onboarding Readiness Checklist will help you spot blind spots and map your path forward.

Policy is a component of organisational AI readiness alongside budgeting, training, culture change management, and decision authority.

Encourage openness and culture change

Recognise that AI adoption is a culture and mindset shift, not only a technology decision. Staff across most humanitarian.

Organisations are already using commercial AI tools. Organisations that make that use visible, supported, and open to challenge will govern it more effectively than organisations whose policies drive it out of sight.

Invest in change management and the internal conditions for staff to raise concerns about AI use without penalty.

Assessing the possible costs of AI implementation

Practitioner perspective:

"Despite the rhetoric around efficiency, costs are usually very high at the beginning. Any benefits tend to come later – and only under certain conditions."

Assign sufficient budget

Assigning a suitable budget will help steer you towards a successful project outcome. This goes beyond direct financial expenditure. AI use requires investment in safeguards, participation and community engagement, monitoring and potentially auditing of models.

AI investments often involve higher initial outlay for governance, training, and safeguards, while potential benefits may accrue gradually over time and vary by context. Some AI capabilities benefit from economies of scale, while others involve ongoing costs for compute resources, storage, or access to external services that increase with usage. These costs may fluctuate and can be difficult to predict or control, particularly where organisations rely on third-party platforms or cloud-based services.

Organisations should therefore be cautious about assuming immediate efficiency gains from AI, and consider whether anticipated longer-term returns are realistic given operational volatility, staff turnover, and changing humanitarian needs.

Such dimensions and trade-offs are useful to surface at the outset.

Common added cost components to consider include:

- Staff time for organisational and data readiness assessments, design, implementation, monitoring, and evaluation.
- Meaningful and ongoing community engagement will require additional time and resources to be allocated, sufficient budget for translation, payment to community participants for their time and expertise, interpreting technical issues (see the section on Co-design on the SAFE AI website), and bringing the insights into the design and overall planning of the AI implementation.
- AI and data literacy training for staff and leadership, particularly if additional in-house technical expertise is deemed necessary.
- Procurement of any AI system, if not developed in-house (see the SAFE AI Technical Assurance on the SAFE AI website).
- Ongoing expense for cloud-based AI services reflecting the amount of use of the AI system after deployment.
- Ongoing monitoring and evaluation of AI behaviour to ensure the system is working as intended, including regular testing and technical assurance (see the SAFE AI Technical Assurance on the SAFE AI website).
- Budget for regular technical re-evaluation, at a frequency determined by system volume and risk tier. This is not a one-off cost and it is a recurring operational requirement for any AI system that remains in deployment.

Practitioner perspective:

"Any meaningful AI engagement requires significant organisational infrastructure, security arrangements, external services, dedicated staff, and sustained finance. Humanitarian organisations are littered with failed IT projects where this was underestimated."

Set up governance and assign decision making authority

Successful and responsible use of AI in humanitarian contexts requires a multi-disciplinary approach. AI projects rarely succeed when treated solely as technical initiatives, as they depend on the interaction among technology, organisational processes, and humanitarian practice.

For this reason, we recommend creating a multi-disciplinary AI project team to work together on the AI implementation, drawing expertise from across the organisation.

Senior leaders are responsible for ensuring that relevant stakeholders are identified and involved, and that decision-making roles are clear. Leadership does not require technical expertise, but it does require sufficient understanding to ask the right questions, weigh risks, and make informed trade-offs.

Roll out AI literacy training

In addition to basic technical fluency for those developing or managing AI systems, organisations should ensure a broader understanding between social and technical teams (known as socio-technical). This includes understanding how technical design choices and AI systems interact with affected communities and individuals, humanitarian principles, protection risks, power dynamics, and local contexts.

Broader AI literacy training should be rolled out across the organisation to demystify AI, build familiarity with the technology you want your staff to engage with, and understand AI's potential for positive impact as well as its limitations and inherent risks. AI-related capacity does not need to be built all at once.

Organisations may draw on external expertise for training or technical support, while gradually strengthening internal capability. Before progressing with their SAFE AI journey, organisations should consider whether they have access to an appropriate mix of:

- IT, data management, AI, and cyber security expertise.
- Humanitarian programme and operational knowledge.
- Protection, gender, and safeguarding expertise.
- Community engagement and accountability to affected populations skills.

What Boards need to know about humanitarian AI use

The SAFE AI Framework expects Boards to apply existing duties for risk, accountability, and organisational mission to AI-specific governance questions. Board oversight is strategic rather than technical so trustees do not need to understand how AI systems work. They do need to understand the governance architecture under which those systems are deployed and the accountability pathways when they fail. As part of organisational readiness, Boards should:

- Be aware of how AI is currently being used across the organisation, including informally, and the basis on which it is being used.
- Apply a strategic test to AI use. Does this advance the organisation's mandate, or is it adoption pressure? Boards should be willing to say no.
- Designate a senior accountable owner for AI-related decisions and risk management, with a clear escalation path to the Board.
- Ensure that humanitarian principles, protection risks, and accountability to affected populations are governance foundations.
- Boards should expect the SAFE AI journey to be completed and SAFE AI Transparency Card to be published and regularly reviewed.
- Expect that affected populations are informed when AI shapes decisions about them, and that routes to challenge and correction exist.

- Recognise the deployer/developer asymmetry. Where the organisation deploys systems built upstream, Boards should expect documented accountability for upstream model choice, data provenance, and redress pathways when the system causes harm.
- Ensure insurance, liability, and indemnity arrangements reflect AI-specific risks, including the risk liability shifts to the deployer for upstream failures.
- Receive periodic updates on significant AI initiatives, incidents, or emerging risks, including any near-misses or decisions to withdraw a system.

Expect independent, external assurance to become a future requirement where AI materially affects decisions about people. Internal sign-off is the minimum. External assurance is the standard SAFE AI is building toward. Board engagement should focus on oversight and accountability, rather than approving individual AI use cases or technical designs. For the underlying strategic analysis of why deployer-level governance carries this weight in humanitarian AI, see the SAFE AI Governance Gap paper (McElhinney et al, 2026).

Once minimum governance, policy, and capacity conditions are in place, organisations can move to planning specific AI implementations as the next step on their SAFE AI journey – use case assessment.

Step 1.3: Use case assessment

What is this?

Assessment of potential use cases and types of AI systems you might encounter or consider in humanitarian work, along with some associated risks.

What does this require for your tier?

Awareness for Tier 1, with deeper and more sustained engagement for Tier 2. It should be considered mandatory for Tier 3.

Who should be involved?

Leaders of AI initiatives introducing AI into back-office functions and/or external operations. This includes senior leaders, IT heads, programme managers, or any staff thinking to initiate AI projects in humanitarian organisations.

By this point in the SAFE AI Journey, you have defined a humanitarian problem and concluded that AI is the right tool to address it. Use Case Readiness asks the next question: which specific AI use case will best serve the problem you have identified, and are you operationally ready to proceed? Where the answer to either question is not yet clear, return to Problem Formulation before going further.

The rapid change in the capabilities of AI tools and systems means that the problems they can address and the use cases where they may prove valuable are changing constantly. This does not mean that organisations should try to implement AI solutions simply because they are available. Capabilities, opportunities, and risks vary across AI and machine learning (ML) systems. You can find some examples on the SAFE AI website.

The use case readiness section of the SAFE AI Onboarding Readiness Checklist will help you identify some of the key things you should consider before you progress to designing the solution to the use case your SAFE AI problem identification step has helped you to define.

As you conduct activities such as co-designing with communities and community engagement to keep your communities in the loop, considering the appropriate AI architecture, and carrying out your SAFE AI Impact Assessment, the project concept is likely to evolve.

Outline the goal of your proposed AI project. Revisit this in light of community engagement work and the SAFE AI Impact Assessment in the next stage of your SAFE AI journey, where new evidence may shift what the project should do.

Key questions to answer about your potential AI use case:

- What specific humanitarian problem or need does the proposed AI implementation address?
- Why is the current approach insufficient, if one exists?
- Which AI capabilities are relevant to addressing this problem?
- What evidence supports the expected improvement, and from what context does that evidence come?

Data readiness

The humanitarian sector already operates within a developed data infrastructure: organisational data protection policies, ICRC's *Handbook on Data Protection in Humanitarian Action*, IASC *Operational Guidance on Data Responsibility*, OCHA *Data Responsibility Guidelines*, and your organisation's data protection focal points and humanitarian protection staff.

AI does not replace this infrastructure. It tests whether it is adequate for the demands AI places on it.

Three questions are enough at this stage:

Question 1: Is your existing data work fit for purpose for this use case?

Confirm that staff handling the data are familiar with applicable data protection policies, that PII and other sensitive data are adequately safeguarded, and that the data needed for the proposed AI use case can be processed lawfully under your existing consent and data protection arrangements.

Where gaps exist, address them in your data infrastructure before progressing the AI use case. The SAFE AI framework cannot substitute for an absent or inadequate data protection foundation.

Question 2: Does AI extend the use beyond what people understood at the point of collection?

In most cases, affected individuals give consent for their data to be used for the delivery of humanitarian assistance. They do not assume the data will be shared outside the humanitarian organisation, used to train AI models, or fed into predictive systems that infer new information about them or about people like them.

Repurposing data collected for service delivery to train, fine-tune, or operate AI systems is a new use, not an extension of existing consent. It requires fresh disclosure to the people whose data is being used, and where the AI may significantly affect individuals' rights, a Data Protection Impact Assessment (which is integrated into the SAFE AI Impact Assessment in the next stage).

This is one of the most under-examined risks in current humanitarian AI practice. Data collected from people in crisis under one set of expectations is being used in ways those people did not anticipate and have not been asked to agree to. Mechanisms should exist for people to question, contest, or withdraw from data uses that extend beyond what they understood at the point of collection, without penalty. Where these mechanisms do not exist, they should be built before the AI use case proceeds, not after.

Question 3: Have you considered the dual-use risk?

The data and model infrastructure used for registration, targeting, and needs assessment may also be accessible for commercial, intelligence, or surveillance purposes. AI intensifies this risk. It increases both the volume of data generated through humanitarian operations and the sophistication of what can be inferred from it, including from metadata that would not traditionally have been considered sensitive.

Local and nationally led organisations, which hold some of the most granular community-level data and often have the least capacity for data protection, are particularly exposed. Where the use case may produce or expose data that could be repurposed against the people the operation is meant to serve, that risk should be named at this stage and carried into the Impact Assessment.

Detailed consent guidance for community-facing AI systems can be found in the Co-design section of the AI website.

Supplier data ownership, access, and contractual terms are addressed in the SAFE AI Procurement Guide on the SAFE AI website.

Step 1.4: SAFE AI Impact Assessment

The considerations you should now have worked through in this first stage of your SAFE AI journey should have helped you think about your organisation's readiness, the topline opportunities for you to address problems with AI, and some of the potential risks of your proposed AI use case. The Stage 1 Decision Gate is the first formal moment in the SAFE AI Journey where the conditions for proceeding are tested.

The SAFE AI Impact Assessment, conducted at this gate, draws on the work of Steps 1.1, 1.2, and 1.3 to assess whether the humanitarian problem is real, the use case is viable, the data foundation is adequate, and AI is the right tool. Each is reviewed against humanitarian principles, protection requirements, and the responsible-refusal conditions that govern whether AI should be deployed at all.

You are now nearly ready to move on to Stage 2: Design, and the next step in your SAFE AI journey, Risk, identification and mitigation.

This next stage will provide further guidance and support for your multidisciplinary SAFE AI Project Team as they move towards procurement. It will dig deeper into how to reduce your organisation's exposure to AI-related risks via the SAFE AI Impact Assessment, and help you scope out your procurement and technical needs via the SAFE AI Architecture and Tech Stack Decision Guide.

Decision Gate: Determining whether your AI use case should proceed

The final part of the SAFE AI Impact Assessment is the Go/no-go decision matrix within the SAFE AI Onboarding Readiness Checklist.

This includes high-level questions to establish whether you have the necessary foundations in place to pursue an AI initiative responsibly.

If you've already worked through the earlier steps in Stage 1, this should provide a final verification that the organisational, governance, and risk conditions required to proceed are in place.

If you decide **not** to go ahead with attempting to deploy AI to address a particular problem, document the decision so you have a record of:

- Why the decision was taken
- What alternative approach will be used instead
- What conditions would need to change for AI to be reconsidered in future

This documentation will promote learning and protects staff from pressure to adopt AI prematurely, or to address problems deemed to have too high a risk.

Having completed the SAFE AI Impact Assessment and Go/no-go decision process, your organisation should now have a clearer view of where and how AI may be appropriate, what level of risk is involved, and what safeguards are required before proceeding onto the next stage.

How Stage 1 embeds trustworthiness into your SAFE AI journey

Stage 1 establishes the organisational, ethical, and governance foundations required for trustworthy humanitarian AI before any system is designed, procured, or deployed

Through problem formulation, readiness assessment, use case assessment, and the SAFE AI Impact Assessment decision gate, organisations are required to demonstrate that AI is appropriate to the humanitarian problem, that risks and safeguards have been considered, that governance and accountability structures are in place, and that affected communities and data practices have been properly considered.

Together, these steps support all seven NIST trustworthiness characteristics by establishing conditions for valid and reliable use, safety, security and resilience, transparency and accountability, explainability, privacy protection, and fairness with harmful bias managed.

Critically, Stage 1 also establishes clear conditions under which organisations should not proceed with AI deployment.

Stage 2: Design

What this stage is for

Once you have identified your problem and intended approach in collaboration with crisis-affected communities, Stage 2 turns to system design. The work here is to translate the agreed problem into an AI system whose architecture, tech stack, and oversight mechanisms are fit for purpose, proportionate to the SAFE AI tier, and accountable to the people the system will affect.

What it will help you produce

- A structured assessment of risks, impact and potential harms
- Clear mitigation strategies and safeguards
- Documented community input and co-design considerations
- An initial system design and architecture approach
- Defined governance, accountability, and oversight mechanisms

Who needs to be involved

- Programme, technical, compliance and community engagement leads, alongside community engagement specialists and, where appropriate, external technical or ethical advisers

How communities are kept in the loop

- Engage affected communities throughout the design process to understand potential impacts, test assumptions, and shape safeguards and system design decisions (see Codesign guidance)

SAFE AI Transpency Card sections filled in

- Section 2 (detailed risks and mitigation)
- Section 3 (system design and safeguards)
- Section 5 (oversight and redress)

Step 2.1: Risk identification and mitigation

What is this?

A structured approach to identifying, assessing, and managing the risks associated with your proposed AI use case. It helps you understand where harm could arise, evaluate the likelihood and impact of those risks, and define practical safeguards to reduce them to an acceptable level. It also supports ongoing review, recognising that many AI risks change over time and require reassessment.

What does this require for your tier?

Applicable for tiers 1, 2 and 3, with increasing depth of assessment, documentation, and mitigation required for higher-risk use cases.

Who should be involved?

Programme, technical, data protection, safeguarding, and compliance leads, alongside community engagement specialists. Where risks are complex or high, organisations may also need input from external technical, legal, or ethical advisors.

AI is imperfect and will always be imperfect. When exploring AI for humanitarian contexts, little effort should be spent considering whether AI will make mistakes – because it will. Focus on when and how the mistakes might occur and what you can do to counter them.

Risk awareness will already have been developing as you worked through Stage 1 of your SAFE AI journey.

This dedicated step helps you understand how different kinds of risk affect your SAFE AI tier (and so which SAFE AI tools and guidance you should be leveraging) – as well as what you can do to mitigate them. The SAFE AI Impact Assessment provides a structured way to document these risks and mitigation strategies.

Throughout your SAFE AI journey, risk assessment, mitigation identification, and community in the loop co-design often need to happen iteratively.

Once you have identified a specific risk, you can develop a corresponding mitigation strategy, and update your SAFE AI initiative's design. After the design update, it may be necessary to conduct a fresh risk assessment again.

Note that any AI use case may contain multiple risks. Each risk may require its own mitigation strategy. New risks can emerge at any time.

When you first identify a potential new risk, confirm whether you need to make changes to your project, such as your SAFE AI risk tier, or whether it is sufficient simply to document the risk you have identified along with why you have decided not to address it.

How risk affects your SAFE AI journey

Risks related to leveraging AI in humanitarian contexts can materialise across several dimensions, and at multiple places within your SAFE AI journey. Depending on their severity and the context of your AI use case, any of these risks could impact how you should approach your AI project – informed by your SAFE AI tier.

Remember that not all AI errors are equal. When AI-based capabilities are used for low-risk scenarios, like grant proposal writing, inaccurate AI model output has relatively little negative consequence. However, when AI is integrated into a high-risk process,

such as early disaster warning or landmine clearance, an error in the AI output can have devastating consequences.

This is why the SAFE AI tier identified for your use case has a direct influence on the proportionality required for your mitigation strategy. This can be found in the tools and guidance section of the SAFE AI website.

To understand why and how your AI use case's risk level may change, it's helpful to understand the main kinds of AI risk:

Static vs. dynamic risk

Some risk levels are predictable and unchanging. For those static risks, once you have identified the Tier for your use case, identified the associated risks, and implemented the appropriate mitigation strategy, it is likely to remain mitigated. (For example, you may have identified that there is a risk of exposing sensitive data in your AI system and have taken the steps to remove this information from what the AI system can access.)

Other risks are more unpredictable and can change rapidly. These dynamic risks may be more sensitive to external factors and pressures, and will require reassessment when certain triggers are met. See more below about triggers for reassessment. (For example, you may have identified that your flood risk forecasting AI system is sensitive to model drift and have taken steps to address this by updating the model, but given the nature of the use case, you will need to continually evaluate to what degree potential model drift requires model updates.)

Temporal risk

Consider the temporal aspect of the risk you have identified. Within which timeframe does the risk have potential to materialise? Immediate, short-term, long-term or open-ended?

Applied to an AI-enabled forecasting system, for example, an immediate risk could lead to misallocation of supplies; a short-term risk could be repeated inaccuracies causing operational inefficiencies; a long-term risk could be that decision-makers become over reliant on the model, reducing analytical capacity; and open-ended could be forecasting models shaping donor expectations and program design for years.

Stakeholder risk

Which stakeholder groups are most likely to be impacted by the risk you have identified is a key factor in assessing your risk exposure. These include:

- Your organisation – e.g. your staff directly, or your organisation's reputation.
- The impacted community – whether directly or indirectly.
- Your partners – including other sector partners or technology suppliers.
- Your donors.
- The sector at large.

Estimating the cost of unmitigated risk

Establishing the cost of error – i.e. the human, financial, legal, or other significant cost if an AI output is wrong – allows you to design your use case to be more resilient. Consider the notion of cost of error as a continuum, from low to high. Now consider where your specific use case lies on this continuum (see the SAFE AI tier guidance in the SAFE AI User Guide for high level help assessing this).

- Where the cost of error is low an d the use case is internal or back-office, AI outputs may be acted on with proportionate human review rather than full expert oversight. Even at low cost of error, outputs should be subject to spot-check, sample review, or other lightweight verification appropriate to the use case.

- Where the cost of error is anything other than low, AI outputs should be treated as recommendations for a human expert to evaluate, not as decisions in themselves.

For most AI outputs in a humanitarian setting, the cost of error should be weighed first and foremost against humanitarian principles and do no harm.

Navigating AI regulation: What you need to know

AI regulation is evolving rapidly. States, international bodies, and industry actors are developing laws, frameworks, and standards to regulate AI development and use.

For humanitarian practitioners designing AI solutions, managing AI-enabled projects, or shaping organisational policy, understanding this shifting landscape is essential for both responsible practice and legal compliance. Humanitarian organisations operate across multiple jurisdictions, often in contexts where AI-specific legislation is absent. Humanitarians should adopt robust ethical frameworks even where legal requirements are minimal, and monitor regulatory developments in operational contexts.

New legislation continues to emerge, and voluntary tech sector commitments have proven inconsistent. Humanitarian organisations and donors should:

- Monitor regulatory developments in crisis contexts they are operational.
- Adopt robust ethical standards even where legal requirements are minimal.
- Build organisational capacity for ongoing AI governance.
- For an overview of current AI laws, frameworks, and standards, see the SAFE AI website.

Mitigation strategies for your SAFE AI journey

Understanding the nature of the risks of leveraging AI in humanitarian action is the first step towards avoiding them. Applying the SAFE AI Framework is itself a path towards mitigation of the identified risks.

Choosing the right mitigation strategy is dependent on the operating context, specific use case and specific AI model chosen. As you go through your own SAFE AI journey, tools like the SAFE AI Impact Assessment will help you identify contextualised risks and plan for mitigating them.

The high-level AI risk mitigation strategies we've included in the SAFE AI Tools and Guidance section on the SAFE AI website are not exhaustive, but provide an introductory commentary on potential mitigation strategies.

Step 2.2: Community engagement

How to co-design AI with crisis-affected communities

What is this?

A structured approach to involving crisis-affected communities in the loop during the design and use of AI systems. It helps you understand how people may be impacted, surface risks and unintended consequences, and shape decisions around data, design, and safeguards through meaningful participation, rather than treating communities as passive recipients.

What does this require for your tier?

Awareness for Tier 1, with deeper and more sustained engagement for Tier 2; it should be considered mandatory for Tier 3.

Who should be involved?

Community engagement and accountability to affected populations (AAP) specialists, programme leads, safeguarding and protection experts, and members of the multidisciplinary AI project team. Where possible, representatives of affected communities should be directly involved, with appropriate facilitation and support.

Keeping communities in the loop is a core principle of the SAFE AI Framework, and should be considered at every stage of your SAFE AI journey – but it is especially important in the Design stage. Not all AI use cases require the same level of community participation, but when systems affect people's access to information, assistance, or protection, meaningful participation is a necessity.

You should apply the guidance in this step in full when an AI-enabled system:

- Directly affects crisis-affected people (e.g., information provision, eligibility, targeting, verification, feedback analysis).
- Influences access to assistance, protection, or services.
- Operates at scale or across multiple partners or locations.
- Cannot offer meaningful redress if it fails or causes harm.

In these cases, co-design should be considered a SAFE AI governance requirement – essential for mitigating harm, maintaining trust – not optional. This includes shared decision influence over system scope, deployment conditions, and continuation where risks to affected people are identified. Document key participation decisions and outcomes for audit and learning.

For stage-by-stage guidance on how to implement community engagement in practice including resourcing requirements, community decision authority, and the conditions under which deployment should not proceed, see the SAFE AI website.

Community participation in SAFE AI is a governance mechanism, not a consultation requirement – and it should make your AI implementations more effective. Crisis-affected communities hold knowledge about how AI systems actually function in their context – misinterpretation in local languages, exclusion patterns, unintended effects on access – that cannot be generated through internal assessment alone. Participation is therefore a source of governance intelligence, not only an ethical commitment.

This is why we believe that failure to apply this guidance in SAFE AI Tier 3 contexts constitutes a governance and risk management failure that could put you and your organisation, in addition to the communities you work with, at risk.

While we recommend keeping communities in the loop anyway as part of fundamental humanitarian principles, full co-design is not expected for low-risk, internal AI uses (e.g., drafting support, translation, internal analysis). Yet you should still consider whether assumptions about affected people are being embedded indirectly, or if outputs could later be reused in higher-risk contexts that would shift tier assessment.

Regardless of your use case's provisional SAFE AI tier, pause and reassess if:

- Community feedback indicates confusion, mistrust, or perceived harm.
- The AI system is repurposed beyond its original scope.
- Data collected for one purpose is reused in another.
- Accountability or redress mechanisms are not functioning in practice.
- Assumptions embedded in the system design are challenged or rejected by affected communities.

In such cases, escalation to fuller participation or reconsideration of deployment may be necessary. Repeated community concerns should be treated as operational governance signals, even where technical metrics remain within expected ranges. For guidance on establishing red lines and documenting community veto rights, see the SAFE AI website.

AI makes keeping 'community in the loop' more important than ever

AI systems built without community input fail in ways that internal testing does not catch. Misinterpretation in local languages, exclusion patterns that only become visible at the household level, assumptions about documentation or registration status that do not match reality on the ground: these are governance failures. They show up as outputs that quietly drop people from lists, and nobody checks the source because the system looked like it was working. Field practitioners recognise this risk. "We're getting into the space of data being the 'word of god' and with AI we're layering another level – risk of imagined realities emerging that bear little resemblance to reality," observed a participant in the SAFE AI regional consultation in Nairobi in June 2025.

The communities most affected recognise it too. "We need to be involved in decisions because we refugees are the ones affected" said a participant in the Kakuma community consultation in December 2024 (Suge et al., 2026). Traditional accountability mechanisms assume harm is visible and attributable. AI can exclude people silently, at scale, across multiple agencies simultaneously and faster than any existing feedback loop can catch.

That is the operational case for community involvement. But it is not the whole case, and organisations that stop there are missing the point. Involving communities to build a better system is not the same as giving communities authority over a system that affects their lives. People whose access to assistance is shaped by an AI process have a right to know that, a right to understand how it works, and a right to challenge what it decides about them. That is a governance requirement, and it sits at the centre of what SAFE AI is trying to do.

SAFE AI research grounded in community consultation in Kakuma Refugee Camp found that where communities have been consulted, their views have sometimes been simply ignored, suggesting that aspirations to participate in AI governance sit against a backdrop

of prior experience that gives communities little reason to trust that engagement will be meaningful.

The honest reality is that genuine co-design where communities holding real influence over whether a system is built, how it operates, and when it stops, is really rare in the sector. Consultation does happen but co-design mostly does not. The same barriers that have historically undermined collective accountability to affected populations such as tokenism, no consequences for non-compliance, or participation reduced to consultation without influence apply directly to community engagement in AI governance (Suge et al. (2026)).

In a period of tightening budgets, creating parallel AI-specific mechanisms is neither realistic nor responsible. Alignment, reuse, and collective approaches are practical necessities. Organisations should build AI accountability on existing collective AAP and coordination structures rather than creating parallel processes where possible.

But individual agency approaches, or even collective approaches when achieved, however well designed, cannot add up to sector accountability. Communities affected by AI systems deployed across multiple agencies need a consistent standard and a shared accountability architecture.

SAFE AI starts from that collective accountability requirement - can communities see what an AI system does, understand how it affects them, and actually contest the outcome? That is the minimum. Co-design is what good looks like. Independent assurance is what then allows donors to verify claims about community engagement.

User-centred design vs. co-design

Both co-design and user-centred design share a common goal: building AI solutions that address user needs and meet their expectations. However, these two approaches differ in important ways – specifically in terms of who participates, how decisions regarding the design and deployment of AI solutions are made, and the level of collaboration among all key stakeholders throughout the design process.

The key difference? User-centred design is created for people affected by crises, whereas co-design is created *with* them.

User-centred design: Users are consulted, but designers make final decisions, with no commitment to act on the feedback received. In co-design, designers and those impacted by the solution are partners, with defined decision influence, throughout the process, from problem definition to governance and evaluation, and their feedback is acted on.

Co-design: Ensures that AI solutions reflect real-world needs, especially in high-stakes or complex environments that impact the rights, liberties, and agency of individuals in a community. User-centred design can maximise usability and accessibility of tech products, but may overlook key opportunities or risks when users and local knowledge do not shape design decisions or buy in.

USER CENTRED-DESIGN	VS	CO-DESIGN
Passive providers of input and feedback	USER ROLE	Active co-creators with equal power and agency
Lead by AI solution designers	Y/N DECISION MAKING	Shared between designers and end users/those impacted by AI solutions
AI solution designers control the process and have final say on the outcome	POWER DYNAMICS	Equal participation

Aligning co-design with humanitarian standards and principles

The SAFE AI Framework's approach is grounded in and aligned with key humanitarian standards and commitments that promote ethical, inclusive, and accountable humanitarian action.

The list below maps how SAFE AI's co-design principles support and reinforce sector-specific standards and commitments.

1. *IASC Commitments on Accountability to Affected People (AAP)*: These emphasise leadership, participation, transparency, and feedback mechanisms that uphold the rights and agency of people in crises.

2. *Core Humanitarian Standard (CHS)*: Particularly Commitments 1, 4 and 5, which affirm that humanitarian responses should be appropriate, participatory, and accountable.

3. *The Grand Bargain Participation Revolution*: A sector-wide effort to shift power by involving people receiving aid in decisions that affect them.

4. *IASC Guidance on Participation (2021):* Encourages meaningful participation throughout humanitarian decision-making.

5. *The Sphere Handbook*: Recommends minimum standards for participation, inclusion and information sharing at all stages of response.

Together, these standards reinforce the need for a paradigm shift in how humanitarian technologies are designed: from designing for communities to designing with them.

The idea of shifting power and incorporating the voice of communities in humanitarian response is not new. For many years, this has been framed as Communication, Community Engagement and Accountability (CCEA).

Humanitarian reform processes and their resulting policy improvements have already led to commitments to collective, response-wide community engagement in emergencies, which in turn inform the humanitarian architecture and leadership.

Yet many staff and coordination structures remain ill-equipped to listen to, engage, and act upon feedback from communities.

There is no reason to suppose that the same individuals, with different technology, will be able to overcome this challenge, unless sufficient consideration is given to what engaging communities via AI, or working with communities to design AI, might look like.

This is why SAFE AI community engagement extends existing AAP, complaints, and CCEA infrastructure rather than building parallel mechanisms wherever possible.

The framework's role is to apply an AI-specific lens to the community engagement work organisations already do, not to add another consultation process.

What is participatory AI?

Participatory AI is not a tick-box exercise, but a foundational principle, grounded in AI ethics that seeks to redefine the notion of expertise, redistribute power, and ensure that those designing and/or deploying AI systems and solutions are held to account.

But including communities in AI decision-making requires care and skill. Yet, there remains little consensus on what constitutes a participatory AI approach and how to evaluate participatory AI efforts.

SAFE AI sets minimum expectations for participation proportionate to risk.

Important questions remain:

- What is the goal of participation?
- Who is included – and who is left out?
- What decisions do communities get to shape?
- What knowledge do participants need to meaningfully contribute?
- How will participants be rewarded for their inputs and knowledge sharing?
- How can communities monitor and evaluate AI solutions over time?

As the humanitarian sector expands its use of AI, it must prioritise the quality of participation over speed and reframe innovation from techno-solutionist interventions to human-centred design.

Guiding principles for co-designing AI solutions

Be inclusive

Engage diverse community representatives and ensure that marginalised voices are valued, heard, and integrated throughout the development process, prioritising equity and continually aiming to redistribute power and agency towards these communities.

This means ensuring that marginalised populations within a community are not only included but that their voices are heard and amplified throughout the process.

This requires actively addressing barriers to participation, fostering equitable collaboration, and designing solutions that are accessible, culturally sensitive, and responsive to diverse lived experiences.

By embedding inclusivity from the outset, humanitarian actors can create AI solutions that better serve and empower those most impacted by crises.

Ensure informed consent

Communities are provided with all the necessary information, including risks, to enable them to make an informed decision about the extent to which, if at all, they want their personal data to interact with an AI, or whether they want an AI involved in decisions that impact their well-being.

Informed consent in the development and deployment of AI solutions acknowledges that individuals have the right to be involved in decisions about their well-being, and aligns with emerging data privacy regulatory standards.

Be transparent

The objectives, methods and potential risks of an AI solution are shared with communities and made publicly available.

This includes not only general awareness and visibility about the AI solution across communities and the humanitarian sector, but also ensuring that communities fully understand both the inherent and potential risks associated with AI.

This also includes a commitment to sharing what works and what doesn't work regarding

community engagement and participatory AI in humanitarian contexts to help other humanitarian actors collectively improve their SAFE AI deployments over time.

Be ethical and accountable by design

Accountability is built into the design, not added afterwards.

This means humanitarian actors take responsibility for the impact of AI solutions on crisis-affected communities, including unintended harms – and that responsibility is reflected in transparent decision-making, fairness embedded from the inception of the design process, mechanisms for community feedback and oversight, and clear pathways for addressing grievances.

Designing for accountability protects human rights and dignity, safeguards privacy, and mitigates potential harms from the outset, rather than at the point of deployment when remedy becomes harder. (For supplier requirements, see the SAFE AI Procurement Guide on the SAFE AI website.)

Be reflexive

Adopting a reflexive stance in research or design involves critically examining one's own assumptions, biases, and position throughout a process.

When co-designing AI solutions for crisis-affected communities, adopting a reflexive stance requires humanitarians to actively question power dynamics and acknowledge how their own perspective shapes decision-making, while remaining open to adapting methods based on community insights and lived experiences.

This can help reduce power imbalances and the impact of unintentional bias.

Engage long term

Co-design is not a one-off activity but rather a commitment to deep, sustained engagement in which power and agency related to key decisions are shared with defined and transparent decision rights between humanitarian actors and crisis-affected communities.

This forms the backbone of the CCEA approach throughout humanitarian engagement. Community engagement is required throughout the design and deployment of the AI solution, from the pre-concept and concept stages through to evaluation – but the more you keep communities in the loop, the more the benefits of these relationships will compound.

Be complementary and collaborative

Avoid duplication, promote complementarity, and leverage existing work related to AI and humanitarian action, wherever possible.

This includes drawing on and integrating scholarship and other work from a range of AI-related disciplines and principles outside humanitarian action – such as AI for social justice, AI and environmental sustainability and feminist AI – as well as the abundance of resources related to community engagement in humanitarian contexts (see CDAC Network's website for more resources).

Draw on community engagement expertise, and where possible, dovetail into existing processes and structures, without displacing or undermining existing efforts, and collective and collaborative action can multiply the humanitarian impact of your SAFE AI initiatives.

For more detailed guidance on how to co-design with communities throughout your SAFE AI journey, see the SAFE AI Tools and Guidance section on the SAFE AI website.

How SAFE AI co-design complements other sector standards and commitments

Guiding principles for co-designing AI solutions	IASC AAP	CHS	Grand Bargain	SPHERE	CDAC CCEA
Be inclusive	✓	✓	✓ Participation	✓ Consideration of diversity & vun	✓ 2-way communication
Ensure informed consent	✓	✓	✓	✓ Info sharing	✓ Risk comms
Be ethical and accountable by design	✓	✓	✓ Implied	✓ Protection principles & complaint systems	✓ Feedback use
Be reflexive	✓	✓	✓ Programme adaptation & responsiveness	✓ Monitoring, adaptation & learning	✓ Context-aware comms
Engage long term	✓	✓	✓	✓	✓ Collective engagement
Be complementary and collaborative	✓	✓	✓ Coordination & common platforms	✓ Coordination	✓ Community leadership
Be transparent	✓	✓	✓	✓ Decision making & reporting	✓ Risk comms

Co-design takes time and resources

This guide offers some pointers on how to structure thinking around community participation in the use of AI. At its core, this requires time and resources from the humanitarian organisation. The Kakuma fieldwork conducted under SAFE AI in December 2024 reinforced this point.

The exercise was a structured pilot, not a deployment, and was made possible by years of prior work by FilmAid Kenya in the camp - relationships and trust accumulated long before any AI conversation began. AI co-design that lacks that foundation has little to build on.

While these are difficult numbers to tabulate in a donor proposal, a commitment to locally led AI systems requires long-term processes and engagement. Co-design systems built merely for the purpose of testing a specific AI solution will have limited value and utility.

Building institutional co-design processes that can adapt to the changing landscape of technology can be difficult, but it is essential for organisations that are committed to shifting power asymmetries in the deployment of emerging technology solutions.

Locally led AI systems are ethically preferable and operationally more durable. Systems built on existing community relationships, local knowledge, and trusted intermediaries are more likely to function as intended, adapt to context, and sustain community confidence over time. Investment in co-design is investment in localisation. The two agendas are inseparable.

Step 2.3: System design

Architecture and Tech Stack Decision Guide

What is this?

A structured approach to designing how your AI system will work in practice. It helps you translate your defined use case, risk assessment, and community input into clear decisions about data, models, infrastructure, and governance, ensuring the system is appropriate, secure, and aligned with humanitarian principles.

What does this require for your tier?

Applicable for tiers 1, 2 and 3, with increasing levels of technical rigour, documentation, and assurance required for higher-risk use cases.

Who should be involved?

Technical leads (including data scientists, engineers, and IT specialists), alongside programme, data protection, compliance, and safeguarding leads. Input from community engagement specialists and, where needed, external technical experts should inform key design decisions. For a guide to technical terms, you can visit the SAFE AI Glossary.

By this point in the SAFE AI Journey, you have worked through risk identification and community engagement. It's now time to assess what kind of system design will fit your purpose, constraints, and principles based on that.

This step helps your SAFE AI Project Team identify the data sources, infrastructure requirements, integration constraints, and AI model architecture most appropriate to the humanitarian problem you are trying to address. This will ensure that architectural and tech stack choices are fit for purpose, risk-aware, and well governed from infrastructure to human oversight.

Working through the tables for this step (which you can find in the SAFE AI Tools on the SAFE AI website) – and applying the necessary thinking these should provoke – will result in a more robust AI use case and a broadened awareness of the risks associated with certain decisions given the context. This will better prepare you for the Design stage of your SAFE AI journey.

Use this process to help you select the most appropriate architecture and model for your humanitarian use case selected in stage 1 of your SAFE AI journey. It will help you make transparent, evidence-based AI architecture choices, anticipate and mitigate context-specific AI risks, and align innovation with humanitarian values, oversight, and governance requirements. You will identify differing trade-offs, evaluate risk across the full AI stack, and understand where and how to set up and follow governance requirements. Thinking about each layer – and the different trade-offs you will need to consider along with them – will help you explore the components of your use case before you start the design stage.

Identify candidates for AI system architecture

The architecture question has four dimensions:

- First, the data sources available to the system, including how data is collected, accessed, governed, and whether its proposed use is consistent with the original

context and consent under which it was gathered. These choices will shape what kinds of AI approaches are feasible and appropriate.

- Second, the AI model type itself – rule-based, machine learning, generative, hybrid, or agentic – which will help determine the shortlist of viable AI architectures.

- Third, the pathway by which the system is built, deployed, and maintained, whether locally built, frugal or small AI, open-weight, or commercial foundation model. These are independent choices. A locally built system can be rule-based or generative. A foundation-model deployment can be commercial or open-weight.

- Fourth, the operating environment and infrastructure within which the AI system will run, whether on local devices, cloud infrastructure, or hybrid systems, each of which creates different operational, governance, security, and resilience considerations.

SAFE AI does not prefer any single pathway, architecture, or deployment model, but expects these choices to be made deliberately, justified in relation to the humanitarian problem being addressed, and documented clearly.

Identify applicable risks across the stack

Different types of risks impact and manifest at different layers of the AI stack - and can affect communities and humanitarian operations in different ways. Taking time to understand and map the risk implications affecting and created within separate and interdependent layers of the AI stack will help you decide whether an architecture is appropriate, and what risks you might need to mitigate and what oversight mechanisms you'll need to establish. In the SAFE AI Tools and Guidance on the SAFE AI website, you will find a list of tech-specific risks that you should assess exposure to before making your choice of which types of AI to deploy.

Some risks you identify during this process may change your SAFE AI tier and affect what guidance you follow and SAFE AI tools you should use.

Align with governance requirements

Cross-check that your choice of AI system will align and comply with legal, ethical, and organisational requirements, as well as humanitarian principles.

AI does not exist without data, and the type, source, and quality of that data can pose risks. Alignment begins with understanding your data and these risks; only then can you identify the governance and oversight mechanisms you'll need. The SAFE AI Tools and Guidance on the SAFE AI website provides a table to help you consider the data aspects to ensure your use case works as intended, exposes no risk to anyone, and complies with laws, internal policies, and humanitarian principles.

Governance and oversight mechanisms can be introduced at different layers of the AI stack to mitigate risk, monitor accuracy, comply with laws, and so on. As regulators continue to adapt to the rapid evolution of AI, governance requirements are changing fast. The SAFE AI website will continue to monitor these changes and report on the most significant.

Once you have worked through the process above, summarise the considerations and decisions, including the AI model and architecture you intend to implement. Document this information in Sections 1: System snapshot and 3: Data and model summary of your SAFE AI Transparency Card. This creates traceable, auditable design record of your AI use case.

Step 2.4: SAFE AI Impact Assessment

The second Decision Gate in your SAFE AI Journey. It sits at the end of Stage 2 and prevents the journey from moving into development until the design work of this stage has been formally assessed.

How this decision gate differs from the Stage 1 SAFE AI Impact Assessment

Stage 1 confirmed that the problem was real, the data foundation was adequate, and AI was the right tool.

Stage 2 examines what has been designed: the risks identified, the mitigations in place, the community engagement and co-design work conducted, the architectural pathway and tech stack chosen, and the oversight mechanisms set out.

Each should be reviewed against humanitarian principles, protection requirements, and the responsible-refusal conditions that govern whether AI should be deployed at all.

Apply tier-proportionate rigour

For SAFE AI Tier 2 and Tier 3 use cases, this Stage 2 SAFE AI Impact Assessment requires greater rigour, including external technical, legal, or protection expertise. The community engagement evidence must demonstrate that affected populations have meaningfully shaped the system's design.

Document the decision

Outputs should be documented in your SAFE AI Transparency Card across Sections 2, 3, and 6. A decision not to proceed is itself a SAFE AI outcome.

Once the gate is passed, the use case is ready to enter Stage 3: Development.

Embedding trustworthiness in Stage 2

Stage 2 translates the governance and accountability commitments established in Stage 1 into concrete design decisions through risk assessment, community engagement, system design, and the revised SAFE AI Impact Assessment decision gate. These steps require organisations to identify and mitigate potential harms, involve affected communities meaningfully in shaping system design, and ensure that technical, data, and governance decisions are evidence-based, proportionate, and properly documented before development begins.

Together, these steps support all seven NIST trustworthiness characteristics by embedding safety, fairness, privacy protection, transparency, explainability, resilience, and accountable governance directly into the design process. Critically, Stage 2 also establishes a second formal decision gate with the authority to prevent progression into development where risks remain insufficiently understood, mitigations are inadequate, or governance conditions have not been met.

Stage 3: Development

What this stage is for

- Once the design has passed the Stage 2 Decision Gate, Stage 3 is where the system is procured, developed, tested, and assured for deployment. The work here is to translate the agreed design into a working AI system, secure governance obligations contractually before deployment leverage is lost, and verify before any external rollout that what has been built matches what was approved.

What it will help you produce

- A configured or developed AI system aligned to your use case and design
- Documented testing and validation results
- A record of how identified risks have been mitigated in practice
- Technical assurance and sign-off for deployment readiness
- An updated SAFE AI Transparency Card reflecting development decisions and safeguards

Who needs to be involved

- Technical teams (including engineers, data scientists, and IT specialists), community engagement and procurement teams, alongside programme leads, data protection, compliance, and safeguarding leads. External suppliers or technical partners may be involved where systems are procured rather than built in-house

How communities are kept in the loop

- Engage communities, where relevant, to test assumptions, validate system behaviour, and identify unintended impacts, particularly for community-facing or higher-risk use cases

SAFE AI Transpency Card sections filled in

- Section 1 (supplier dependencies)
- Section 2 (updated risk summary)
- Section 3 (system design and safeguards – refined)
- Section 4 (development, testing, and validation)
- Section 5 (supplier accountability)
- Section 6 (sign off)

Step 3.1: Procurement

What is this?

A structured approach to selecting and engaging external suppliers or partners for your AI system. It helps you assess whether proposed solutions meet your technical, ethical, and governance requirements, and ensures that responsibilities, risks, and safeguards are clearly defined before entering into any agreement.

What does this require for your tier?

More rigorous due diligence, contractual safeguards, and oversight required for higher-risk tiers, more autonomous systems, and community-facing use cases.

Who should be involved?

Procurement and legal teams, alongside technical leads, data protection, compliance, and safeguarding specialists. Programme leads and community engagement specialists should also be involved where AI systems may affect crisis-affected populations. For higher-risk use cases, organisations may require external technical or legal expertise to support supplier evaluation and contracting.

Thinking about the issues outlined in this vital step on your SAFE AI journey prior to starting an AI procurement process will help you engage more confidently with AI suppliers, ensuring that technology implementations align with operational requirements, ethical standards, humanitarian principles, and the needs of the communities you serve. Where sufficient information about system behaviour, limitations, dependencies, or governance arrangements cannot reasonably be obtained, organisations should treat this as a governance risk within procurement and deployment decisions.

Remember

In the SAFE AI Tools and Guidance you can find a list of questions to ask your supplier when speaking with them, along with others for you and your SAFE AI team to consider during your AI procurement process.

As a potential customer, prior to signing a contract, you have the greatest leverage to secure detailed and specific information about the system or model in question and guarantees related to its safety, accuracy, training data, and other key issues outlined here and in the risk and mitigation step, the SAFE AI Architecture and Tech Stack Decision Guide, and elsewhere on the SAFE AI website.

Once you are locked into a contract, such discussions might be harder to secure and changes difficult to make.

Identify who should build what

Given your specific use case and the available in-house technical expertise, determine which stakeholder will be responsible for building which components of the targeted implementation.

Consider the long-term impact and dependencies of these decisions so that you have a clear understanding of the sustainability and maintenance of any implementation, including associated costs.

Procurement choices should reflect the architecture pathway selected at design stage. Procuring access to a commercial foundation model is a different procurement exercise from contracting a partner to build a locally-trained system, deploying an open-weight model on infrastructure the organisation controls, or commissioning a lightweight AI tool for offline field use.

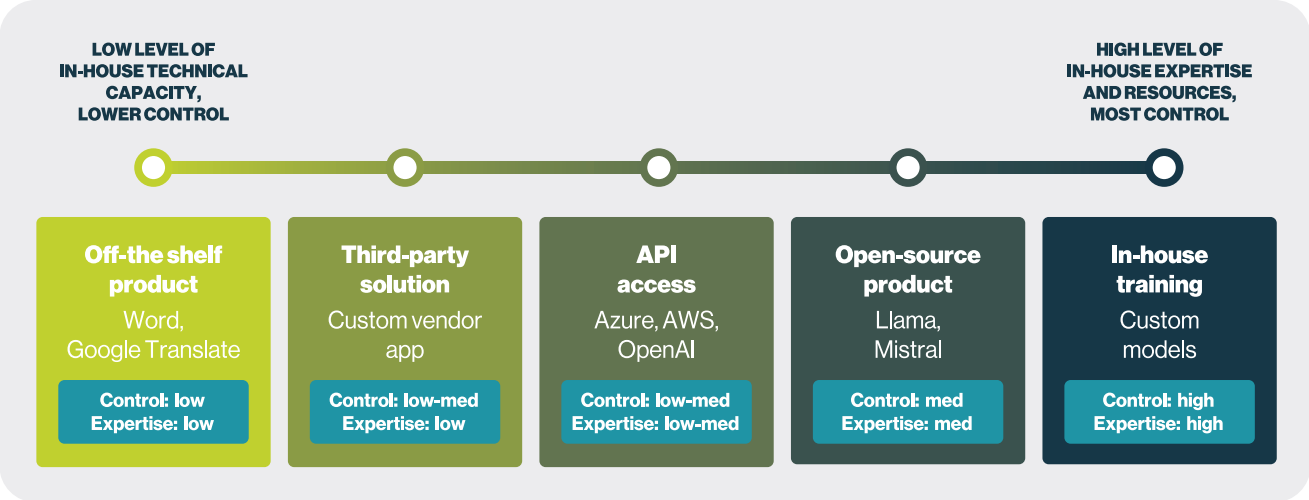
The control, sustainability, and exit terms appropriate for each pathway differ materially. Where the architecture pathway has not been deliberately chosen at design stage, procurement should pause and return to the SAFE AI Architecture and Tech Stack Decision Guide before going further.

Where the pathway selected is open-weight or open-source, additional procurement and operational considerations apply because there is no supplier relationship to mandate governance through. See the data governance guidelines on the SAFE AI website.

In the high-level graphic below, 'control' refers to the ability to build, train, fine-tune, further develop, or adapt the model or system architecture in other ways.

Review this to help you consider the long-term impact and dependencies of these decisions so that systems are sustainable and easily maintained.

AI model control selection vs level of expertise needed in-house



Confirm your implementation requirements

These are things that only procurement can secure via contracts, cannot be fixed later by policy, impact assessment, or assurance, and directly reduce systemic risk in a funding-constrained environment.If they are not locked in before contract signature, they are usually lost.

These include change control (model updates) and exit and portability (avoiding lock-in or single point of failure).

Procurement is also the moment to secure audit access for the lifetime of the contract. Humanitarian organisations and the sector as a whole will increasingly need the ability to commission independent audits of deployed AI systems for compliance verification, incident investigation, and the broader assurance work that Phase 2 of SAFE AI will establish.

Audit rights should be secured contractually at procurement, before deployment leverage is lost. At Tier 1, right-to-audit may be a short clause covering the principle of audit access. At Tier 2 and Tier 3, contractual depth should include third-party access, audit cycle frequency, and provisions for sharing findings with sector-level assurance bodies.

In your discussions with a potential supplier:

- Specify the functional requirements of the AI service/implementation, e.g., accuracy, integration options, language support, customisation options.
- Specify non-functional requirements, e.g., scalability, interoperability, data portability.
- Specify transparency and AI explainability requirements.
- Specify data protection and data sharing requirements.
- Specify the need for the supplier to report AI changes, including model updates and new capabilities (typically through an updated model card), including an explanation of how they may impact key metrics.
    - The supplier should be able and willing to explain how such changes may affect performance, accuracy, bias, or risk profiles relevant to the context.
    - Where AI systems are community-facing or high-risk, contracts should include the ability to pause or defer updates until impacts have been assessed.
- Specify that your organisation, or a qualified third party acting on its behalf, may audit the deployed AI system, including model documentation, training data provenance, system logs, and performance records, throughout the contract and for a defined period after termination.

Findings may be shared with sector-level assurance bodies where applicable.

At minimum, AI contracts for humanitarian deployments should include:

- Explicit usage disclosure requirements (what the system does and who it affects, in plain language)
- Clear data and model ownership terms (who owns outputs, training data derived from your deployment, and system improvements)
- Defined liability for AI errors causing harm to affected populations
- Opt-out consent conditions for affected people where technically feasible
- Supplier obligation to notify of material changes to the model, training data, or system, with updated documentation
- Right to audit the deployed system throughout the contract and after termination
- Exit rights including data export in usable formats and system documentation handover
- Penalties for material breach of governance commitments

Where a supplier refuses to include these terms, treat refusal as a procurement red line.

Remember

Only contracts can require advance notice from suppliers and only contracts can give pause and rollback rights in the event of concerns.

Ensure your contract bakes in provisions for changes and exits

In humanitarian systems, funding ends, suppliers exit markets, political contexts change, and platforms are defunded.

Procurement teams should assume that systems may need to be exited, paused, or replaced, and ensure this can be done without creating new risks for affected people. Ensure ability to exit effectively from the outset:

- Ensure your organisation retains the ability to modify, scale, replace, or discontinue the AI system over time, both from a technical perspective and in terms of intellectual property rights.
- Build contractual provisions that enable safe and orderly exit, including data export in usable formats, documentation handover, and clear timelines for data deletion or transfer at the end of the agreement.
- Include transparency and monitoring clauses that provide access to data required for oversight and risk management, with the ability to pause or terminate the contract if risks emerge, without losing access to essential data or system knowledge.
- Require supplier cooperation with internal reviews, donor inquiries, or regulatory assessments related to AI performance, safety, or harm.

SAFE AI for after the procurement contract has been signed

Procurement does not end at contract signature. Establish a process for ongoing monitoring of your AI supplier that includes changes in:

- Pricing structure or licensing terms.
- Corporate ownership, investor base, or parent company that could affect data governance or political exposure.
- Supply chain affiliations relevant to humanitarian principles, including links to state surveillance infrastructure or defence contracts.

Document the monitoring schedule and assign clear internal responsibility for it.

Remember to consider regulatory and governance requirements

Procurement decisions should clarify how responsibility and accountability are allocated between the organisation and the supplier where AI influences decisions, services, or access to assistance.

Where applicable, assess whether the AI system may fall within the scope of emerging AI regulation (e.g., the EU AI Act) and what obligations are placed on the organisation as a deployer.

Contracts should reflect these responsibilities, including roles in incident reporting, cooperation with audits or investigations, and escalation where risks or harms emerge. The challenge the SAFE AI Procurement Guide addresses is not only contractual but structural.

Organisations can be deemed accountable for decisions influenced by AI systems they cannot independently inspect – systems whose behaviour may change through updates, retraining, or interaction with new data in ways that are not visible to the deploying organisation.

SAFE AI Procurement and Business and Human Rights (BHR) frameworks

Assess whether the supplier has a published human rights due diligence policy, whether it has been subject to credible allegations of rights violations in AI-related work, and whether its governance structures include meaningful accountability for harm caused by its systems.

These are baseline conditions for responsible procurement in humanitarian contexts. Revisit BHR compliance as part of ongoing supplier monitoring, not only at the point of contract.

If you enter an agreement, share with your supplier only the data they need for the project and include contractual wording about how the data will be used and secured. Set data access to expire by default, with the option to renew access if an appropriate reason is provided.

Remember: Contractual terms establish rights but cannot substitute for governance architecture that requires verification at deployment and throughout the lifecycle. The technical assurance section of the SAFE AI Framework is the operational response to that constraint.

SAFE AI and responsible AI practice

Where AI outputs may affect people's access to assistance, services, or information, require the supplier to support your organisation's accountability and redress approach by ensuring:

- AI outputs can be explained in plain language to non-technical staff and, where appropriate, to affected people.
- Sufficient logging and audit trails exist to investigate errors, disputes, or harm.
- Human decision-makers can override, pause, or correct AI-influenced outcomes in practice.
- System design does not prevent affected people from raising concerns through existing feedback or complaint mechanisms, utilising existing sector structures.

In practice, suppliers rarely design redress mechanisms for humanitarian contexts, and generic appeals processes often stop at the technical user (staff), not affected people. Accountability breaks down when harm emerges because no one has authority or information to act. This creates a risk that AI-enabled decisions affect people without any meaningful way to question, correct, or challenge them.

Procurement is the moment when leverage is highest, so responsibilities to affected populations can be contractually clarified, and redress can be made operationally possible, not just theoretically desirable.

Where dominant technology suppliers are unwilling to provide this information, collective advocacy by different organisations can have greater influence than one trying to change the terms alone.

The SAFE AI Framework is intended to contribute towards and encourage this process of contractual improvement by raising the level of AI Procurement expectations across the humanitarian sector.

Step 3.2: System development

What is this?

A structured approach to build and configure the AI system in line with agreed design decisions, ensuring that what is implemented matches the intended use case, safeguards, and governance approach.

What does this require for your tier?

Higher-risk (Tier 2 and Tier 3) use cases will require more rigorous documentation, oversight, and validation during the build process.

Who should be involved?

Technical leads (internal or supplier), data and engineering teams, and the accountable programme lead. Procurement or supplier management teams should remain involved where external suppliers are responsible for delivery.

This stage translates the decisions made in design and procurement into a working system.

SAFE AI does not prescribe how systems should be built. Organisations should follow existing software development best practices, whether developing their systems in-house or through suppliers. What matters at this stage is not the specific development approach, but the traceability and integrity of what is built.

The system should reflect the design choices already agreed, including model selection, data sources, safeguards, and governance constraints. Any deviations from those decisions should be documented and justified.

A clear record should be maintained of the system at the point of build, including model version, configuration, data inputs, and dependencies. This provides a baseline for testing, technical assurance, and ongoing monitoring, and ensures that changes over time can be identified and assessed.

Step 3.3: Testing

What is this?

A structured approach to test and validate the AI system before deployment, ensuring that it performs as expected, that risks have been identified and mitigated, and that it is safe to proceed.

What does this require for your tier?

Higher-risk (Tier 2 and Tier 3) use cases require more rigorous testing, including closer scrutiny of performance across different groups, failure modes, and potential harms.

Who should be involved?

Technical teams (internal or supplier), data and evaluation specialists, and the accountable programme lead. Where relevant, community engagement, safeguarding, or protection specialists should be involved to assess real-world risks and impacts.

This step tests whether the SAFE AI system, as built, performs as intended and is safe to use in its intended context.

The SAFE AI Framework does not prescribe specific testing methodologies. Organisations should follow their existing practices for testing and validation, whether conducted internally or through suppliers. What matters is that testing is sufficiently rigorous to identify potential risks, limitations, and unintended outcomes before deployment.

Testing should consider not only technical performance, but also how the system behaves in real-world conditions. This includes how it responds to different inputs, how it performs across different groups, and how it fails. Particular attention should be paid to the potential for harmful outputs, bias, or misleading results.

Where the system is community-facing or assessed at Tier 2 or Tier 3, testing must include affected communities, not only technical experts.

Communities can identify failure modes that internal testing cannot generate: misinterpretation in local languages, exclusion patterns visible only at the household level, assumptions about documentation or registration status that do not match reality on the ground.

Community feedback during testing should be treated as governance evidence, not user research, and recorded in the SAFE AI Transparency Card. Where community testing surfaces concerns that cannot be mitigated, that is a finding for the Decision Gate, not a usability issue to be designed around. Detailed guidance is set out in the Co-design step within Stage 2.

Testing outcomes should be documented clearly and used to inform decisions about whether to proceed, redesign, or pause the implementation. This includes defining and assessing performance indicators such as accuracy, reliability, and robustness, as well as identifying any safeguards or human oversight required.

Where feasible, organisations should define indicative performance thresholds alongside validated baselines to clarify what constitutes acceptable system behaviour in operational context. These thresholds should support consistent deployment, reassessment, escalation, and suspension decisions over time.

In Stage 1, the SAFE AI Impact Assessment confirmed whether AI was the right tool for

the problem. The Stage 2 Impact Assessment confirmed whether the design was sound. Testing produces the evidence on whether the system as built performs as the design promised.

Where testing surfaces material differences from the design – such as performance gaps, bias patterns, or failure modes that change the risk profile – against the SAFE AI Impact Assessment, that is the step 3.4 Decision Gate. This is the mechanism through which the gate is held accountable to evidence, rather than to the design intent that came before testing.

Key findings and limitations from tests must be recorded in the SAFE AI Transparency Card, which carries forward as the system's governance record into deployment and monitoring. Requirements for formal verification of testing are set out in the technical assurance guidance.

Step 3.4: Pre-deployment technical assurance

Technical assurance Decision Gate

What is this?

Provides formal verification that the AI system has been tested, risks have been identified and addressed, and appropriate safeguards are in place before deployment. This step acts as a decision gate to confirm whether the system can proceed, requires redesign, or should not be used – this is the final official sign-off before deployment.

What does this require for your tier?

For higher-risk (Tier 2 and Tier 3) use cases, technical assurance should be more rigorous, may require independent or specialist review, and should include explicit sign-off before deployment.

Who should be involved?

The accountable senior decision-maker, technical leads, and risk, safeguarding, or protection specialists. For higher-risk use cases, independent reviewers, governance groups, or external experts may be required to provide assurance and support decision-making.

General AI governance frameworks do not provide sufficient basis for verifying how systems behave in specific humanitarian contexts.

Performance in controlled conditions does not reliably predict performance in fragile and conflict-affected environments, and there is currently no consistent method for verifying system behaviour in deployment.

This technical assurance process is designed to address that gap – providing context-specific verification rather than relying on general standards or supplier-provided documentation alone.

Technical assurance is a key step that helps you verify and validate your AI implementation and ensure its components perform as designed and expected. The technical assurance process builds trust in the AI implementation, which is needed by all stakeholders before they can collectively sign off on the AI system's external deployment.

Note: Pre-deployment technical assurance happens in Stage 3, at this step. Ongoing technical assurance happens in Stage 4 through monitoring, reassessment, and incident review. Together, they form the assurance pattern that runs through every SAFE AI deployment:

1. During Stage 3: Development (your current stage), when you have the initial end-to-end implementation in place so that you can test the AI system in an as close to the real-world environment as possible before deploying externally.

2. Again in Stage 4: Deployment, monitoring and evaluation, when you have completed the full external deployment of your AI system and your target audience has started using.

A compliance check will help ensure that your deployed AI system adheres to both your internal AI policies and any applicable external regulations or legal requirements.

Review your internal AI policy for compliance and check the SAFE AI website for pointers to the relevant regulations you need to comply with.

How to make contingency plans and handle discrepancies

Where testing surfaces conditions relevant to the Stage 1 responsible refusal criteria, this decision gate has authority to halt deployment regardless of prior stage approvals.

If the AI model or service fails to meet agreed-upon criteria or there is a mismatch between supplier claims and your findings, you should:

- Notify the supplier of specific discrepancies and request clarification or remediation.
- Consider additional rounds of testing if updates or fixes are provided by your supplier.
- Engage internal or external experts for deeper analysis if needed.
- If unresolved, reconsider adoption of the model or seek alternative solutions.
- If the mismatch poses a safety issue, halt deployment.

Additional technical assurance considerations for Agentic AI systems

Agentic AI systems – those that take sequential autonomous actions rather than generating a single output – require additional scrutiny at the technical assurance step.

Unlike discrete tools, agentic systems can initiate actions, call external services, and chain decisions without real-time human review.

In humanitarian contexts this includes automated logistics coordination, case management sequencing, and anticipatory action thresholds. When assessing an agentic system, additionally confirm:

1. There are defined boundaries on what actions the system can take autonomously versus what requires human authorisation.
2. A full action log exists and is reviewable after the fact.
3. A reliable interrupt or 'kill-switch' mechanism is in place and tested.
4. Accountability for each automated action step is clearly assigned before deployment begins.

This technical assurance process should include confirming that your key metrics meet the target values you identified in the procurement step.

This is also an opportunity for you to verify the claims about the AI model, service performance, and behaviour that your supplier has provided.

The process described here enables you to document key decisions and known risks, and to confirm AI system behaviour before you deploy it to your target stakeholders, especially when that involves vulnerable communities.

Technical assurance will require some dedicated resources, so make sure that you have budget, skills, and time secured to carry out this process. (See the section Assessing the possible costs of AI implementation in Step 1.2 for more on budget.)

Keep in mind

Technical assurance is only one part of the continuous oversight of the AI implementation needed during its deployment. AI models and systems may evolve over time, as well as the conditions of the crisis context in which you are operating. See Step 4.3: Ongoing monitoring, evaluation and technical assurance for more information.

Your conclusion may be that everything works as expected and no change is needed. Often, the process will point out improvements you can make.

By following this technical assurance process and using the rest of the SAFE AI framework to better inform your decisions, your organisation can more confidently adopt AI systems that align with your mission and values.

The Stage 3 Decision Gate: Signing off on your SAFE AI deployment

Once all criteria are met and the findings are satisfactory, you should:

- Document formal approval from relevant technical and programme leads.
- Name the individual with ongoing accountable authority for this AI system post-deployment – the person with authority to pause, modify, or decommission it and record this in the SAFE AI Transparency Card.
- Before sign-off is finalised, document the agreed performance baseline – the specific metrics and threshold values that constitute acceptable functioning in your context; without a documented baseline, subsequent re-evaluations have nothing to compare against and drift cannot be detected.
- Retain all assurance artifacts (SAFE AI Transparency Card, test results, correspondence) for audit and accountability purposes.

Communicate approval to the supplier and proceed to deployment as planned.

Embedding trustworthiness in Stage 3

Stage 3 translates governance and design commitments into operational reality through procurement, development, testing, and pre-deployment technical assurance.

These steps require organisations to demonstrate that AI systems are being built, configured, tested, and governed in ways that align with the humanitarian purpose, safeguards, and accountability conditions established in earlier stages, with any deviations documented and justified.

Together, these steps support all seven NIST trustworthiness characteristics by verifying that systems are safe, valid and reliable, secure and resilient, privacy-enhanced, explainable, transparent, and subject to meaningful human oversight before deployment. Critically, Stage 3 also establishes a final pre-deployment decision gate with the authority to halt deployment where systems do not meet the conditions required for responsible humanitarian use.

Stage 4: Deployment, monitoring and evaluation

What this stage is for

- Stage 4 is where the trustworthiness commitments made in Stages 1, 2, and 3 are sustained against operational reality. A system signed off for deployment may operate for months or years across changing contexts, evolving populations, and shifting threat environments. Stage 4 sets out what governance looks like across the operational life to create an audit trail so that risks identified before deployment continue to be managed, new risks are detected, and decisions to continue, modify, pause, or decommission the system are made on evidence rather than assumption.

What it will help you produce

- A live SAFE AI Transparency Card kept current through deployment.
- Documented evidence that human oversight, community feedback, and override mechanisms are functioning in practice.
- Re-evaluation triggers and a defined response when they fire.
- A documented decision at any point of material change, including the decision to decommission.

Who needs to be involved

- The oversight body, technical leads, programme leads, community engagement specialists, and the multi-disciplinary AI project team. Independent or specialist reviewers may be required for Tier 2 and Tier 3 systems.

How communities are kept in the loop

- Communities are notified at deployment in the channels they already use. Feedback mechanisms are operational from day one and the loop is closed: feedback received reaches system designers, and changes resulting from feedback are communicated back. See the Stage 4 elements of the Co-design AI section.

SAFE AI Transpency Card sections filled in

- Section 4 (deployment, monitoring outputs, version history)
- Section 5 (community feedback, redress, oversight in practice)
- Section 6 (version control, sign-off record updated through the lifecycle)
- The governance record that Stage 4 produces via your SAFE AI Transparency Card should be updated through the deployment lifecycle, monitoring outcomes documented at every reassessment, incident reports filed and resolved, ongoing technical assurance findings recorded.

Step 4.1: SAFE AI Transparency Card consolidation

Pre-deployment governance review

What is this?

The point at which the SAFE AI Transparency Card moves from working draft to published governance record, immediately before the system goes live.

What does this require for your tier?

Card consolidation is required at every tier and is the precondition for proceeding to deployment and ensuring the right to know is met.

Who should be involved?

The named accountable post-deployment authority, technical leads, programme leads, community engagement specialists.

Your SAFE AI Transparency Card is built across the SAFE AI Journey, accumulating evidence at every step: use case rationale at problem identification, risk assessments at design, community engagement records at co-design, design decisions at system design, procurement terms at procurement, test results at development, and assurance sign-offs at the Decision Gates.

Before deployment is the point at which this accumulated record should be fully consolidated, finalised, and published as the system's governance record before any deployment begins. This establishes a public, auditable foundation on which the rest of Stage 4 depends. Three things must be confirmed at this step:

1. **Your SAFE AI Transparency Card is complete**
   Every section reflects what was decided at the relevant stage of the journey. Where decisions were revised in later stages, the Card reflects the final version, not the historical sequence. Cross-references between sections are accurate.

2. **The Card is accessible**
   Communities affected by the system should be proactively notified before deployment, in accessible formats and through channels they already use. Notification should make clear what the system does, what it does not do, how decisions can be challenged, and who is accountable. The relevant sections of your SAFE AI Transparency Card should be available to communities at this point, not only on organisational websites. The accessibility test is whether a frontline staff member could share the Card with a community member during a routine interaction.

3. **The named accountable authority has signed off on the Card**
   Your completed SAFE AI Transparency Card is the system's governance record going forward. That authority's name, role, and contact pathway should be visible on the published Card. Sign-off is the act through which the authority accepts that, from this point forward, the Card is the record against which the system will be monitored, assessed, and held to account.

This is also the moment at which monitoring parameters are locked in: the seven trustworthiness domain thresholds, the trigger conditions, the reassessment frequency, and the incident response protocol. These should be documented in your SAFE AI Transparency Card at consolidation and become the baseline against which Step 4.3 (ongoing monitoring and evaluation) will operate.

Where this process surfaces a gap, e.g. a missing risk assessment, an undocumented design decision, or a community engagement record that is incomplete, that gap needs to be filled.

A consolidation that goes ahead with known gaps creates a governance record that cannot be relied on, and a deployment that cannot be held to account.

Once the consolidation is complete, the system is ready to proceed to Step 4.2: Deployment.

Step 4.2: Deployment

What is this?

The point at which the SAFE AI system goes live with users or affected communities, and ongoing governance begins.

What does this require for your tier?

Tier 2 and Tier 3 systems require communities to be notified before deployment and feedback mechanisms to be operational from day one.

Who should be involved?

The named accountable post-deployment authority, technical leads, programme leads, community engagement specialists.

Deployment is a governance event.

Before the system goes live, confirm that everything documented at the Stage 3 pre-deployment technical assurance decision gate is in place: the named accountable authority, the agreed performance baseline, the feedback and redress mechanisms, the incident response protocol, and the published SAFE AI Transparency Card.

For the conditions under which Tier 3 deployment should not proceed without co-design in place, see the tier reassessment guidance and Stage 4 of the SAFE AI Co-design tool on the SAFE AI website.

Communities affected by the system should be notified before deployment, in accessible formats and through channels they already use. Notification should make clear what the system does, what it does not do, how decisions can be challenged, and who is accountable. The relevant sections of your SAFE AI Transparency Card should be available to communities at this point, not only on organisational websites.

Feedback mechanisms should be operational on day one of deployment. A feedback channel that is built but not staffed, or staffed but not connected to the people who can act on what comes through it, is a governance gap that will surface as harm.

Step 4.3: Ongoing monitoring, evaluation and technical assurance

Perpetual Decision Gate

What is this?

Where ongoing monitoring and evaluation, including internal technical assurance work feed a documented decision to continue, modify, pause, or decommission the deployed AI system. This step should operate continuously across the system's operational life, with formal reassessment at defined intervals and immediate reassessment at any trigger.

What does this require for your tier?

Should be monitored by all tiers, with proportionate frequency and depth. Tier 2 and Tier 3 systems require more rigorous assessment, more frequent reassessment cycles, and may require independent or specialist review.

Who should be involved?

The named accountable post-deployment authority, technical leads, programme leads, community engagement specialists, data protection lead, and the multi-disciplinary AI project team. For Tier 2 and Tier 3, organisationally independent reviewers may be required.

This step sets out the governance requirements for ongoing monitoring and evaluation: the triggers that should activate reassessment, the process for verifying each of the seven core trustworthiness domains, architecture-specific considerations for different AI types, and the incident monitoring and reporting process that must be in place from the first day of deployment.

The depth and frequency of monitoring activity should reflect your SAFE AI risk tier. Tier 3 deployments require more intensive and regular review than Tier 1, as they are likely to encounter more risks that could require mitigation.

Ongoing monitoring and evaluation is the retrospective complement to the forward-looking SAFE AI Impact Assessment. Where your impact assessment will make assumptions about how the system will behave in deployment, Stage 4 produces evidence about how it actually does. Together they form the assurance loop that runs across the system's full lifecycle: assumptions tested against evidence, with your SAFE AI Transparency Card carrying the record forward.

For community engagement during deployment, monitoring, and evaluation including pilots, real-world testing, feedback mechanisms, third-party evaluations, and sharing learnings with communities see the Co-design section.

The SAFE AI Transparency Card should be updated through the operational life of the system at every re-evaluation trigger, supplier change, model update, and incident.

This ensures trustworthiness is sustained across the full lifecycle of a deployed AI system. Pre-deployment technical assurance at Step 3.4 verified that the system was ready to go live, now Step 4.3 verifies that it remains safe, fair, and accountable in the real-world conditions of humanitarian operation, conditions that change, that the system was never fully tested against, and that no supplier assessment can predict.

Monitoring and evaluation processes should be linked to defined escalation and response

pathways where significant risks, failures, or unintended impacts are identified. These pathways may include additional review, mitigation measures, deployment modification, or suspension of use where appropriate.

Keep in mind

AI systems can perform better in controlled test conditions than when operating at scale in live deployment, and suppliers have structural incentives to demonstrate good behaviour during assessment. Where possible, supplement pre-deployment technical assurance with spot-check evaluations at intervals after full deployment.

This is also why the SAFE AI Framework's continuous evaluation requirement applies throughout the deployment lifecycle, not only at sign-off.

Why ongoing monitoring is different in humanitarian contexts

AI systems deployed in stable commercial environments face a known set of operational conditions: user demographics are relatively fixed, data distributions shift gradually, and the cost of error, while significant, is rarely life-altering.

Humanitarian AI systems must operate differently. The populations they serve change as displacement and crisis evolve. The data environments they depend on are often sparse, incomplete, or subject to sudden disruption.

The trust relationships through which communities engage with humanitarian actors and through which feedback reaches system designers are fragile and depend on conduct that may have nothing to do with the AI system itself.

This means that monitoring and evaluation must be sensitive to three distinct categories of change:

1. Technical changes in how the system performs.
2. Socio-technical changes in how the system interacts with the people who use and are affected by it.
3. Humanitarian changes in the context in which the system operates.

A significant trigger in any of these categories can materially alter what the system does to affected communities, and each requires a different form of assessment and response. You can find examples of each kind of trigger on the SAFE AI website.

The governance task is therefore not simply to check that the system is still doing what it was doing at launch. It is to verify, continuously, that what the system is doing remains appropriate given who it is serving, how it is being used, and what is happening in the operational environment around it.

For some AI architectures – particularly generative and agentic systems – this task is qualitatively different from monitoring traditional software: the system's outputs may shift in ways that are not captured by any metric, and the harm potential of undetected drift is high.

Frequency and scheduling

The appropriate frequency of formal reassessment depends on SAFE AI tier, system volume, architecture type, and the pace of change in the operational context.

As a baseline:

- **High-volume or community-facing Tier 2 and Tier 3 systems:** formal reassessment across all seven domains every three months as a minimum, with continuous metric monitoring between assessments.
- **Lower-volume internal Tier 1 systems:** formal reassessment every six months as a minimum.
- **Generative and agentic systems at any tier:** qualitative output review and action log review monthly, regardless of tier. These architectures require more frequent human scrutiny because metric monitoring alone is insufficient.
- **Any system:** immediate reassessment following a trigger event, regardless of the scheduled cycle. Triggers override the schedule.

Technical leads should advise on frequency at the point of deployment sign-off and document the agreed schedule in the SAFE AI Transparency Card.

Any deviation from the schedule including a decision to defer reassessment must be documented and justified by the named accountable authority.

Keep in mind

Scheduled reassessment assures trustworthiness is maintained throughout the system's deployment.

In fast-changing humanitarian contexts, the appropriate response to a trigger is immediate assessment, not waiting for the next scheduled cycle.

A six-month reassessment schedule does not mean a system can operate for six months without governance attention. It means that, in the absence of triggers, formal reassessment occurs at that interval.

Incident monitoring and reporting

Ongoing monitoring identifies gradual drift and scheduled reassessment detects patterns over time. Neither is designed to handle the immediate governance requirements of an incident – an event in which an AI system has produced, or is suspected to have produced, harm, error, or a significant unexpected output affecting one or more individuals or communities.

Incident monitoring and reporting is a distinct process that must be in place from the first day of deployment, not constructed in response to an event after it has occurred.

An incident, for the purposes of the SAFE AI Framework, is any event in which:

- An AI system produces an output that causes, or is credibly suspected to have caused, harm to an individual or community – including harm to access to assistance, protection, or information.
- An AI system produces an output that is acted upon and subsequently found to be materially wrong in a way that had real consequences for affected people.
- A security or privacy breach occurs that involves data processed by or for the AI system.
- An autonomy boundary is breached by an agentic system, or an action is taken that was not authorised.

Feedback from affected communities or frontline staff indicates that the system is causing confusion, harm, or exclusion at a scale or severity that requires immediate response.

Any of the monitoring trigger conditions in the previous section materialise in a form that constitutes an immediate risk rather than a scheduled reassessment need.

More detail on how to respond to incidents can be found in the SAFE AI Tools and Guidance on the SAFE AI website.

Updating the SAFE AI Transparency Card following an incident

The SAFE AI Transparency Card must be updated following every incident, not only those resulting in system changes. Update the following:

- Section 4 (deployment and lifecycle): Record the incident date, severity, brief description, and any resulting changes to system configuration, version, or operational scope.
- Section 5 (community in the loop and governance): Update redress and feedback records to reflect community notification and any changes to grievance mechanisms.
- Section 6 (version control and sign-off): Add the incident reference number, updated sign-off date, and name of the accountable authority.

Where an incident results in significant system changes – re-training, architectural modification, or changes to the autonomy boundaries of an agentic system – the updated card should be treated as equivalent to a new deployment in terms of the governance review required.

A logged, investigated, and resolved incident is evidence of a functioning governance system. Organisations that have deployed AI in humanitarian contexts and have no incident record after an extended period should question whether their identification and reporting mechanisms are working.

Absence of incidents is not evidence of safe operation.

The Stage 4 Decision Gate: When monitoring findings require action

This is the Decision Gate that closes Stage 4 in Phase 1 – and is activated when monitoring and evaluation lead to findings that require action.

Monitoring, evaluation, and ongoing technical assurance findings – whether from scheduled reassessment, community feedback, or incident response – should lead to one of five documented outcomes:

1. **Continue without change**
   Monitoring confirms that all seven domains are within agreed thresholds. Document findings and confirm the next reassessment date.

2. **Implement operational adjustment**
   A specific concern has been identified that can be addressed without system modification, for example, retraining staff on override procedures, restoring a feedback mechanism, updating community communications, or clarifying accountability assignments. Document the adjustment and confirm it has been completed.

3. **Escalate to supplier**
   A technical issue requires supplier action, performance degradation, a model update that has introduced bias, a security change, or a gap in audit trail provision. Document the escalation, the supplier's response, and the agreed remediation timeline. Do not resume full deployment until remediation is independently verified.

4. **Pause deployment**
   Monitoring has identified a concern that cannot be resolved through operational adjustment and requires formal review. Pausing deployment is a governance decision, not an admission of failure: it is the correct response when the alternative is continued deployment of a system that may be causing harm. Document the reasons for the pause, the governance process being initiated, and the conditions under which deployment may resume. Notify affected communities.

5. **Decommission**
   Monitoring or evaluation has shown that the system cannot be modified to operate safely, fairly, or accountably in its current context. See section on decommissioning for the process that applies.

Cross-functional accountability is required to pass the Decision Gate:

- The Accountable Senior Owner holds decision-making authority for all outcome types.
- The AI Governance Team, functioning as the oversight body, the formal reassessment process and maintains the monitoring record.
- The Data Protection Lead should be involved where monitoring identifies privacy or data concerns.
- Technical and IT leads manage supplier escalation and verify remediation.
- Humanitarian Programme and Community Engagement staff are responsible for ensuring community feedback flows into the formal monitoring process and that affected communities are informed what has changed as a result, the implications and what they can do if they do to seek redress.

The outcome chosen must be recorded in the SAFE AI Transparency Card, with the rationale, the evidence on which the decision was based, and the name of the Senior Responsible Owner accountable who made it.

Perpetual Decision Gate: Monitoring and community in the loop

Monitoring is a governance function that affects the communities the system serves.

Communities should not be passive subjects of the SAFE AI governance process.

At minimum:

- Community feedback is a monitoring input, not only a monitoring output. Findings from community engagement channels must flow into the formal reassessment process and the incident identification process.
- Communities should be informed when a formal monitoring review has been completed, in accessible language and through channels they already use.

- Where monitoring identifies an issue that has affected communities, including bias, data errors, privacy breaches, or harmful outputs, those communities must be told what was found, what was done, and what protection is in place going forward.
- Where monitoring leads to a pause or decommissioning, communities that depend on the system for access to assistance or information must be notified in advance, with clarity about what alternative provision is available.
- Where appropriate, evaluation results should be made public and shared both with the humanitarian sector and with affected communities. Public sharing is the mechanism through which sector-wide learning becomes possible and the foundation for comparative oversight.

For detailed guidance on structuring community engagement through the monitoring phase, see the SAFE AI website.

The monitoring record as accountability evidence

The SAFE AI Transparency Card's monitoring and incident record is the evidence base through which donors, partner organisations, and affected communities can verify that responsible governance has been maintained throughout the operational life of the system. A card that is updated at every trigger event, every reassessment, and every incident, and that shows clearly what was found, what was decided, and who decided it, is the practical expression of the accountability that the SAFE AI Framework requires.

Phase 2 of SAFE AI will establish the sector-level independent assurance function that uses this record. Where Phase 1 produces the audit-ready evidence, Phase 2 does the auditing.

This step works alongside the SAFE AI Technical Assurance tool in the Tools and Guidance section, which provides templates and detailed checklists for the assurance work conducted at Step 3.4 (pre-deployment) and at this step (ongoing). The monitoring frequency, triggers, and assessment domains for ongoing technical assurance are set out in more detail in the Tools and Guidance section of the SAFE AI website.

Step 4.4: SAFE AI Phase 2

SAFE AI Phase 2: Moving towards independent auditing of AI systems

What is this?

Step 4.4 names the assurance function that Phase 2 of SAFE AI is being built to provide: independent external review of deployed humanitarian AI systems by a body organisationally independent of the deploying organisations entirely with costs shared to enable responsible options for all.

What does this require for your tier?

Sector-level independent assurance is needed across the spectrum of humanitarian AI deployments with depth and frequency calibrated to risk.

Who should be involved?

Once Phase 2 is operational, an independent assurance body acting on behalf of the sector.

The SAFE AI Framework establishes deployment-level governance and produces the evidence base through which assurance becomes possible (SAFE AI Transparency Cards, Impact Assessments, monitoring records, incident reports).

What Phase 1 does not yet provide is the function that turns that evidence into operational assurance: independent external review by a body organisationally independent of the deploying organisation, accountable to sector-level standards rather than to the organisation paying for the audit.

In SAFE AI Phase 1, deploying organisations monitor themselves, sign off their own deployments, and decide on their own continuation. This framework makes that work transparent but not yet independent.

Phase 2 of SAFE AI will establish an independent assurance function that unlocks shared, cost-efficient assurance for the sector. This is in development through consortium and donor consultation.

This framework is the foundation. Phase 2 of SAFE AI builds the independent assurance function on top of it: shared infrastructure, accessible to Global South actors at no cost, available to the wider sector at a fraction of what individual assurance costs today. The right investment now unlocks responsible AI adoption and the confidence to back it at scale.

What to do when you need to decommission a SAFE AI system

Decommissioning a humanitarian AI system – whether at the planned end of its lifecycle, in response to provider exit, or because monitoring or evaluation has shown the system should not continue – is itself a governance act.

The named accountable authority, technical leads, procurement and legal teams, programme leads, and community engagement specialists should all be involved, and communities informed through the same channels used at deployment.

The decision should be documented in the SAFE AI Transparency Card, including the reason for decommissioning, what alternative arrangement is in place for the function

the system was performing, what happens to data held by the system, and how affected communities are informed.

Where decommissioning is the result of a supplier exit rather than organisational choice, the protocol developed at procurement applies. See Exit strategy from suppliers mid-deployment in the SAFE AI Procurement Guide. The same principles apply: communities are informed, service continuity is maintained or safely wound down, data is retrieved and secured.

The SAFE AI Transparency Card should be updated at decommissioning to reflect the system's last deployed status, then archived for record-keeping and learning.

Embedding trustworthiness in Stage 4

Stage 4 is where the governance commitments built across the previous three stages are tested against operational reality, not once but continuously.

Through Card Consolidation at Step 4.1, deployment governance at Step 4.2, ongoing monitoring, evaluation, and technical assurance at the Step 4.3 Decision Gate, and the future-facing commitment to independent external assurance at Step 4.4, organisations are required to sustain trustworthiness in conditions that change, that the system was never fully tested against, and that no vendor assessment can predict.

Together, these steps support all seven NIST trustworthiness characteristics by ensuring that valid and reliable performance, safety, security and resilience, fairness with harmful bias managed, privacy protection, explainability, and transparent and accountable governance are sustained operationally rather than declared at launch.

The SAFE AI Transparency Card, updated at every trigger event and every reassessment, is the evidence that trustworthiness has been maintained, not merely intended. Phase 2 of SAFE AI will subject that evidence to independent external scrutiny; Phase 1 ensures it exists and is audit-ready.

Closing the governance gap: From framework to tools

The SAFE AI Framework makes the case for how AI has to be governed differently in humanitarian contexts.

The SAFE AI tools are what help organisations enact, document, and produce governance to do this. They are published digitally at cdacnetwork.org/safe-ai. These new governance elements fall into three clusters:

1) Accountability that reaches the people AI affects

- When AI shapes decisions about a person's access to assistance, that person has a right to know automation is in play, to understand how decisions are being made, and to contest those decisions.
- When multiple organisations deploy AI in the same context, communities, donors, and partners have a right to know whether the systems being used are held to the same standard.
- When AI is used to make decisions about people who never interact with the system at all, accountability has to reach them too.

> *"We've learned in the digitalisation of aid over recent years that the preferences of communities have often been overlooked... often they prefer face to face and trust building."*
>
> **– Local NGO Leader, SAFE AI Regional Dialogue, June 2025**
>
> *"AI needs to be installed by some cultural information so that it can follow the guidelines of the community."*
>
> **– Focus group participant, Kakuma Refugee Camp, December 2024**

2) Procurement and architecture decisions made deliberately

- The architecture choices organisations make shape whether the right to know is real; the choice between locally built, frugal, open-weight, and commercial foundation models has direct consequences for who can see, audit, and improve a system.
- Right-to-audit clauses, model change notification, data ownership terms, and exit rights are AI-specific contractual conditions that have to be secured at procurement.
- Data collected for service delivery is increasingly being processed by AI in ways the original consent did not anticipate.

3) Operational governance that recognises AI is dynamic

- AI risks change after deployment through model drift, behaviour shifts, retraining, and repurposing, in ways existing risk frameworks could not have foreseen

- AI systems operate in a wider information environment they do not control - both producing outputs into it (AI content mistaken for trusted humanitarian communication, AI-driven misinformation and disinformation) and taking in inputs from it (training data, real-time feeds, scraped third-party data) that can be compromised by adversarial actors. Compromised inputs produce systematically compromised outputs in ways monitoring cannot easily catch.

- Pre-deployment and ongoing technical assurance for AI requires assessment of trustworthiness, performance against baselines, and bias drift over time.

- Responsible refusal, the decision that AI is not the right tool for a particular humanitarian problem at any point in the journey, is a legitimate outcome the framework supports.

These governance commitments, applied through existing functions, are what the SAFE AI tools make operational.

How to use the SAFE AI tools

Most humanitarian organisations have substantial governance already: risk committees, procurement functions, data protection officers, safeguarding leads, AAP infrastructure, programme quality monitoring and evaluation. These structures are mature and continue to do important work.

> *"What is often missing in our work are tools that we can integrate into our workflows that will start to achieve some of the standards."*
>
> **– SAFE AI Expert Pool member, virtual meeting, May 2025**

SAFE AI does not replace the governance practices humanitarian organisations already have – but it does stress test whether they are adequate for what AI specifically introduces. The framework's task is to close those gaps.

The SAFE AI Framework expects organisations to apply its tools through risk committees, programme monitoring and evaluation teams, data protection functions, and the equivalents that already exist.

The tools can then extend what organisations already do.

- **For organisations with strong existing governance**, the tools are AI-specific extensions that connect to functions already in place.

- **For organisations with weaker existing governance**, the tools provide the depth and detail needed to set up responsible AI deployment from the ground up.

Where governance practice is most constrained, sector-shared infrastructure becomes most valuable.

The SAFE AI documentation standard then helps organisations produce evidence in a form that is consistent and comparable across organisations using the framework. That consistency is what gives sector-level assurance its force.

The SAFE AI tools support different parts of the framework's work:

- **Journey-aligned tools** support the work of each stage of the SAFE AI Journey, including:

- The SAFE AI Onboarding Readiness Checklist.
- The SAFE AI Use Case Assessment Checklist.
- The SAFE AI Risk Mitigation Strategies.
- The SAFE AI How to Co-design AI with Crisis-Affected Communities.
- The SAFE AI Architecture and Tech Stack Decision Guide.
- The SAFE AI Procurement Guide.
- The SAFE AI Technical Assurance.

- The **SAFE AI Impact Assessment** is the structured analytical exercise conducted at each pre-deployment Decision Gate before deployment to test whether the conditions for proceeding are met.
- The **SAFE AI Transparency Card** is the framework's central governance record. It accumulates across the system's full lifecycle, from initial design through deployment and operational monitoring, holding the outputs of the pre-deployment Impact Assessments alongside the design decisions, risk assessments, community engagement records, procurement conditions, deployment sign-offs, ongoing technical assurance findings, monitoring outcomes, and incident records. The Card is what makes responsible AI deployment visible, comparable, and improvable across organisations and over time.

Donor use of SAFE AI in the humanitarian funding cycle

The largest humanitarian donors fund AI-enabled humanitarian work directly through bilateral grants and indirectly through pooled and pass-through mechanisms. SAFE AI is designed to give donor staff a practical means of applying their organisations' responsible AI commitments to the programmes they fund.

This section is for the people inside donor organisations who design calls, review proposals, sign grant agreements, conduct due diligence, and manage programme oversight.

It does not require AI expertise. It assumes the reader is already familiar with humanitarian funding workflows and is looking for the points at which AI governance fits into them.

What donors get from SAFE AI

SAFE AI gives donor teams four things they can use directly:

- A risk tier that tells them how much governance evidence is proportionate to request from a partner organisation, scaled to the actual risk of the AI use case rather than to the size of the grant.
- A standard set of governance artefacts (the SAFE AI Transparency Card and SAFE AI Impact Assessment) that can be requested, reviewed, and compared across partners, reducing the burden of bespoke due diligence on every grant.
- A defensible basis for proceed, redesign, pause, or decline decisions where AI is proposed as part of a humanitarian response.
- A set of harm signals to look for in monitoring and reporting, before they become incidents.

Key principles for donor application of SAFE AI

- Expectations should scale with the risk tier of the AI use case, not the size of the grant.
- Focus on whether governance is functioning in practice, not whether documentation is complete on paper.
- Treat SAFE AI as a basis for dialogue with partners, not a checklist to be returned.

Where SAFE AI fits in the donor funding cycle

Donors engage with AI governance at five points in the funding cycle. At each point, SAFE AI provides a structured basis for what to require as the evidence base for responsible funding decisions.

1. **At proposal and technical review:** Require a plain-language use case description with SAFE AI risk tier classification, a statement on community involvement and consent, and confirmation that the organisation has the governance capacity – legal, technical, programme, and community engagement – to implement AI safely. A proposal that cannot answer these questions clearly is not ready for funding.

2. **At the design stage:** Require a completed SAFE AI Impact Assessment identifying foreseeable harms and mitigations for the specific use case and context. Assess whether mitigation strategies are substantive – addressing bias risks, data gaps, community exclusion, and conflict sensitivity – or whether they describe intent without concrete mechanisms.

3. **At the procurement and development stages:** Require sight of the SAFE AI Transparency Card confirming that governance obligations are contractually embedded, that audit rights exist, and that accountability for AI-influenced decisions is clearly assigned. Responsibility that is shared between human judgement and system output without clear assignment is a governance gap.

4. **At deployment:** Require documentation confirming that human oversight is functioning in practice that caseworkers or programme staff are applying overrides, community feedback is being received and acted upon, and monitoring is operational. Planned safeguards that are not functioning are not safeguards.

5. **During monitoring and evaluation:** Actively seek harm signals rather than waiting for them to appear in routine reporting. These include consistent override rates among caseworkers, community complaints that are not reaching system designers, bias audit findings that are documented but not acted upon, model drift in a changed operational context, and cybersecurity incidents with protection implications.
   Where harm signals emerge and the partner organisation cannot demonstrate an adequate governance response, donors have the authority and the responsibility to require that response as a condition of continued funding. Where harm cannot be mitigated, requiring decommissioning is a legitimate and sometimes necessary funding condition.

> *"Owing to cuts in the sector we're going to have to be much more specific and granular about the cost efficiency analysis. It's not going to be enough to say it's going to save you money…we're going to have to be specific."*
>
> **– SAFE AI Expert Pool member, virtual meeting, May 2025**

This framework gives donors the tools to apply their AI governance commitments to the programmes they fund now.

Conclusion

What the SAFE AI Framework delivers

Phase 1 of SAFE AI gives humanitarian organisations and their funders a defensible basis for AI deployment decisions, and the sector its first shared operational standard for how those decisions are made, recorded, and held to account.

It is built on empirical research, sector consultation, and direct engagement with crisis-affected communities.

It is designed to be used now, with the resources and capacity humanitarian organisations actually have.

Phase 1 operationalises one fundamental norm: that people whose lives are shaped by AI-influenced decisions have a right to know when automation is in play, a right to understand how it is shaping decisions about them, and a right to contest those decisions. This right operates at both individual and collective level.

Where multiple organisations deploy AI in the same context, as is routine in humanitarian operations, communities also have a right to know whether the system being used on them by one organisation is held to the same standard as the system being used by another.

How do we know AI governance is working in humanitarian action if we cannot see, compare, or improve the systems being deployed?

- Phase 1 of SAFE AI is what makes the answer possible.
- The framework makes humanitarian AI deployments visible.
- The documentation standard makes them comparable across organisations.
- The decision gates and ongoing monitoring create the conditions for AI to improve over time, in public, and at scale.

The framework translates these rights into governance artefacts that can be inspected, audited, and used as a basis for accountability:

- The SAFE AI Transparency Card, published before deployment and updated through the lifecycle, is the central record.
- The SAFE AI Impact Assessment is the structured forward-look.
- The tier logic ensures governance effort scales to actual risk.
- Right-to-audit clauses secured at procurement, proportionate to tier, ensure that the evidence base produced under Phase 1 remains accessible for independent assurance.

What the SAFE AI Framework cannot do

Phase 1 sets the standards organisations follow when they deploy AI. It does not yet provide independent verification that they are doing so. The sector still carries structural gaps that no framework, applied within a single organisation, can close:

- Internal governance processes cannot verify themselves.

- Sector-shared frameworks lose force when funding pressure and capacity constraints create incentives to defer.
- Pooled and pass-through funding mechanisms, the channels closest to affected communities, currently carry no AI governance conditions.
- Communities operating in lower-resource languages remain structurally less protected by the safeguards built into commercial AI systems.

> *"Ninety percent of the training data used for language AI comes from just 17 dominant languages."*
>
> **– SAFE AI FCDO roundtable, March 2025**

What's next for the SAFE AI Framework

Closing the governance gap is what unlocks responsible innovation.

Phase 2 of SAFE AI will support roll out and uptake of the framework, establish the independent assurance function, including the sector-level standards, and the comparative evidence base, that means SAFE AI can evolve into shared, cost-efficient assurance for the sector.

The governance infrastructure the sector needs goes beyond what any voluntary framework can generate alone. It requires investment and political commitment to establish the independent assurance function the sector currently lacks.

Without that infrastructure, right to know stays a principle on paper. It cannot be checked or made operational across the many organisations and systems that shape people's lives.

> *"Merely discontinuing a system is not enough to stop the harms from it now… we talk a lot about AI life cycles, but how many of us have actually heard about afterlife cycle stages?"*
>
> **– SAFE AI Expert Pool member, virtual meeting, May 2025**

We have two asks:

- **For organisations using this framework:** Apply it, document what you find, and share what you learn. Publish your SAFE AI Transparency Cards before deployment, not after. The evidence base that makes independent assurance possible depends on sector-wide transparency about where governance works and where it fails.
- **For donors and sector leadership:** The gap between this framework and the accountability architecture the sector needs is a funding and governance decision. It requires three things: backing uptake of SAFE AI; embedding SAFE AI conditionality in humanitarian funding, including pooled and pass-through mechanisms; and investing in the independent assurance function that turns this framework into accountability. The right architecture unlocks responsible AI innovation and the confidence to invest in it at scale.

Right to know is the test the framework is built to meet. Phase 1 makes that achievable from now for organisations adopting SAFE AI. Phase 2 will make it accessible to local, national and international actors on equal terms, and sustainable across the sector.

SAFE AI will continue to evolve

For the latest insights and information about developments that may impact your SAFE AI journey, as well as the latest version of the SAFE AI tools, SAFE AI Glossary, and supplementary resources, visit the SAFE AI website.

Appendix 1: SAFE AI tools and guidance

Why the SAFE AI tools are published digitally

The SAFE AI tools are published digitally at cdacnetwork.org/safe-ai. This is a deliberate choice.

Once an AI system is deployed, it does not stay the same. The model gets updated. AI capability is added to existing software through agentic features, plug-ins, and supplier-driven updates. The situation the system is being used in changes. Problems no one foresaw at launch start to appear.

Tools that govern AI deployment have to keep up with both the technology and the operational realities they are designed for. A printed tool fixed at the moment of publication starts to lose accuracy from the day it is printed. The exception is the SAFE AI Onboarding Readiness Checklist, included here, which addresses the strategic and organisational readiness questions that need to be in front of senior leadership and Boards before any AI deployment proceeds. Those questions change less than the tools that follow them. We also include a snapshot of the central governance artefact, the SAFE AI Transparency card, with full template available online.

Digital release also continues to ensure the framework is participatory in its development. As organisations apply the tools, they generate evidence about what works, what does not, and what is missing. That evidence informs revision. Each version of the tools is sharper than the last because adoption has shaped it. The SAFE AI Community of Practice on AI will continue to meet and be one of the places to support this process.

The SAFE AI Framework approach laid out in this document sets out the case and the architecture. The digital tools deliver against it, version-controlled, dated, and improved through use.

At its heart, the framework is an instrument of legibility. It makes AI deployment legible to the people it affects, comparable across the organisations deploying it, and improvable through use. The governance infrastructure humanitarian organisations already have is not replaced; it is stress-tested, extended, and made fit for what AI specifically introduces.

Together they constitute the SAFE AI Framework as a living standard, designed to mature alongside the technology it governs and the sector it serves.

The SAFE AI Onboarding Readiness Checklist

Use these questions to assess your readiness at the organisational level for AI use in the following areas.

Gaps identified here do not prevent AI use, but may indicate the need for clearer guidance, escalation, or additional safeguards, as set out in the following sections.

These considerations should help organisations anticipate and manage financial, operational, and reputational costs, rather than responding reactively once issues arise.

As noted in the main document, the SAFE AI Framework offers different value to organisations at different stages of AI governance maturity:

- Comparability for those with established AI governance.

- AI-specific extension for those with general governance.
- Full operational architecture for those building from the ground up.

For those organisations – particularly UN agencies and larger INGOs – who have already developed internal AI governance documentation, ethics review processes, and compliance frameworks, consider using these prompts to review existing practice.

Checklist 1: Taking stock

What AI is already in use, at what level, and whether your organisation has a clear position on it.

Confirm	Yes / no / unclear	Notes / actions needed
Strategy and organisational position Is there an AI policy outlining your organisational position on use of AI in humanitarian action? If so, does it define a clear position on the role of AI in supporting humanitarian aims and its relationship to accountability to affected people? Does your organisation have processes to assess the *reliability* of information used in needs assessment, targeting, or operational decisions, including where those sources may have been influenced by automated systems?		
Existing AI use, guidance and controls Do you have a working inventory of AI systems in use across your organisation, including AI features embedded in platforms and tools you already use for other purposes? Do existing rules or guidance govern staff use of AI tools and data, including commercial AI services? Have you taken steps to understand how staff are actually using AI in their day-to-day work, including through personal or commercial tools not sanctioned by the organisation?		

Checklist 2: Leadership and accountability

Who owns AI governance, who can make decisions, and where authority sits.

Confirm	Yes / no / unclear	Notes / actions needed
Leadership and accountability Is there a senior leader accountable for AI-related decision-making and risk? Where a small number of people carry multiple functions, is it still clear who has final accountability for AI decisions and who covers that responsibility when they are absent?		
Budget and capacity (high-level) Does your budget for AI uses cover more than tooling (e.g. talent, training, change management, data infrastructure, insurance, ongoing support including audits of models)?		
Existing governance maturity Can existing governance mechanisms (e.g. safeguarding, data protection, digital risk) be extended to cover AI-related risks?		

The goal of clear rules is not to prohibit AI use but to make it visible, supported, and open to challenge.

Rules that drive use underground are harder to govern than rules that bring it into the open.

Checklist 3: Rules and permissions

What staff are and are not allowed to do, and what requires escalation or approval.

Confirm	Yes / no / unclear	Notes / actions needed
Scope Do your rules apply to both organisation-wide AI initiatives and individual staff use of commercial AI tools? Do your rules apply across all functions (programme, operations, support, headquarters)?		
What is permitted and what is not Have you defined which categories of data must never be input into AI systems (e.g. PII, beneficiary data, sensitive or operational information)? Have you clarified which types of AI use are permitted for routine work, and which are not? Have you defined which types of AI use require prior approval or escalation? Have you distinguished between lower-risk uses and those requiring additional review or safeguards (e.g. using SAFE AI tiers)?		
Escalation and approval pathways Is there an escalation path when AI uses move beyond low risk or internal applications? Are staff explicitly required to seek advice or approval where AI use goes beyond routine or where risks are unclear? Is it clear who is accountable for decisions to proceed, redesign, pause, or stop an AI use? Is it clear how decisions and rationales should be documented (e.g. using the Impact Assessment and Transparency Card)?		

Checklist 4: Cost and resource implications

What AI actually costs, including governance, monitoring, and failure.

Confirm	Yes / no / unclear	Notes / actions needed
Cost and risk awareness Are the potential costs of data breaches, misuse, or regulatory non-compliance understood and taken into account when budgeting for AI use? Are there sign-off and escalation procedures for procurement, development, or expansion of AI systems? Have you considered the resource implications of escalation, review, or redesign if risks increase? Have insurance, liability, or indemnity implications been considered, particularly where AI outputs influence decisions affecting others?		

Checklist 5: Assembling the right team

Assigning roles and equipping your team to use AI responsibly.

Confirm	Yes / no / unclear	Notes / actions needed
Skills and capability Do you understand what AI-related skills and experience already exist within the organisation? Do staff involved in AI have sufficient socio-technical understanding to identify risks and trade-offs? Have you identified the most significant knowledge gaps and how to address them (e.g. training or external support)? Is there a visible internal advocate for the SAFE AI approach?		
Training requirements Have you introduced proportionate AI literacy training for staff based on their roles and risk exposure? Has AI adoption been treated as a sustained culture and change management challenge, with dedicated resource, rather than addressed through a one-off training module? Does training cover organisational rules, humanitarian risks, data protection, and escalation pathways? Is AI literacy integrated into existing mandatory processes (e.g. safeguarding, data protection, compliance)? Has AI literacy been considered for crisis-affected people directly?		

The SAFE AI Transparency Card: Key components

This snapshot provides an overview of what's captured in the SAFE AI Transparency Card. Further details and templates are available online at cdacnetwork.org/safe-ai.

Section	Subsection	Focus	Lead	Purpose/output
1. AI system snapshot	Overview and intent	System name, purpose, and scope	Programme and technical lead	Establish context, intent, and expected benefit/ harm.
	System stack and infrastructure	Components, vendors, hosting, and interoperability	Technical and procurement teams	Map dependencies and potential systemic risks
2. Humanitarian principles alignment and context	Context, participation, and humanitarian alignment	Community engagement, inclusivity, and adherence to humanitarian principles	Programme / humanitarian team	Ensure contextual fit and adherence to humanitarian principles
	Risk, impact, and mitigation summary	Identified ethical and operational risks, mitigations, and residual risk	Joint (all teams)	Integrate risk management to minimise risk of harm to users and communities and decision rationale
3. Data and model summary	Data and model summary	Data sources, model design, training, and performance	Technical team	Capture technical transparency
4. Deployment and lifecycle	Lifecycle and decommissioning	Development and deployment timeline, revision, and decommissioning plans	Technical team	Demonstrate technical and community accountability through the AI system lifecycle
5. Community in the loop and governance	Operational readiness, oversight, and accountability	Roles, ownership, review mechanisms, redress, and audits	Programme Lead with Legal/ethical	Demonstrate institutional legal and ethical accountability
6. References and versioning	Documentation references	Linked assessments, versioning, decommissioning plans	Joint (all teams) / oversight body	Provide audit trail, version management and continuity over time.

Appendix 2: How SAFE AI was developed

SAFE AI came into being during the most severe period of contraction the humanitarian sector has experienced in a generation. Between 2024 and 2026, traditional donor funding fell sharply, agencies undertook large-scale staff reductions, and operational presence in active crises was scaled back or withdrawn.

At the same time, AI deployment across the sector accelerated, driven by efficiency pressure and by tech-sector partnerships stepping into space vacated by public funding. The conditions under which a sector-wide governance framework could be developed were narrow, and the conditions under which AI was being adopted were not matched by equivalent investment in governance.

That context shaped the SAFE AI methodology and theory of change.

The framework's development was deliberately agile and evolving, designed to absorb learning from a fast-moving operational and policy environment rather than fix a single specification at the outset. The framework draft presented at the FCDO Durbar Court roundtable in March 2025 was materially different from the framework launched in May 2026, and that is by design.

AI use in creating the SAFE AI Framework

The SAFE AI Framework was authored by McElhinney, Spencer, Mazumder, Tjalve, and Madigan, supported by the consortium and expert community named below. Throughout the drafting process, AI tools were used for research synthesis, structural analysis, and editorial drafting support. Substantive intellectual contribution, strategic positioning, framework architecture, and final editorial judgement remained with the named authors. This note is included as a matter of consistency: the SAFE AI Framework asks humanitarian organisations to make AI use legible. Its own authors do the same.

Research, consultation and communities in the loop

SAFE AI was built through three commitments held constant across the development period:

- Original empirical research with crisis-affected communities.
- Sustained sector-wide practitioner consultation.
- SAFE AI pool of experts engagement spanning academia, tech, industry and multidisciplinary humanitarian operations.

The journey began at CDAC Network's event at The Alan Turing Institute's AI UK Fringe in 2024, where Dr Abeba Birhane (cognitive scientist, algorithmic bias researcher, and one of TIME's 100 most influential people in AI) delivered a keynote that set the frame for the project's approach to AI, power and accountability. A panel discussion followed, titled 'Do we need a humanitarian manifesto for AI?', in which Dr Birhane was joined by Sarah Spencer, Billy Perrigo (TIME Magazine), Foni Joyce Vuni (Refugee-Led Research Hub Nairobi) and Stella Suge (FilmAid Kenya).

The intellectual grounding for the community participation work was established at the CDAC Network Public Forum on 2 December 2024 at iHub Nairobi. There, Nyalleng Moorosi of the Distributed AI Research Institute delivered the keynote on 'Building for

ourselves' and panellists Patrick Gathara of The New Humanitarian, Monica Nthiga of Humanitarian OpenStreetMap, and Data Values Advocate Angeline Akai set out a Global South perspective on data, power and participation in humanitarian AI.

Two days later, fieldwork began in Kakuma Refugee Camp and Kalobeyei Settlement, conducted in partnership with FilmAid Kenya and trained local facilitators. That fieldwork engaged 193 participants across structured focus group discussions in December 2024. It is, to our knowledge, the first-time crisis-affected community testimony has been brought directly into the design of a sector-wide humanitarian AI governance framework. The findings are published in the academic paper From experimentation to engagement: on the paradox of participatory AI and power in contexts of forced displacement and humanitarian crises. The residents of Kakuma and Kalobeyei who gave their time, candour and expertise made that work possible. Their voices are documented in this film.

Sector consultation was delivered through eight bimonthly Communities of Practice, an expert pool methodology session in June 2025 and ongoing bilateral or small group discussions with expert pool members, a regional dialogue convened in Nairobi in June 2025 hosted by the Inter-Agency Working Group with local NGO leaders from Adeso Africa and WASDA, collaborations at panels at Nethope and Humanitarian Networks and Partnership Week with AccessNow, Nesta and UNHCR, and a hybrid roundtable at FCDO Durbar Court in March 2025 attended by c.90 participants from over 30 institutions. A design workshop at the Alan Turing Institute in October 2025 with media development and humanitarian organisations including Save the Children, the Norwegian Refugee Council, Internews, the British Red Cross, and Valent further grounded the framework.

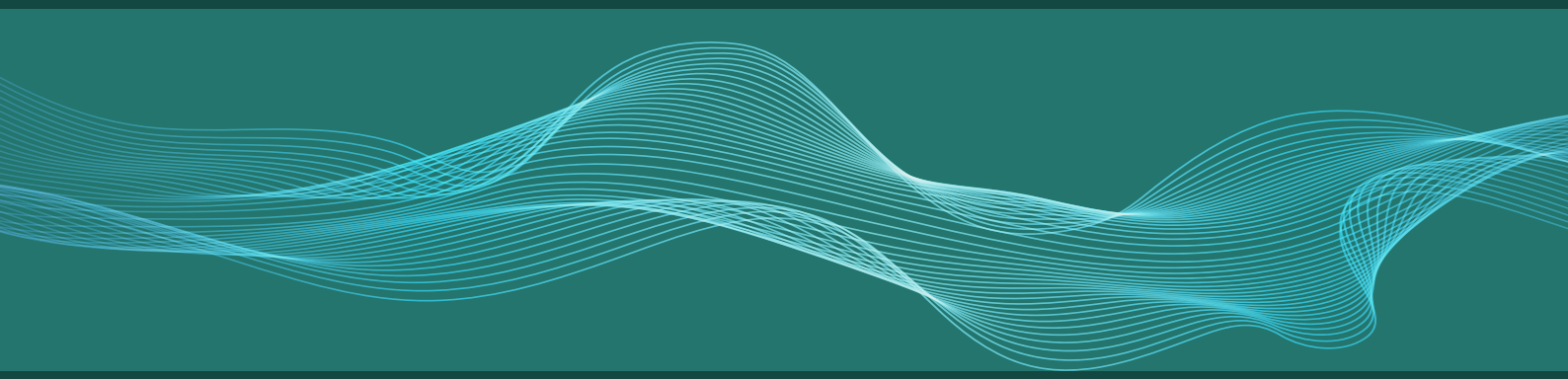
A session at UN-OCHA's Humanitarian Networks and Partnerships Weeks 2025 convened dialogue with Access Now, Nesta and UNHCR on responsible AI in humanitarian response, and a panel at the NetHope Summit 2025 – 'Operationalising Responsible AI to Protect and Empower Communities' – enabled members of the SAFE AI project team to examine AI-driven efficiencies, accountability, and the risks of introducing AI systems into complex humanitarian contexts. On the margins of the India AI Impact Summit, Shruti Viswanathan, a CDAC expert consultant, presented SAFE AI's community in the loop approach at the Participatory AI Research & Practice Symposium (PAIRS) India in 2026.

Acknowledgements

SAFE AI also rests on the time and expertise contributed by a broad pool of experts to whom we are very grateful: Practitioners, researchers, standards organisations, think tanks, tech and humanitarian journalists, media development actors, donor and industry experts who engaged across the development period

Zineb Bhaby – Norwegian Refugee Council; Zahira El Marzouki – Gates Foundation; Yolande Wright – Give Directly; Wendy Hall – UK AI Council and University of Southampton; Vera Lummis – Give Directly; Upol Ehsan – Northeastern University Harvard; Tim Davies – Connected by Data; Tamara Low – Save the Children; Sue Black – Durham University; Scott White - UK Foreign Commonwealth and Development Office; Stuart Campo – International Organisation for Migration; Steph Wright – Our AI Collective CIC; Stella Suge – FilmAid Kenya; Stefaan Verhulst – New York University, GovLab's Data Program; Shruti Viswanathan – CDAC; Shannon Vallor – University of Edinburgh; Roslyn Kratochvil Moore – Deutsche Welle Akademie; Robert Philips – UK Foreign Commonwealth and Development Office; Richard Lawne – FieldFisher; Ray Eitel–Porter – Intellectual Forum, Jesus College Cambridge and Lumyz Advisory; Rafik Elouerchefani – Organisation for the Coordination of Humanitarian Affairs (OCHA); Rachel Maher – OCHA; Philippa Spencer, BAE Systems; Pierrick Devidal – International Committee of the Red Cross; Nyalleng Moorosi – Distributed AI Research Institute; Noor Lekkerkerker – Upinion; Nasim Motalebi – World Food Programme (WFP); Muge Ozkaputen – BBC Media Action; Mirca Madianou – Goldsmiths, University of London; Megan McGuire – Médecins Sans Frontières; Matthew Smith – Canada's International Development Research Centre; Matthew Thomas – British Red Cross; Mathias Philip Reuter – Federal Foreign Office, Germany; Marcia Wong – CDAC Board and Private Sector Humanitarian Alliance; Lumi Tuchel – UK Foreign Commonwealth and Development Office; Linda Raftree – MERL Tech Initiative; Leonardo Milano – OCHA; Kristin Bergtora Sandvik – Peace Research Institute Oslo; Kim Malfacini – OpenAI; Julia Irwin – IASEAI; Jorge Fernandes – WFP; Jeremy Boy – United Nations Development Programme; Jeannie Annan – International Rescue Committee; Jatinder Singh – Cambridge University; Jacki O'Neill – Microsoft; Heather Leson – International Federation of the Red Cross and Red Crescent; Giulio Coppi – AccessNow; Geoff Loane – Former CDAC Chair; Gaia Marcus – Ada Lovelace Institute; Fahim Sultan Al Qasimi – Independent; Fergus Thomas – UK Foreign Commonwealth and Development Office; Elizabeth Shaughnessy – NetHope; Elizabeth Seger – Demos; Divij Joshi – Overseas Development Institute; David Sully – AdvAI; Daniel Bongard – Swiss Agency for Development and Cooperation; Colum Wilson – Independent Consultant; Christina Wille – Insecurity Insight; Chris Meserole – Frontier Model Forum; Charlotte Lindsey – Independent Researcher; Charles Antoine Hoffman – United Nations Children's Fund; Caroline Vuillemin – Fondation Hirondelle; Camilla de Coverly Veale – Mozilla Foundation; Brandon Sternquist – Goal; Aubra Anthony – Carnegie Endowment for International Peace; Amina Hamshari – UNESCO; Amil Khan – Valent; Allaine Cerwonka – Alan Turing Institute; Alexi Drew – Independent; Alex Ward – UK Humanitarian Innovation Hub; Aimee Ansari – CLEAR Global; Dr Abeba Birhane – Trinity College Dublin.

Any omission is unintentional.



SAFE AI

cdacnetwork.org/safe-ai