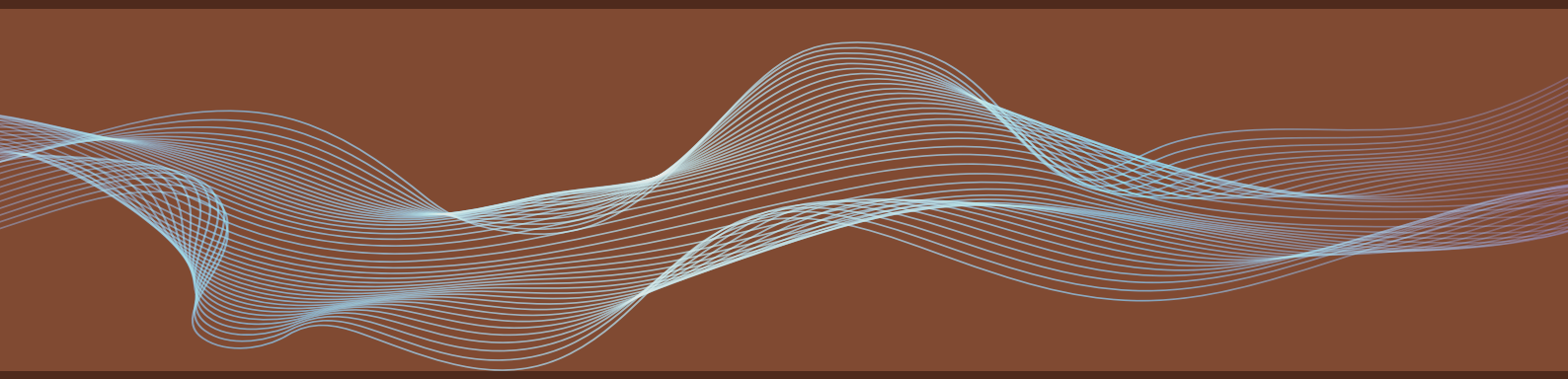
Standards and Assurance Framework for Ethical AI

SAFE AI

Tools and Guidance

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

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

Introduction

Why the SAFE AI tools are published digitally

In this document, we've compiled useful tools and more detailed guidance that can help you as you progress along your SAFE AI journey.

The latest version of the SAFE AI Tools and Guidance will be maintained and published digitally at **cdacnetwork.org/safe-ai**. This is a deliberate choice.

As no SAFE AI journey is alike, not all the SAFE AI Tools and Guidance will be used in the sequence they appear in the SAFE AI Framework, and not all will be necessary for every potential AI use case. Some, like the SAFE AI Transparency Card – the central record of your SAFE AI decisions and activities – you will need to return to repeatedly.

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 and guidance 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 and guidance that follow them. We also include a snapshot of the central governance artefact, the SAFE AI Transparency card, with full template available online.

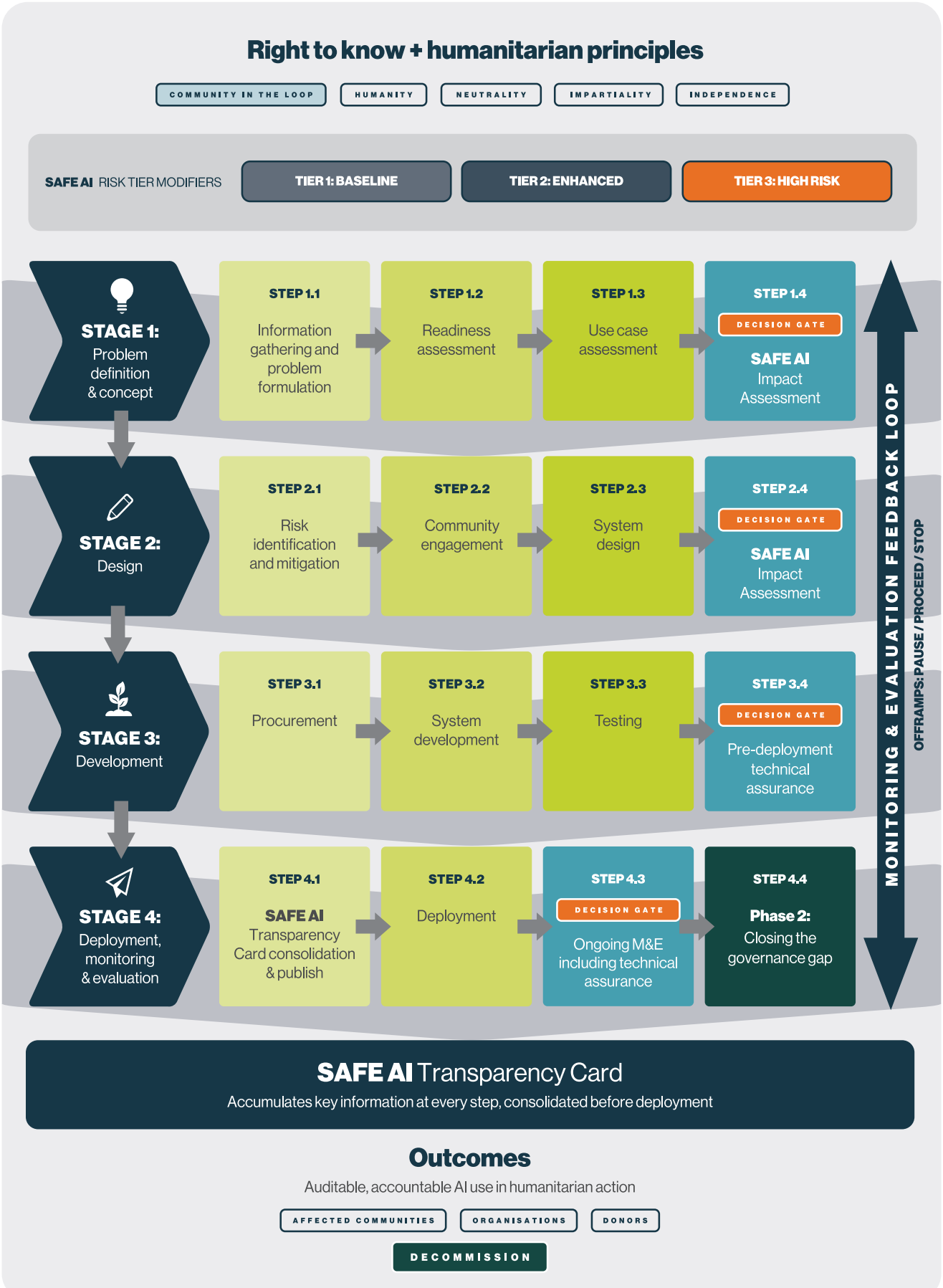
Digital release also continues to ensure the framework is participatory in its development. As organisations apply SAFE AI, they generate evidence about what works, what does not, and what is missing. That evidence informs revision. Each version of the SAFE AI Tools and Guidance is sharper than the last 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 *A Governance Framework for Humanitarians using AI: How to turn SAFE AI principles into practical action* sets out the case and the architecture. The digital tools and guidance set out in this document and on the SAFE AI website 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 Framework



You can find more details about the framework in *A Governance Framework for Humanitarians using AI: How to turn SAFE AI principles into practical action.*

Keeping communities in the loop through your SAFE AI journey

How to co-design AI with crisis-affected communities

This guidance sets out how to put community engagement into practice across the SAFE AI Journey. It complements Step 2.2 of the SAFE AI Framework and applies across all four stages, from problem definition through deployment, monitoring, and decommissioning.

Where the framework body establishes that community engagement is a governance requirement at Tier 2 and Tier 3, this guidance sets out what that requirement looks like operationally. Each section addresses a stage of the SAFE AI Journey. Readers working on a specific stage can land directly on the section they need.

The guidance is designed to draw on the AAP, CCEA, and community engagement work humanitarian organisations already do, not to require a parallel AI-specific community engagement infrastructure. What it adds is the AI-specific lens those existing functions need to govern AI deployment well. Each stage below identifies what existing community engagement work already covers and what the AI-specific addition is.

Community participation under SAFE AI is a governance instrument. It exists to surface intelligence that internal assessment cannot generate, and to hold deployment to account against the standards the framework requires. Where consultation happens without communities holding real decision-making authority, that is participation washing. The framework treats it as a governance failure, not as community engagement.

Throughout the SAFE AI Journey, participation decisions, risks, and mitigations should be recorded in the SAFE AI Impact Assessment. Key choices should be reflected publicly in the SAFE AI Transparency Card.

Working through existing AAP and CCEA structures

SAFE AI community engagement should be conducted through the AAP, CCEA, complaints, and community feedback infrastructure organisations already operate. The framework does not require parallel AI-specific community engagement infrastructure. Where collective AAP processes operate in the context, including inter-agency complaints mechanisms, country-level AAP coordination, and joint feedback systems through clusters or coordination structures, SAFE AI community engagement should connect to them. Multi-agency AI deployment creates accountability questions that individual organisational engagement cannot answer alone.

What SAFE AI does require is that existing AAP and CCEA work explicitly covers AI deployment. AI raises specific community engagement issues that traditional AAP frameworks may not yet address:

- The right to know that automation is shaping decisions about people.
- The comparability of AI standards across organisations operating in the same context.
- The consent question for AI uses of data originally collected for service delivery.
- The structural risk of participation washing.

Whether an AI system operates in the languages spoken by affected communities, whether its outputs reach people with limited connectivity, and whether it includes

or excludes people at the intersections of language, literacy, and infrastructure are core community engagement questions. They should be documented in the SAFE AI Transparency Card where significant. Tier 2 and Tier 3 deployments should treat these as conditions for deployment, not as enhancements to be added later.

The actions set out in this section assume that those engaging in co-design are willing to abandon an AI solution if either the community or humanitarian actor concludes it is no longer appropriate for the context. The actions are not exhaustive. They aim to enable the highest levels of community participation the deployment context allows, including co-creation, ownership, and authority.

SAFE AI journey Stage 1: Community engagement at problem definition and concept

This section is for confirming what the problem actually is, and what the conditions for any response would need to meet. Communities affected know the problem, how it shows up in their context, and what the priorities are. The framework treats this as authority.

What is new for AI. Existing problem identification already engages communities on what they need. The AI-specific addition is the explicit question of whether AI is the right tool, surfaced with communities before AI is presented as the option, and the right to refuse AI as a legitimate outcome.

Begin engagement before AI is introduced as an option. This stage is designed to surface whether AI is the right solution at all. Abandoning an AI approach on the basis of community input at this point is a legitimate and sometimes necessary outcome, not a failure. Where communities identify priorities AI cannot address, or where the failure modes of previous interventions suggest AI will not work in context, document the finding and apply responsible refusal.

Resource participation properly. Community participants should be compensated for their time and expertise. Budget for this at the outset, alongside translation, facilitation, and interpretation costs. Co-design that depends on unpaid community labour is participation washing in another form.

Build AI awareness and familiarity. Conduct information sessions and community-based workshops covering how AI works, where it fails, and the specific risks involved. Use scenario-based storytelling, metaphor, non-technical explanations, and demonstrations of AI tools. Communities cannot bring authority to a system they have not been given the means to understand.

Identify the constituent base. Work with local organisations, community groups, and leaders to identify individuals who may be directly or indirectly impacted by the AI solution, as well as those who will participate in or be consulted through the co-design process. Identify potentially marginalised populations at the outset and methods to engage them safely in non-stigmatising ways.

Map community-based knowledge and infrastructure. Surface what is already in place: technology familiarity, existing participation mechanisms, local AI work, and the expertise community members and local organisations already hold.

Work through existing mechanisms but remember that conversations with crisis-affected communities do not stay within neat lanes. Issues raised about other aspects of humanitarian action should be referred to the relevant actors and acted upon.

Establish and communicate a level of ambition. Define what is open to community decision-making: only programme decisions, or also AI infrastructure choices including data, model selection, and governance arrangements. Be clear about which decisions sit outside your authority to share. Aim for the highest level of community participation the deployment context allows. Communicate this publicly.

Clarify the scope of community decision-making authority. Define how much control communities have over AI usage, its parameters, and potential modifications. Determine

whether communities can object to, and trigger review of, AI design, deployment, or supplier selection decisions. Where supplier responses to community-driven changes are required, ensure these are set out in contracts (see SAFE AI Procurement Guide).

SAFE AI journey Stage 2: Community engagement in design

What this stage is for. Communities are co-designers of the system. By Stage 2, AI has been chosen as the response and the work is designing it. Decisions about how the system will function in context, what it must and must not do, how its outputs will reach people, and how it will be monitored once deployed are structural design decisions, not technical choices to be optimised by the deploying organisation alone.

What is new for AI. Existing community engagement in programme design already shapes interventions with communities. The AI-specific addition is co-design of system behaviour, red lines, the data the system uses, and the feedback loops embedded in the system itself.

Plan multiple co-design meetings and workshops. Work with a core group of constituents over time to agree priorities, aims, and expectations for the AI solution. Allow adequate time to collectively agree the purpose of the system, how it will work, and its anticipated benefits and risks. One-off consultation is not co-design.

Account for people the AI system affects who will not participate in co-design. AI systems often make decisions about people who never interact with the system directly, through inference, prediction, or analysis based on data they did not provide. The community participating in co-design may be different from the population the system affects. Identify these populations early and consider how their interests and rights are represented in the design decisions co-design produces. This is the right-to-know norm applied operationally.

Establish red lines. Define what the system must never do and what ethical risks cannot be tolerated. Examples include barring AI outputs from life-altering decisions, preventing contact with children and their personal information, or forbidding processing of personally identifiable information. Humanitarian actors and communities define these together, anchored in dignity, safety, and agency. Document the red lines and test them through the SAFE AI Impact Assessment.

Agree the key aim for the AI solution. Before development begins, align with communities on the system's primary purpose, who it serves, and how it functions in context. The aim must respond to community-identified needs, not impose a technology-first approach. Set practical, measurable objectives aligned with humanitarian priorities.

Identify risks at individual, community, and societal levels. Distinguish risks related to AI models, deployment contexts, and the humanitarian actors deploying the system. Work with communities to surface concerns early through scenario-based assessment. Embed monitoring that adapts as risks emerge. Develop mitigation measures alongside the risks identified.

Co-design feedback loops and embed them in governance. Build on existing feedback processes or create new ones to continuously surface community insights. Make the loops public where safe to do so. Be explicit about how feedback translates into design changes and what the communication protocol is with the technology partner.

Assess costs, benefits, and trade-offs with direct community participation. Cover financial costs, ethical trade-offs (efficiency gains versus risks to privacy, consent, or labour displacement), and operational feasibility. Name the expected benefits in concrete terms. Common tensions include efficiency versus equity, automation versus human oversight, and data-driven decision-making versus surveillance. Identifying tensions early lets humanitarian actors weigh compromises in favour of dignity, inclusion, and adaptability over technical convenience.

Ensure all materials and processes are accessible. Communicate in the languages and modalities communities actually use. Budget for translation, interpretation, and culturally

sensitive approaches including storytelling, multimedia, and metaphor. Account for lower literacy, lower digital familiarity, and the digital gender divide.

SAFE AI journey Stage 3: Community engagement in development, testing, and procurement

By Stage 3, the system is being built. Community engagement at this stage tests whether the design decisions agreed in Stage 2 hold up against the system as built, and surfaces failure modes that internal testing will not catch.

What is new for AI. Existing community engagement already validates approaches before scale-up. The AI-specific addition is community testing of the system as built, in operational languages, with feedback treated as governance evidence at the Decision Gate.

Treat community feedback during testing as governance evidence. The framework body is explicit on this point. Community feedback at Stage 3 is not user research and should not be designed around. It is governance evidence and should be recorded in the SAFE AI Transparency Card. Where testing surfaces concerns that cannot be mitigated, that is a finding for the Stage 3 Decision Gate, not a usability issue.

Test the system's performance in the operational languages of affected communities. Humanitarian operations disproportionately occur in multilingual environments where many affected populations speak low-resource or under-digitised languages. Most commercial foundation models are trained predominantly on English and a small number of other high-resource languages. Performance in low-resource and under-digitised languages can degrade in ways that supplier benchmarks will not show. Look for subtle mistranslations, dropped negations, missed cultural context, and outputs that are incorrect.

Communities who speak the operational languages are the source of intelligence here. They will spot inadequate performance in ways others cannot. Where the model's language coverage cannot meet requirements, reconsider your architecture choice or apply responsible refusal. Translation alone cannot fix a model that does not work in the language.

Surface failure modes through community testing. Misinterpretation in local languages, exclusion patterns visible only at the household level, and assumptions about documentation or registration status that do not match reality on the ground are governance failures that internal testing cannot generate. Community testing is the mechanism through which these are caught.

Verify procurement obligations reflect community priorities. Where suppliers have been engaged, confirm that contractual terms reflect the commitments made to communities during Stage 1 and Stage 2. This includes data ownership terms, right-to-audit clauses, supplier obligations on model change notification, and exit provisions. See the SAFE AI Procurement Guide.

Safe AI journey Stage 4: Community engagement at deployment and ongoing

Communities are co-decision-makers in an ongoing way. Decisions about whether to continue, adjust, escalate, pause, or decommission the system are made at the Decision Gates throughout Stage 4. Communities affected by the deployment bring intelligence that no monitoring system can replicate. They encounter the system in ways no other actor does.

What is new for AI. Existing AAP and CCEA work already operates feedback mechanisms through deployment. The AI-specific addition is community authority at Decision Gates to require the system to pause or stop, and community notification through decommissioning.

Launch with a small-scale pilot. Before full-scale deployment, test usability, impact,

and real-world effectiveness with crisis-affected communities. Use iterative pilots, with feedback shaping refinements before broader implementation.

Notify communities before deployment. Communities affected by the system should be notified before it goes live, 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 the SAFE AI Transparency Card should be available to communities at this point, not only on organisational websites.

Establish governance with transparent decision-making and accountability structures. Disclose key decisions about how the system is applied, updated, or modified. Give communities visibility into AI lifecycle changes and clear pathways for raising concerns. Define the scope of community decision-making authority at Stage 4, including the authority to trigger review at Decision Gates. Align feedback and redress with existing AAP systems where present, ideally through collective processes.

Actively raise awareness of the feedback mechanism. Feedback channels only work if communities know about them, trust them, and understand how to use them. Use multiple formats: community meetings, multilingual messaging, visual guides, direct outreach. Show how previous community input has shaped the system, particularly to reach those with lower digital familiarity.

Address challenges raised by community feedback. Integrate feedback into post-deployment refinements rather than treating it as one-time consultation. Maintain mechanisms for continual adjustment based on lived experience. Be prepared to redesign the system if feedback requires changes to the underlying model, the data, or the overall aim.

Treat repeated community concerns as Decision Gate signals. Where community input identifies failure modes that performance metrics miss, outputs that are systematically wrong for particular circumstances, or assumed consent where it was not given, this is a Stage 4 Decision Gate signal. Responsible refusal, including decommissioning a deployed system, is the framework's defined response.

Conduct regular performance assessments, audits, and model evaluations. AI models should be subjected to continuous performance assessments to ensure they remain accurate, ethical, and contextually relevant. Regular stress testing helps identify vulnerabilities before they lead to harm. Routine model evaluations should assess bias drift, performance degradation, and unintended impacts on crisis-affected communities. Results should be made public and shared with humanitarian actors and communities, supporting accountability and refinement. See SAFE AI Technical Assurance and the SAFE AI Transparency Card.

Build local capacity for long-term oversight. Ensure crisis-affected communities and local humanitarian actors understand the system, its risks, and its limitations. Invest in technical training, public awareness, and skill development so community participation in monitoring and evaluation is informed and sustained.

Commission third-party evaluations and independent assessments. Engage external experts, ethicists, or independent evaluators alongside community members to assess the system against its stated goals. These should include community interviews, bias assessments, and technical audits, supporting accountability beyond internal project teams.

Share learnings with the sector and with communities. Document successes, failures, and lessons learned. Make insights publicly available so other humanitarian actors and community groups can build on them. Coordinate data from feedback mechanisms and improve accountability through participatory design. Listen to communities at every stage of evaluation.

Sustain feedback mechanisms for iterative improvement. AI solutions must remain adaptable to evolving needs. Test and refine long-term feedback loops where communities can raise emerging concerns, ethical risks, or contextual changes that affect usability.

Engage communities through decommissioning. When a system is decommissioned, whether at planned end of life, in response to supplier exit, or because monitoring has shown the system should not continue, communities should be informed through the same channels used at deployment. The notification should make clear 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 their existing access to assistance is protected through the transition.

Retroactively embed co-design in existing AI deployments. Where AI is already in use without co-design having been conducted, retrofit co-design processes by integrating community engagement into governance and oversight. Even where the system cannot be restructured, applying co-design principles ensures communities have authority over adaptation and refinement.

The conditions for community in the loop decision-making authority

For Tier 3 deployments, if co-design – where communities hold real influence over whether the system is built and how it operates – cannot be achieved, that is a governance red line. Responsible refusal applies. Deployment should not proceed until the conditions for community authority are in place. See When to reassess your SAFE AI tier in the SAFE AI Framework. The conditions for genuine community decision-making authority are:

- Decision-making is documented, with rationale in the SAFE AI Transparency Card.
- Communities have access to the information they need to make decisions, including on the system's design, performance, and limitations.
- Where the decision is shared between the deploying organisation and communities, the basis on which it is shared is explicit.
- Communities have routes to escalate concerns when the organisation is not honouring its commitments.
- Communities have the authority to trigger review at Decision Gates, including the authority to require the system to pause or stop.
- Where the decision is the organisation's alone, that is stated and the reasoning given.

Where these conditions cannot be met, the SAFE AI tier should be reassessed and responsible refusal considered.

How to mitigate common community co-design risks

The risks below are not risks related to AI. They are risks that arise within the co-design process itself. Identified risks should be recorded in the SAFE AI Impact Assessment, alongside mitigation measures and decision points.

Potential risks	Mitigation measures
The constituents engaged in the co-design process may not adequately reflect the diversity of experiences,	Identify potentially marginalised populations at the outset and find methods to safely consult with and engage them in non-stigmatising ways. Use multiple means of engagement to capture a broader range of perspectives. Regularly reassess representation gaps and adjust outreach and engagement approaches accordingly.

Communities do not fully understand how the AI solution works, including its constraints, risks, and potential consequences.	Use culturally relevant and plain language to explain technical and non-technical aspects. Provide visual and interactive demonstrations of how the AI solution functions. Clearly articulate expected outcomes of using AI, including benefits and potential harms, supporting informed decision-making.
The feedback mechanism established to solicit inputs from the community fails to work as intended, or feedback is either not included or only partially included.	Use both digital and non-digital methods of communication to improve accessibility. Identify trusted local intermediaries to help gather, translate, and relay community feedback. Regularly assess the effectiveness of feedback loops and iterate based on engagement levels and response quality.
Power dynamics between humanitarian actors and communities may skew decision-making, undermining meaningful participation.	Establish a governance structure that shares power equally between community representatives and the humanitarian actor deploying the AI solution. Define clear measures that ensure equitable influence over AI design and implementation, with options for community recourse when these measures are not met. Prioritise transparency in decision-making, documenting and clearly communicating choices and rationale.
Changes to the design of the AI solution are not sufficiently or consistently shared with communities, degrading trust in the solution and engagement	Develop and agree a protocol on how information will be shared about the design, modification, or improvement of the AI solution, including method and frequency. Inform co-design constituents about forthcoming modifications in advance and invite input. Maintain clear documentation and an accessible record of how the AI solution was designed and the decisions made.
Co-design discussions are designed for a predetermined outcome or limit people's ability to meaningfully contribute to the process.	Ensure early-stage conversations focus on community priorities and facilitate open-ended discussions rather than structuring them around predefined AI use cases. Empower communities with decision-making authority and define clear pathways for communities to challenge, redirect, or reshape AI solution objectives based on their lived experiences. Build in ways to incorporate feedback on both the AI solution and the co-design process itself.
Humanitarian actors fail to sustain communities' expectations of shared decision-making, influence, and control of co-design processes, eroding trust and reducing engagement in future co-	Set realistic and transparent expectations about the level of decision-making power communities will have throughout the co-design process. Build on existing or new long-term participation mechanisms, including community advisory boards, governance committees, and iterative engagement. Develop accountability mechanisms, such as independent oversight structures or co-design audits, that enable communities to evaluate whether humanitarian actors have honoured participatory commitments.
Co-design processes increase communities' exposure to AI without improving their understanding of how AI functions, its limitations, and its potential dangers	Introduce long-term AI literacy initiatives that equip communities with a clear understanding of how AI functions, including its strengths, limitations, and ethical concerns. Use real-world examples, accessible language, and participatory workshops that actively engage communities in discussions about AI's potential risks including bias, automation pitfalls, and unintended impacts on decision-making.
Participation increases protection risks, including direct threats to participants, increased surveillance, or reprisals.	Conduct scenario-based risk assessments with local partners and protection actors. Identify power dynamics, surveillance threats, and potential triggers for reprisals before initiating any participatory activities. Offer remote, pseudonymised feedback mechanisms, including secure messaging or community intermediaries, and avoid public documentation that could identify contributors without informed consent. Design safety protocols with both communities and humanitarian protection actors. Ensure emergency response or legal support mechanisms are in place and aligned with existing community safeguarding strategies or referral pathways.

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 set out in *A Governance Framework for Humanitarians using AI: How to turn SAFE AI principles into practical action*, the 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, and 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.

Section 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?		

Section 2: Leadership and accountability

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

Confirm cost and risk awareness	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.

Section 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)?		

Section 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?		

Section 5 - Assembling the right team

Assembling the right team, assigning roles and equipping them 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 Use Case Readiness Checklist

Once you have identified the problems you want to try and solve using AI, it's time to reflect on which use cases have the most potential for your organisation and its specific circumstances.

Section 1: Identifying your use case

Identifying use cases

Question / check	Yes / no / unclear	Notes / actions needed
Have you identified which specific AI capabilities are relevant to solve the problem(s) you've outlined in the SAFE AI Framework's Problem identification step?		
What is possible with the proposed solution that the current approach doesn't allow?		
Do you have evidence to support this claim?		
Have you critically evaluated sources of evidence for claimed productivity gains?		
Have you performed a feasibility study to determine whether the proposed AI implementation is practical, achievable, financially viable, contextually appropriate, and responds to an identified community need?		
Would the proposed AI approach replace valuable human-to-human contact?		
Based on the above, have you considered value for money and opportunity cost for deployment?		

Section 2: Ensuring participatory and compliant data stewardship

Confirm you are ready for data compliance

Question / check	Yes / no / unclear	Notes / actions needed
Do you know which data your organisation has access to?		
Do you know where it lives?		
Do you know who has access?		
Are you clear on its intended and potential uses?		
Does it contain any personally identifiable information (PII) or sensitive information?		
Do you know which data will be used with the AI solution?		
Will any data be used to train a model or just to interact with the model?		
Given the intended use cases, do you need any additional data?		
Have you defined a clear process for acquiring or collecting this data?		
Is there a risk to individuals that a technology company has access to and control over collected data?		
If data collected from communities is reused or combined for AI-enabled analysis, would people still expect and accept this use?		
Are there clear routes for communities to understand, question, or challenge how their data is being used?		
Does your organisation know when continued data reuse should trigger renewed engagement?		

Section 3: SAFE AI use case go/no go decision matrix

Confirm if your use case is ready to progress

Question / check	Yes / no / unclear	Notes / actions needed
Do you have a clearly defined humanitarian problem where AI can play a role?		
Can you describe how and why AI will provide value over current approaches?		
Have you identified a multi-disciplinary team to conduct the SAFE AI Impact Assessment?		
Have you identified individuals responsible for reviewing risks and impact throughout implementation?		
Does your organisation have appropriate human resources, budget, and expertise?		
Do you have a plan for engaging affected communities?		
Does your organisation have a governance group with senior leadership support?		
Do you have a clear estimate of total budget and resources required, and do you have that budget?		
If the system uses community data, are humanitarian data protection protocols in place?		
If the system processes sensitive data or PII, are legal and GDPR protections in place and understood?		

The SAFE AI Impact Assessment

You should conduct a SAFE AI Impact Assessment twice in your SAFE AI journey.

The first time, before you commit to AI. You answer a short set of questions to decide whether AI is the right tool for the problem you are trying to solve. If you decide it is not, you stop and use a different approach. If you decide AI is right, you carry on.

The second time, before you build or deploy. You answer a longer set of questions to confirm the design is sound and the safeguards are in place. If you find problems you cannot fix, you redesign, pause, or stop.

How much of the this section need depends on the risk of what you are doing:

- Lower risk use case (internal, advisory, easily reversible): you only need the first round.
- Medium risk use case (AI shapes operational decisions, with people overseeing it): you need both rounds.
- Higher risk use case (AI directly affects people's access to assistance, protection, or information): you need both rounds, with more depth and outside expertise.

If you are not sure which risk tier your use case falls into, apply the higher one.

Who should be in the room

Stage 1, lower risk: A small team. Someone with technical or AI knowledge, someone with humanitarian programme experience, and your data protection lead. **Medium or higher risk:** The same plus community engagement, protection, and safeguarding leads.

Stage 2, medium risk: All of the above plus legal and policy input. **Higher risk:** all of the above plus external advisors with technical, legal, or protection expertise. You should also have documented evidence from your co-design work that affected communities have shaped the design.

How to run the session

Host a workshop at each round. Discuss the questions together, in plenary and in breakout groups. Have someone capture the answers and any actions or mitigations. Before you start, describe in plain language what the AI system will do and who it will affect. This becomes the foundation record for your SAFE AI Transparency Card.

Keep humanitarian principles top of mind throughout. They are the test for whether a proposed system belongs in humanitarian work.

The questions are not a checklist. Where you find a problem you cannot fix, the right answer is to redesign, pause, or stop. Documenting a decision not to proceed is a successful outcome of a SAFE AI Impact Assessment.

What SAFE AI Impact Assessement covers and what it does not

Your SAFE AI Impact Assessment addresses the questions specific to AI: how AI systems behave, the harm they can do at model level, the inferential harms they can cause to

communities, and the question of whether AI is the right tool at all. It is designed to build on the work your organisation already does.

Most of what you need to answer the questions here can be found in documents your team already uses or produces: E.g. Humanitarian Needs Overview and Humanitarian Response Plan; Multi-Sector Needs Assessment findings; Country context analysis; Community priorities and feedback; Innovation strategy and data strategy; Protection assessments and protocols; Gender analysis; Safeguarding assessments and protocols; Conflict sensitivity analysis; Political economy analysis; Risk and impact matrix.

You should not need to create new evidence from scratch. The work of SAFE AI Impact Assessment is to apply an AI-specific lens to what you already have.

You may also need other assessments alongside this one

For most use cases, you will typically need a **Data Protection Impact Assessment (DPIA)** alongside this Impact Assessment.

The two do different work. A DPIA addresses how data is collected, stored, accessed, and protected across your AI system. It does not assess what the AI does with the data, whether outputs are accurate and unbiased in context, how AI-shaped decisions can be explained to affected people, or whether those decisions can be challenged. Those questions are answered by SAFE AI Impact Assessment. Both are needed.

For Tier 2 and Tier 3 AI systems, your organisation is typically in the role of "Data Controller" and the DPIA work usually involves project leads, IT, legal, and a Data Protection Officer if one exists.

The DPIA asks you to think about how data is used, accessed, and protected across the whole life of your AI system, and to identify the risks. The questions a DPIA typically covers include:

- What safeguards are in place to protect data and systems?
- Is a Data Protection Officer engaged to support the AI implementation?
- Are data protection policies and safeguarding protocols in place, and are staff trained to implement them?
- Do you have a data governance process that can monitor the system and its behaviour over time?
- How will data retention limits be decided and acted upon?
- How will informed consent for data collection be achieved?
- Do your data retention guidelines limit storage of data to only when it is strictly necessary?
- Will the implementation access or process sensitive data?
- Could your anonymised data be re-identified if combined with other data sets?
- Where are user inputs and model outputs stored?
- Does the implementation rely on live data signals?

As you work through these questions, the SAFE AI Framework also asks you to think about **participatory data stewardship**: how decisions about data use can shift power to people affected by crises.

People whose data is being used participants with a stake in how the data shapes decisions about them.

For higher risk use cases, you may also want to conduct a **Human Rights Impact Assessment (HRIA)** to address broader human rights implications. Useful references include Business for Social Responsibility's *A Human Rights Based Approach to Impact Assessment*, the Ontario Human Rights Commission's *Human Rights AI Impact Assessment*, and the Government of the Netherlands' *Impact Assessment Fundamental Rights and Algorithms*.

The environmental impact of AI deployment including the energy and carbon costs of training, fine-tuning, and operating models is increasingly a governance consideration. For environmental impact, the **ML CO2 Impact Tool** provides a calculator for estimating the energy and carbon costs of training and running models.

Keeping communities in the loop

Engage affected communities directly in the problem identification work that feeds into stage 1, through participatory methods and in partnership with local groups. Include the right-to-know, redress, and accountability questions in that engagement, and feed the outputs into your assessment.

Work through your existing community participation mechanisms wherever possible. Add new processes only where necessary. The aim is to avoid technology-first solutions that do not address the actual problem, and to keep open the option of a non-AI solution where it would work better.

For medium and higher risk use cases, your stage 2 work must include documented evidence that affected communities have meaningfully shaped the system's design. If that evidence does not exist, the design is not ready for Round 2.

See How to Co-design with Crisis-affected Communities for stage-by-stage guidance.

Stage 1: Before you commit to AI

Required for all use cases.

Is the problem real, and is AI the right tool?

This tests whether the problem is real, the data foundation is adequate, and AI is the right tool. The output is a decision: proceed, redesign, or stop.

	Yes / No	Notes / Mitigations
Has this problem been identified through your existing needs assessment work, not invented to fit an AI solution?		
Can you describe the problem in plain language without mentioning AI?		
Have you completed the SAFE AI Organisational Readiness Checklist, and have you addressed or scheduled the gaps it surfaced?		

Is the data you need adequate, including its quality, how representative it is, and whether you have consent for the proposed AI use?		
If AI extends the use of existing data beyond what people originally agreed to, have you put in place a way for people to be told and to opt out without penalty?		
Have you considered the dual-use risk, particularly where data the operation produces could be repurposed against the people it is meant to serve?		

If any of the following could happen and cannot be prevented, the use case does not proceed:

	Yes / No	Actions required
Red line: Could the system generate or materially increase protection or safeguarding risks to affected people or staff?		
Red line: Could the system plausibly generate or amplify dangerous misinformation or harmful guidance at scale without reliable safeguards?		
Red line: Could this system systematically exclude or disadvantage specific groups from assistance, protection, or information in ways that cannot be detected and corrected?		
Red line: Would reliance on suppliers, data providers, or infrastructure undermine organisational independence or create unacceptable conflict-related exposure?		

Accountability and the right to know

	Notes and actions
Could the system undermine accountability to affected people? Consider transparency, explainability, community participation, and feedback.	
What is the cost of error if outputs are wrong or misleading, and who could be harmed?	
Does the system work in the operational languages of the people it will affect? AI systems trained predominantly on high-resource languages may degrade significantly in low-resource and under-digitised languages, in ways	
Is there a real way for people affected by the system to challenge a decision it has shaped, including people who never interact with the system directly?	

What benefit do you expect, and what is the evidence?

	Notes
What specific humanitarian benefit do you expect this AI system to deliver? Describe it in concrete terms, with measurable outcomes where possible.	
What evidence supports the expectation that AI will deliver this benefit in your operational context, as distinct from supplier claims or comparable deployments in different contexts?	

What conditions need to hold for the benefit to be realised, including staff capacity, change management, integration with existing workflows, supplier reliability, and ongoing resources? Are those conditions in place or planned?	
What is the cost of pursuing this AI use case (direct costs plus the staff time, infrastructure, and oversight costs identified in the SAFE AI Organisational Readiness Checklist), and does the expected benefit justify that cost?	

Deciding whether to proceed

You should not proceed if any of these apply and cannot be fixed:

- There is no meaningful way for affected people to challenge or seek redress.
- Accountability cannot be clearly assigned to a named person or function.
- Data use stretches beyond what people would have agreed to, with no way to fix it.
- The system would significantly affect people who cannot opt out or refuse.
- AI would replace human judgement in ways that matter.
- The main reason to use AI is donor expectation, urgency, or cost pressure rather than evidence that AI will help.

A documented decision **not** to proceed is a successful outcome. Record:

- Why you decided not to use AI.
- What alternative approach you will use instead.
- What would need to change for AI to be reconsidered in future.

The outputs of Stage 1 feed into your SAFE AI Transparency Card. Specifically: system snapshot, rationale, initial risks, governance arrangements.

If your use case is lower risk, this completes the work. You should now complete the Lean SAFE AI Transparency Card. If your use case is medium or higher risk, continue to Stage 2 after you have completed your design and co-design work.

Stage 2: Before you build or deploy

Required for medium and higher risk use cases. Complete before development begins or deployment proceeds.

This stage examines what has been designed – the risks identified, the safeguards in place, the community engagement work conducted, the architecture chosen, and the oversight set up. The output is a decision whether to proceed, redesign, pause, or stop. Before you start, check your Stage 1 conclusions still hold. If the design has changed the risk picture, return to the Stage 1 red lines.

What kind of AI are you actually building?

	Notes and actions
Which type of AI system have you chosen: locally built, frugal or small AI, open-weight, or commercial foundation model? How did you make that choice? If the choice was assumed rather than deliberate, go back to the SAFE AI Architecture and Tech Stack Decision Guide before going further.	
Where the system uses a foundation model or external supplier, what can you independently inspect, and what must you take the supplier's word for? When something goes wrong, who carries the accountability?	
Has the architecture been tested against the languages the affected populations actually speak? If the model does not cover those languages well, reconsider the architecture or stop.	

Protection and safeguarding

This focuses on AI-specific protection questions. Wider protection considerations are tested in other subsections that follow, particularly right-to-know in Accountability, and through your protection assessments and safeguarding protocols.

	Notes and actions
If the AI gets it wrong, who could be harmed and how serious could that harm be?	
How are affected people protected from outputs that could put them at risk, including unsafe advice, dehumanising treatment,	
What protection risks could the system create for people indirectly, including people who never see its outputs?	

Fairness and inclusion

	Notes and actions
Where and how could bias show up in this AI implementation?	
How will you detect and address poor performance or unintended behaviour affecting particular groups?	
Which groups might be excluded because of language, connectivity, disability, gender norms, displacement, or documentation status?	

Data privacy and security

	Notes and actions
What data will the AI system access or collect from users or affected communities?	
Why is this data needed? Would your objectives fail without it?	
Will personal data be protected and anonymised, and could anonymised data be re-identified if combined with other data sources?	
How will you get and maintain informed consent for data collection, storage, and use, including where AI extends the use beyond what people originally agreed to?	

Conflict sensitivity

	Notes and actions
Which groups in the community could this AI system advantage or disadvantage, and could that cause tensions?	
Could data sources, suppliers, hosting, or deployment make your organisation look aligned with one side in a conflict, or actually create that alignment?	
Could any of your suppliers themselves be adversarial actors, through state pressure, coercion, or deliberate manipulation of inputs or	
Could outputs or data be intercepted and used for surveillance, coercion, or political inference?	
Where trade-offs exist between neutrality and delivering assistance, is the rationale explicit and are safeguards in place?	

The wider information environment

	Notes and actions
Does this AI system operate in an environment where other AI systems are also active, including ones outside humanitarian control? If multiple systems are running in the same context, what is the risk of conflicting outputs, AI-derived outputs being treated as independent verification, or trust in humanitarian information eroding?	
What does this system rely on for its inputs, and how confident are you those inputs are not compromised, through error, bias, or deliberate manipulation such as data poisoning or coordinated disinformation? Compromised inputs produce compromised outputs in ways monitoring does not always catch.	

How can affected communities flag when outputs do not match their experience, when the system is not working in their language, or when AI content is circulating around them outside humanitarian channels?	

Accountability to affected people

	Notes and actions
How will affected people know that automation is shaping decisions about them, understand how those decisions are made, and have a way to challenge them? This is a right-to-know question, not a disclosure choice.	
Are system outputs explainable to affected people in plain language and in the languages they speak?	
Are there clear, working mechanisms for people to challenge or correct AI-influenced decisions, including for people who never interact with the system directly?	
Where shared accountability structures already exist in the context, such as inter-agency complaints mechanisms, coordinated AAP, or joint feedback systems, how does this AI system connect to them? Where multiple agencies are using AI in the same context, affected people have a right to comparable accountability across all of them.	
Who else is involved in the system, including suppliers for data, cloud, or other services?	
Who is responsible when the model is inaccurate or produces an outcome someone wants to challenge?	
How will complaints about AI outputs be handled? Does your existing Feedback, Complaints and Response Mechanism need to be adapted?	

Community participation in decision-making

	Notes and actions
How have affected communities shaped the design of the system, with documented evidence of meaningful influence?	
How were affected communities involved in decisions about benefits and risks, before those decisions were made rather than after?	
How will affected communities give informed consent in relation to the system, including the AI literacy support they need to do so meaningfully?	
Where AI-generated outputs reach affected communities directly, for example through a chatbot, what authority do communities have over whether and how that happens?	
What AI literacy support do affected communities need to use the system safely and ask questions about how it makes decisions?	

Please note

The questions below look more technical but need both technical and humanitarian expertise to answer well. Keep your team working together on them.

Whether the model is fit for purpose

	Notes and actions
Is the chosen model genuinely fit for your specific use case?	

Can you secure access to evidence that confirms the system's performance claims? Will the evaluation data be different from the training data, representative of the populations you will deploy against, and available to your team or to an independent reviewer? If access cannot be secured, record this as a finding and surface it before deployment.	
Does the AI system need an internet connection to work? What happens if you lose connectivity? Would a hybrid approach with partial offline functionality work?	
Has the model been tested in the languages affected communities actually speak? If language coverage is inadequate, reconsider the architecture or stop.	

Decommissioning plan

	Notes and actions
What is the planned process for decommissioning the system, and what would trigger it?	
What is the plan for data your organisation holds after the system is decommissioned, including affected people's data, model outputs, and assessment records?	
Who is responsible for deletion or anonymisation, by when, and how will affected communities be told?	
What after-effects could there be in affected communities once the system is gone, and how will you monitor those?	
If affected communities rely on the system for a service, what will replace it at the point of decommissioning?	

Impact on staff and local actors

	Notes and actions
If the system makes significant efficiency gains, is there a job re-skilling programme in place for your staff?	
If you reduce staff in-house or with external partners, what local knowledge or local employment do you lose?	
Could a smaller staffing footprint open up new local partnership opportunities?	
If the system gets significant use, are you set up to absorb the ongoing cost and the scale-up or maintenance needs?	

Deciding whether to proceed

You should **not** proceed if:

- The architecture choice cannot support the use case responsibly, and a viable alternative pathway does not exist.
- Accountability cannot be clearly assigned, including for harms higher up the supply chain.
- Language coverage cannot meet the operational requirement.
- Co-design has not produced meaningful community influence over the design.
- A red line from Stage 1 has been triggered by what has been designed.

A documented decision **to stop**, or **to decommission** an already-deployed system, is a valid outcome. So is a decision to return to architecture or design before going further.

The outputs of Stage 2 feed into your SAFE AI Transparency Card. Specifically: detailed risks and mitigation, data and model summary, governance and oversight, and decision record and version history. Once the Stage 2 Decision Gate has been passed, the use case is ready to move into development.

SAFE AI Risk Mitigation Strategies

Use this guide when you have identified a specific risk in your AI use case and need a starting point for how to address it. This section sets out mitigation strategies organised by SAFE AI tier and by risk dimension. Work through it in three steps:

1. **Confirm your provisional SAFE AI tier.** The summary table below is enough to navigate this section. Full tier characteristics are in the SAFE AI Framework.

2. **Locate the risk** dimension you are addressing in the mitigation table for your tier. The risk dimensions used here are the same ones used in the SAFE AI Impact Assessment.

3. **Apply the mitigation guidance** proportionate to your tier, working downwards. Tier 1 guidance applies at every tier. Tier 2 guidance adds to Tier 1. Tier 3 guidance adds to Tier 2.

Consider this section as reference material rather than a structured exercise. Practitioners typically return to it more than once as risks emerge through the SAFE AI Journey.

Who should be in the room

Risk mitigation work is led by the SAFE AI project team, with input from programme, technical, data protection, safeguarding, and community engagement leads.

For Tier 2 and Tier 3 use cases, decisions about which mitigations to apply should involve the named accountable authority, with risks and mitigations documented in the SAFE AI Impact Assessment and recorded in the SAFE AI Transparency Card.

Risk mitigation does not sit with a single function. Choices about mitigation reflect trade-offs between operational, protection, accountability, and resource considerations that no team can weigh alone.

When to use it

This guidance is used iteratively across the SAFE AI Journey:

At Stage 1, to test whether the residual risks of a proposed use case are mitigable at the tier you are working at, and whether responsible refusal applies

At Stage 2, to identify the mitigation strategies your design and architecture choices need to address

At Stage 3, to confirm that the mitigations agreed at Stage 2 have been implemented before deployment

At Stage 4, to revisit mitigations when monitoring or community feedback surfaces new risks, or when the SAFE AI tier needs to be reassessed

The higher the tier of your use case, the more present the option of responsible refusal should be.

Where a risk cannot be mitigated at the tier you are operating, the framework's defined response is to redesign, pause, or stop. A documented decision to apply responsible refusal is a SAFE AI outcome.

For organisations with established risk management practice, existing risk management continues to handle organisational, programmatic, and operational risk. The SAFE AI tier logic and mitigation strategies add the AI-specific risks that may not be captured by existing risk registers:

- Dynamic risk where AI behaviour changes over time.
- Dual-use risks where AI systems can be repurposed beyond their intended scope.
- Surreptitious experimentation risks where AI is used to make decisions about people who never interact with the system.
- Bias-pattern risks that may emerge or change after deployment.
- Cumulative risks where multiple AI systems deployed in the same context produce conflicting outputs, false redundancy through AI-derived outputs being read as independent verification, or degraded trust in humanitarian information sources.

These risks may not be captured by existing risk frameworks because they are AI-specific, dynamic, and increasingly systemic.

SAFE AI's tiered approach is consistent with Adaptive Risk Management (ARM) logic, which applies proportionate governance responses based on risk level rather than uniform requirements across all situations.

Organisations already familiar with ARM frameworks will recognise the underlying logic. SAFE AI operationalises that logic specifically for AI deployment in humanitarian contexts, with community accountability and humanitarian principles as the governance foundation.

Locating your use case in the SAFE AI tiers

Before working through the mitigation strategies below, confirm your provisional SAFE AI tier.

The full tier characteristics are set out in the SAFE AI Framework. The summary below is enough to identify the appropriate level of risk mitigation. If in doubt about which tier your use case falls into, always opt for the higher tier.

Tier	Use case characteristics	Primary risk profile
TIER 1: BASELINE	Lower-risk, enabling uses. Primarily internal. Outputs are advisory and easily reversible. Does not directly influence access to assistance, protection, or services. Does not require beneficiary data.	Data leakage or misuse; over-reliance on incorrect outputs; informal or inconsistent staff use.
TIER 2: ENHANCED	Programmatic decision-support. Influences prioritisation, targeting, or operational choices. Human oversight retained, but AI shapes decisions. Requires affected communities' data.	Indirect exclusion or bias; automation bias eroding human judgement; opacity in decision-making; accountability gaps; trust erosion if outputs are wrong or misunderstood.
TIER 3: HIGH-RISK	Rights-impacting or community-facing uses. Directly affects people's access to assistance, protection, or accurate information. High cost of error. Operates in sensitive or conflict-affected contexts. AI output may be used without human oversight. Requires affected communities' data.	Direct exclusion from life-saving support; protection and safeguarding harms; misinformation or unsafe guidance; neutrality and independence risks; legal and reputational exposure.

Warning

Internal use alone does not guarantee Tier 1.

Beware of interpreting too bluntly here.

You will need to think about both the inherent risks of the capability, how people might interact with it, and about how the AI output is to be used.

For full tier characteristics including staff involvement, illustrative use cases, and the SAFE AI tools required at each tier, see the SAFE AI Framework.

Risk mitigation for Tier 1 use cases

Risk dimension	Mitigation guidance
Protection and safeguarding Example: A staff member uses a free AI drafting tool to write a high-level report on the protection situation accidentally includes enough identifying data on community members for them to be reidentified and located; the tool's terms permit the vendor to store inputted content. Example: Writing an internal briefing note, a staff member inadvertently includes the name and details of a colleague involved in a sensitive HR matter which are stored on the commercial AI system overseas.	■ Ensure the AI tool does not collect or process sensitive personal data without explicit consent. ■ Work with protection and gender colleagues to apply an intersectional do-no-harm review before deploying any tool. ■ Assess whether AI outputs, or the data used to generate them, could be, a) used to identify, target, or exclude individuals from aid, b) intercepted or misused by hostile or unauthorised parties in your operating context (conflict sensitivity check).
Information reliability, including misinformation, synthetic media, narrative manipulation, over-reliance on incorrect outputs. Example: generated summary of report may be incorrect, miss or mispresent key points. Example: A staff member shares an AI-generated summary of a situation report with colleagues as a basis for a movement planning discussion. The summary omits a key development, and the error is only caught when a team member checks the original document.	■ If output is assumed correct and acted upon as if correct, any process or individual downstream will suffer from this. ■ Use as is for triaging of documents, e.g. read summary and discard from downstream usage. ■ If used downstream, manually verify accuracy of content, incl. confirm sources. ■ Human oversight is recommended but not mandatory. The AI capability can act autonomously, and the AI output can be presented to the end user without human intervention. This also applies to the use of Agentic AI.
Technical reliability, including technology effectiveness and limitations Example: An AI document analysis tool used by the grants team fails following a routine software update, disrupting internal reporting workflows and requiring staff to revert to manual processing at short notice, causing delays to local partners receiving funds and delaying services to communities.	■ Inform your users and those depending on the service about the issue while you work on improving the technology implementation. Deploy the update when available.

Bias, representation, and exclusion Example: An AI tool trained primarily on English-language data performs significantly worse for speakers of minority languages, producing lower-quality outputs when summarising qualitative key informant interview reports. Example: A document summarisation tool consistently underweights vulnerability indicators more common in female-headed households, skewing internal risk	■ Confirm level of access to training data during vendor contract negotiations to be able to conduct bias audits (see SAFE AI Procurement Guide for the supplier evaluation questions to ask) ■ Review whether training data reflects the demographic range of your intended beneficiary population ■ Test outputs across gender, age, location, and language groups to identify disparate performance. ■ Where bias is identified, limit deployment scope or apply
Data privacy and security Example: A free AI productivity tool used by staff requires broad access to calendars and communications, creating unnecessary exposure of sensitive operational information.	■ Confirm that data used to train or run the system complies with your organisation's data protection policy and applicable law. ■ Avoid using or storing personally identifiable information (PII) unless strictly necessary. ■ Apply basic access controls to restrict who can interact with the system or view its outputs.
Legal and reputational harm Example: AI-assisted drafting of situation reports introduces incorrect figures leading to inaccuracies that undermine credibility with partners and communities. Example: AI-enabled translation of community-facing materials includes subtle mistranslations the cause confusion or unintended tone.	■ Require human review before publication. ■ Clearly label AI-generated text. ■ Strengthen due diligence before deployment ■ Complete the transparency process and fill out the Transparency Card ■ Complete the Technical Assurance process to identify liability issues.
Organisational independence Example: A deeply integrated AI tool that relies on a broader tech provider ecosystem to generate summaries of annual reports.	■ Leverage finished, off-the-shelf applications, also known as Software-as-a-Service or SaaS, without the need for customisation. This allows you to use technology options from different suppliers without risking lock-in.
Autonomy, including goal divergence, self-modifying systems, agentic AI Example: An agentic AI tool authorised to send routine stakeholder updates sends out an incomplete draft before a human has reviewed it, requiring urgent correction and eroding stakeholder confidence.	■ Ensure all AI outputs at Tier 1 are reviewed by a human before being acted upon. ■ Do not deploy agentic or autonomous AI without escalating to a Tier 2 or Tier 3 review. ■ Consider whether autonomous action, even in a low-stakes context, could undermine accountability to affected people if decisions or outputs cannot be explained or contested. ■ If used, confirm that the system's scope of action is clearly
Cognitive shortcuts, including automation bias, confirmation bias, decision dependency Example: Caseworkers rely on an AI triage score to prioritise protection cases	■ Raise awareness of common cognitive shortcuts ■ Request model to provide reasoning for recommendation.

Risk mitigation for Tier 2 use cases

Note: where called out with "Lower tier guidance applies", all the Tier 1 mitigation guidance for the given risk dimension also applies.

Risk dimension	Mitigation guidance
Protection and safeguarding Example: A case triage AI used internally by caseworkers ranks protection cases by estimated severity. Due to a data gap, cases involving women with mobility constraints are systematically under-scored, meaning they receive lower-priority follow-up despite acute need.	Lower tier guidance applies. ■ Engage affected communities in assessing and mitigating protection risks prior to an internal impact assessment and during co-design and deployment, building in mechanisms for ongoing community feedback. (see How to Co-design AI with Crisis-Affected Communities for stage-by-stage guidance) ■ Establish a clear reporting channel for safeguarding concerns arising from AI use. ■ Where the AI system makes inferences or decisions about people who do not interact with it directly, document who is affected, how they would learn that automation has shaped a decision about them, and what route they have to challenge it. The right-to-know norm reaches beyond the people who use the system to the people the system affects. ■ Assess conflict sensitivity: could AI outputs, the data used, or vendor relationships be exploited by hostile parties or create perceived alignment with a conflict actor? ■ Ensure affected communities have been informed of and consented to how AI systems will use their data, and that a mechanism exists for them to withdraw consent. (see SAFE AI Impact Assessment, Stage 2 data privacy questions).
Information reliability, including misinformation, synthetic media, narrative manipulation Example: An AI system used to analyse protection monitoring data generates a confident summary that misrepresents the severity of an emerging pattern, delaying a protection escalation response. Example: A chatbot used by the partnerships team to research potential donors gives a confident but inaccurate summary of a foundation's funding priorities. The team misses the opportunity to apply for funding directly applicable to their programme.	■ Human-on-the-loop design which allows the AI capability to act autonomously but leaves the option for human supervision and intervention.

Technical reliability, including technology effectiveness and limitations Example: A logistics optimisation tool used to plan field visit routes produces increasingly inefficient schedules after a map data update introduces errors in travel time estimates. Longer journeys cause a reduction in household visits, and the cause is only identified weeks later.	Lower tier guidance applies. ■ Constrained deployment, i.e. limit functionality of implementation or limit access to implementation. ■ Carry out Technical Assurance. ■ Make it clear to the user when they are interacting with AI content. ■ Where AI relies on proxy variables (e.g. mobile phone data, satellite imagery), validate that proxies remain valid as conditions change; sudden population movement or crisis escalation can rapidly invalidate proxy assumptions and produce systematically wrong outputs without visible system failure. ■ Ensure the AI Transparency Card is publicly accessible, written in plain language, and updated if material changes occur – failure to disclose AI use in rights-impacting contexts can cause avoidable harms and may itself constitute a legal and reputational risk.
Bias, representation, and exclusion, opacity in decision-making, accountability gaps Example: The model may be biased against certain demographics, due to historical bias in training data. Example: A vulnerability-scoring tool uses proxy indicators including employment status and reported mobility to estimate need: people with disabilities who are employed or under-report mobility barriers are scored as lower-need, because the model has no variable capturing disability-related costs or reliance on others for daily care, systematically underestimating their vulnerability	Lower tier guidance applies. ■ During vendor negotiations, request model documentation (e.g. model card or data sheet) from the vendor to assess training data composition and known limitations, and maintain ability to repeat requests as training data is updated ■ Where outputs influence decisions about affected people, document how those decisions will be explained to the people affected in plain language and in the operational languages of the populations involved. Explainability that exists only for technical staff does not meet the right-to-know standard at Tier 2. ■ Conduct structured bias auditing on outputs across demographic groups before and during deployment (see SAFE AI Risks and Mitigation Strategies). ■ If reliability concerns exist, restrict the AI tool itself rather than the programme: either switch off specific features until they are validated, or limit which teams use the tool while piloting, before expanding to wider programme operations.

Data privacy and security Example: Data and insights generated in humanitarian contexts may contribute to systems used far beyond those contexts, including in commercial and military applications. Value accrues upstream to those controlling models and infrastructure, while risks and accountability remain downstream with implementing organisations and affected populations.	Lower tier guidance applies. ■ Constrain or remove use of and access to PII data from affected communities. ■ Write protections into vendor contract to mitigate risk from company being sold, e.g. data will be deleted, ■ Build contractual provisions with your supplier 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. (see SAFE AI Procurement Guide, exit strategy from suppliers mid-deployment).
Legal and reputational harm Example: You have publicly announced a partnership with a vendor for a project to leverage AI for more effective disaster response. A new report in the news describes ethically dubious practices at your vendor. Example: AI chatbot answers general FAQs on a program website and provides incomplete or misleading answers, leading to perception of poor service quality.	■ Check for supplier's third-party certifications, audits, or adherence to standards. ■ Ask yourself whether you are comfortable aligning yourself with the supplier in a public forum or operational context. Ask the supplier to provide a written response regarding any concerns you have on a potential partnership. ■ Limit chatbot scope to non-sensitive topics. ■ Regularly review usage logs to correct recurring issues.
Organisational independence Example: A deeply integrated AI tool that allows the user to ask questions about your organisation's data which relies on a broader tech provider ecosystem, making it harder to leverage other competitors' capabilities. Example: A key AI vendor is acquired by a company with ties to a government involved in the conflict; the perceived alignment with a conflict party undermines community trust and the organisation's neutrality claim.	Lower tier guidance applies. ■ 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. (see SAFE AI Procurement Guide, supplier evaluation checklist) ■ Include open technical standards and APIs to support interoperability and portability, and avoid dependency on a single supplier. ■ 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.(see SAFE AI Procurement Guide and SAFE AI Architecture and Tech Stack Decision Guide for the architecture implications).

Autonomy, including goal divergence, self-modifying systems, agentic AI	Lower tier guidance applies. ■ Ensure that any agentic or semi-autonomous functions are clearly bounded and cannot take consequential actions without a human-on-the-loop review step. ■ Log all automated decisions and retain audit trails for post-hoc review. ■ Define and test override procedures to halt or roll back automated actions if unexpected behaviour is detected. ■ Encourage a culture of critical thinking and confidence in staff experience
Cognitive shortcuts, including automation bias, confirmation bias, decision dependency Example: Program officers use an AI forecasting tool that consistently predicts higher needs in areas already prioritized, reinforcing existing assumptions and reducing attention to emerging or under-reported areas.	Lower tier guidance applies. ■ Include diverse data sources and independent human reviews.

Risk mitigation for Tier 3 use cases

Note: where called out with "Lower tier guidance applies", all the Tier 1 and Tier 2 mitigation guidance for the given risk dimension also applies.

Risk dimension	Mitigation guidance
Protection and safeguarding Example: A community-facing eligibility system in a conflict zone is exploited by a hostile party to identify and locate individuals seeking protection, directly enabling harm. Example: An AI targeting system makes decisions about access to life-saving assistance based on criteria a beneficiary cannot understand, challenge, or appeal, violating their right to explanation and redress.	Lower tier guidance applies. ■ Engage affected communities in assessing and mitigating protection risks prior to an internal impact assessment and during co-design and deployment, building in mechanisms for ongoing community feedback. ■ Carry out a full protection risk assessment, including review of how AI outputs could affect the safety, dignity, or rights of individuals in the affected community. (coordinate with the SAFE AI Impact Assessment Stage 2 protection questions and your organisational protection assessment) ■ Where the AI system shapes access to assistance, protection, or information, ensure affected communities are notified that automation is in play, in formats and channels they actually use, before deployment. Tier 3 systems cannot operate behind a notification gap. Where the supplier or architecture pathway makes accessible notification impossible, apply responsible refusal. ■ Establish clear safeguarding escalation pathways that include AI-related incidents, ensuring communities can always reach a human for redress.(see SAFE AI Technical Assurance, incident surfacing and decommissioning triggers in the Stage 4 trustworthiness table)

Information reliability, including misinformation, synthetic media, narrative manipulation Example: An AI system providing psychosocial support generates inappropriate advice that a vulnerable individual acts upon, with serious personal consequences.	■ Strict human-in-the-loop design. AI output should only be a recommendation for a human expert to evaluate and decide on. ■ Affected communities must be able to tell when they are interacting with AI-generated content, including content delivered through chatbots, automated advice, or AI-mediated communications. Disclosure is a Tier 3 condition for deployment, not a refinement. ■ This also applies to the use of Agentic AI and autonomous AI systems, which for Tier 3 use cases should only be leveraged with the utmost caution, if at all.
Technical reliability, including technology effectiveness and limitations Example: Community-facing chatbot or automated advice. If output is assumed correct and acted upon as if correct, any incorrect or misleading advice may have real-world consequences on the individuals and communities using the service.	Lower tier guidance applies. ■ Pause or cancel deployment until technical deficiency has been addressed. ■ Carry out Technical Assurance. ■ For chatbot/automated advice, limit to low-risk topics, e.g. office hours. For more sensitive topics, provide static content verified by human expert. ■ For high-risk topics, e.g. mental health advice, leveraging generated content, recommend responsible refusal and provide non-AI alternative. ■ Make it clear to the user when they are interacting with AI content.
Bias, representation, and exclusion, opacity in decision-making, accountability gaps Example: If model output is assumed correct and acted upon as if correct, certain communities with dire and urgent needs may be prioritised lower than communities which have received before.	Lower tier guidance applies. ■ Complex mitigation path. It requires insight into training data, understanding of what and who is represented in the training data, and which specific data was selected vs. excluded for model training. Based on findings, model retraining and updated data set may be needed.(this is largely dependent on the architecture pathway chosen, see SAFE AI Architecture and Tech Stack Decision Guide. For commercial foundation model pathways, training data access is typically constrained and the mitigation may not be available).
Data privacy and security Example: AI tool processes beneficiary registration data including names, locations, and household details stored on an unsecured server, exposing individuals to targeting or surveillance.	Lower tier guidance applies. ■ Require secure data transfer, storage, and processing from supplier. ■ Ensure supplier compliance with data protection (e.g., GDPR) and intellectual property rights.
Legal and reputational harm Example: Community-facing chatbots or automated service providing mental health advice. Any incorrect or misleading advice may harm individuals, erode the trust in your organisation, and hinder your ability to effectively work with the communities you serve.	Lower tier guidance applies. ■ Make it clear to the user as soon as they are interacting with AI content.
Organisational independence Example: AI system integrated with a partner's digital ID registration platform leads to dependence on single partner's technology offering.	Lower tier guidance applies. ■ To reduce dependency on a technology supplier, invest in in-house capacity building that allows you to maintain and update the AI system independently.

Autonomy, including goal divergence, self-modifying systems, agentic AI Example: AI agent autonomously reallocates convoy routes to "maximize delivery efficiency", optimizing for speed and coverage, but unexpectedly routes convoys through insecure areas.	Lower tier guidance applies. ■ Restrict AI to advisory model for all sensitive and security-related decisions. ■ Require human expert sign-off on any critical decisions. ■ Implement hard constraints on which decisions the AI agent is allowed to take.(see SAFE AI Technical Assurance, agentic AI considerations)
Cognitive shortcuts, including automation bias, confirmation bias, decision dependency Example: An AI-enabled recommendation system for disaster response resource allocation increases dependency on field teams on AI decision-making and reduces ability to handle situation when system behaves unexpectedly.	Lower tier guidance applies. ■ Add some friction into the design of the AI system to make it harder to accept the AI-generated output without having considered its accuracy, a manual step that forces the human expert to actively review it before it can be used in downstream processes. ■ Require staff to record brief justification when overriding or accepting AI output. ■ Provide training on common cognitive shortcuts. ■ Display confidence level from the model (if available) along with AI output.

What to do with the work you have done

Risk mitigation work produces three outputs that need to land elsewhere in the SAFE AI Journey for the work to be operational.

- **Residual risks and chosen mitigations are recorded in the SAFE AI Transparency Card.** Section 2 (humanitarian context, risks, participation, and safeguards) and Section 5 (community engagement, governance, accountability, and redress) are the destinations. The Transparency Card is the auditable record that mitigation decisions were made deliberately and applied at the depth the tier required.

- **Mitigations that require contractual action are taken into the SAFE AI Procurement Guide.** Right-to-audit, supplier reporting obligations on model change, data exit clauses, and exit protocols are mitigations only if procurement secured them.

- **Mitigations that require ongoing monitoring are taken into the SAFE AI Technical Assurance work.** Bias drift, performance degradation, language adequacy, and incident response are tested at the Stage 3 pre-deployment assurance gate and again continuously through Stage 4.

Risk mitigation work is iterative across the SAFE AI Journey.

The mitigations identified at Stage 2 are tested against the design at the Stage 2 Decision Gate, against the built system at Stage 3, and against operational reality at Stage 4.

Where mitigation strategies prove inadequate at any point, the framework's defined response is to redesign, pause, or stop. A documented decision to apply responsible refusal is a SAFE AI outcome.

Where this guidance sits in the wider governance picture

Risk mitigation is one part of a wider governance practice. The guidance supports but does not replace organisational risk management, community engagement, and the inclusion of contextual expertise and voices from affected populations.

Phase 1 of SAFE AI puts the mitigation logic and the tier-proportionate approach in place. Phase 2 will establish the independent assurance function that verifies mitigation work has been done to the standard the framework requires. Until that function is operational, the responsibility for honest, documented mitigation sits with the deploying organisation, audited by the named accountable authority and recorded in the Transparency Card.

The SAFE AI Architecture and Tech Stack Decision Guide

This guide helps you decide what kind of AI system to build or buy, and how to structure it. It is used at Stage 2 of the SAFE AI Journey, after the Stage 1 Impact Assessment has confirmed that AI is the right tool for the problem you are trying to solve.

Work through the guide in four steps:

1. **Choose your architecture pathway**. The four pathways are locally built, frugal or small, open-weight, and commercial foundation model. Each has different implications for governance, cost, sovereignty, and what you can inspect.

2. **Work through the stack-layer decisions** that follow from the pathway choice. Infrastructure, data, model architecture, training, deployment, application, and oversight.

3. **Test your decision against the data dimensions**. How you handle training, validation, evaluation, and deployment data is shaped by the architecture you choose.

4. **Document the decision** in the SAFE AI Transparency Card, including what you decided, what you considered, and what you ruled out.

The depth of analysis expected scales with your SAFE AI tier. Tier 1 use cases can work through this guide at a high level. Tier 2 and Tier 3 use cases require structured documentation of trade-offs and rejected alternatives, and Tier 3 use cases require input from technical leadership with experience of comparable deployments.

Who should be in the room

This guide is applied by technical leads working with programme leads, procurement, data protection, and the named accountable authority. The architecture decision is too consequential to sit with technical functions alone. Organisations that lack technical leadership equipped for this decision should refer back to the SAFE AI Organisational Readiness Checklist. Where the technical capacity is not present, the responsible step is to address that gap before making architecture decisions, not to defer to a supplier's default.

What AI architecture decisions add to standard IT and procurement decisions

Standard IT and procurement decision-making already covers system requirements, supplier evaluation, cost estimation, integration with existing infrastructure, and security review. AI architecture decisions extend that work in seven specific ways:

- **Pathway sovereignty.** Whether the AI system you deploy is something your organisation can inspect, modify, and operate independently, or whether you depend on upstream actors you cannot see.

- **Training data provenance.** Where the data that built the model came from, under what consent, and whether the resulting model carries embedded biases that affect the populations you serve.

- **Evaluation evidence access.** Whether you can independently test the system's performance against your operational context, or whether you must rely on supplier representations.

- **Language coverage.** Whether the system works in the operational languages of the populations the deployment will affect, which is shaped by the pathway choice as much as by the supplier.
- **Interpretability and the right to know.** Whether the system's outputs can be explained to affected people in ways they understand, which different architectures make more or less possible.
- **Model change exposure.** Whether the system you deploy today will behave the same way tomorrow, after supplier updates, retraining, or capability additions outside your control.
- **Exit and portability.** Whether you can move away from the system at end-of-life or in response to harm, and what you keep when you do.

The framework body's Stage 2 design work treats these as governance questions, not technical refinements. The decisions you make at architecture stage flow into the SAFE AI Impact Assessment (Stage 2), the SAFE AI Procurement Guide, and the SAFE AI Technical Assurance.

Section 1: Choose your architecture pathway

Before working through stack-layer trade-offs, consider the broader architecture question: what kind of AI system is appropriate for this problem and this context.

The default assumption that "AI" means deploying or accessing a commercial foundation model is one option among several. SAFE AI does not prefer any single architecture, but expects organisations to make the choice deliberately and document it.

Four pathways are relevant in humanitarian contexts. Hybrid combinations are common.

1. Locally built systems

Smaller, purpose-built models trained or fine-tuned on data appropriate to the deployment context. Typically narrower in capability than foundation models, more interpretable, and better aligned with data sovereignty. Require technical capacity to build and maintain.

Often the strongest choice where the problem is well-defined, the data is governed, and the cost of error is high.

2. Frugal and small AI

Lightweight systems designed to run on limited compute, low-bandwidth environments, or on-device. Prioritise reliability under field conditions, lower energy use, and reduced supplier dependency.

Often appropriate where infrastructure is constrained, where connectivity is intermittent, and where long-term sustainability matters more than capability ceiling.

3. Open-weight and open-source models

Models whose weights are publicly available and can be deployed, modified, and audited by the deploying organisation. Reduce supplier dependency, support data sovereignty, and enable deployment in contexts where commercial terms or data residency requirements are incompatible with foundation-model APIs.

Place greater responsibility for governance, monitoring, and updates on the deploying organisation, because there is no supplier to require a model card from or to escalate harm to.

4. Commercial foundation models

Large general-purpose models accessed through commercial providers. Offer broad capability, fast deployment, and lower upfront technical cost. Carry the highest dependency on upstream design choices the deploying organisation cannot inspect or change, and the highest exposure to supplier terms, data handling practices, and pricing changes.

Often appropriate for low-stakes, internal, or productivity use cases. Often the wrong default for community-facing or high-stakes deployments.

Architecture pathway choice does not substitute for assurance

All four pathways remain subject to the full SAFE AI Journey, including community engagement, Impact Assessment, and the decision gates. The choice between pathways should be made before working through the stack-layer decisions below, documented in the SAFE AI Transparency Card, and revisited as the system moves through the deployment lifecycle.

Section 2: Work through the stack-layer decisions

Once the pathway is chosen, work through the decisions at each layer of the AI stack. The pathway shapes what is possible at each layer.

Use the table below to identify, at each layer, what option you have chosen, what you are optimising for, what risks the choice carries, and what governance controls apply.

For example, for the 'infrastructure' layer, you have the option to use a cloud deployment or an on-device deployment. The trade-offs to consider might be latency (likely longer latency with cloud, shorter with on device) and reliability (likely more reliable with cloud if connectivity is good).

Work through this with a technical lead who can advise on the trade-offs.

Stack layer	Options	Key trade-offs (what to optimise for)	Typical risks	Governance controls
System architecture pathway	■ Locally-built; frugal/small; open-weight ■ Commercial foundation model ■ Hybrid	■ Capability vs interpretability ■ Vendor independence vs deployment speed; sovereignty vs scale	■ Vendor lock-in, capability mismatch, technical capacity gap, sustainability after grant ends	■ Architecture decision documented in Transparency Card ■ Build-versus-buy review ■ Sovereignty assessment
Infrastructure	■ Cloud vs. on-device (edge) deployment, bandwidth, energy, sovereignty	■ Latency vs. reliability ■ Centralisation vs. local autonomy	■ Cybersecurity, energy use, vendor lock-in	■ Sustainability standards, data localisation policy
Data	■ Type (structured/ unstructured), sensitivity, access, consent	■ Availability vs. privacy ■ Global vs. local datasets	■ Bias, privacy breach, misinformation	■ Data ethics, consent protocols, GDPR, ICRC data norms

Model architecture	▪ Rule-based, ML, LLM, hybrid, agentic systems	▪ Accuracy vs. interpretability ▪ Autonomy vs. control	▪ Bias, hallucination, lack of transparency	▪ Model cards, explainability tools, red teaming
Training	▪ Dataset scope, optimisation, localisation	▪ Generalisability vs. specificity	▪ Bias amplification,	▪ Bias audits, dataset documentation
Deployment	▪ APIs, integration pipelines, monitoring	▪ Modularity vs. complexity	▪ Data drift, failure in scaling	▪ Continuous evaluation, drift detection
Application (humanitarian)	▪ Decision support, chatbots, early warning, field tools	▪ Automation vs. human oversight	▪ Misinterpretation, automation bias	▪ Human-in-the-loop, UX testing
Human and governance	▪ Oversight models, accountability, stakeholder inclusion	▪ Speed vs. deliberation	▪ Misuse, politicisation, erosion of trust	▪ Ethical/ oversight boards, transparency mandates

Cross-cutting risks at the stack level

Once you have worked through the layer-level decisions, review the cross-cutting risks below. These are risks that can manifest at multiple stack layers and that carry direct humanitarian consequences. Use the table to identify which apply to your use case and where in the stack they are likely to surface. This may affect your SAFE AI tier and should inform the next steps in your Impact Assessment.

Risk domain	Manifestation	Stack layers affected	Humanitarian impact
Technical	Model drift, overfitting, hallucination, lack of robustness	Deployment	Reliability of predictions in emergencies, misinformation, misclassification of needs
Operational	Integration failure, latency, scaling, overreliance on automation	Infrastructure, deployment, application	Interruption in field delivery, errors in aid targeting, decision delays
Ethical and social	Bias, discrimination, exclusion, privacy breaches	Data, model, governance	Harm to vulnerable populations, trust erosion
Political and security	Misuse, surveillance, data weaponisation, and exploitation	Infrastructure, governance	Threat to neutrality and protection, targeting of populations
Environmental	Compute intensity, e-waste	Infrastructure	[Un-]sustainable humanitarian digital operations
Economic	Supplier dependency, licensing cost	Deployment, governance	Access inequities, long-term viability, supplier lock-in

Section 3: Test your decision against the data dimensions

The data choices you make at architecture stage shape every downstream stage of the SAFE AI Journey.

The framework body treats data as one of the four architecture dimensions, alongside model type, pathway, and operating environment.

Work through the questions below before moving to procurement.

What kind of data does your AI system use?

An AI system uses data for different purposes at different stages.

Each carries distinct protection, consent, and bias considerations, and each requires a separate decision.

- Training data is what the model learns from.
- Validation data is used during development to tune the model.
- Evaluation data is what the model is tested against to confirm it performs as expected.
- Deployment data is the live data the model encounters in use.

Data dimensions and risks to consider

Use the table below to identify the data risks specific to your use case.

Data dimension	Examples	Risks to consider
What type of data do you have?	Geospatial, text, sensor, biometric, social media	Privacy, accuracy, bias
What is the source of your data?	Public datasets, partner data, user input	Licensing, consent, reliability
What is the quality of your data?	Completeness, timeliness, validation	Data drift, bias amplification
What data security measures do you have in place?	Encryption, storage, access controls	Leakage, manipulation, re-identification

How the architecture pathway shapes your data position

The architecture you choose constrains the evaluation evidence available to you. Closed proprietary systems usually permit only input-output sample testing. Open-weight or open-source systems permit deeper access. Where evaluation evidence is constrained by architecture, the constraint is itself a finding to surface in the SAFE AI Impact Assessment and to record in the SAFE AI Transparency Card.

Independent evaluation against a representative dataset is one of the strongest signals of trustworthy AI in humanitarian use. The SAFE AI Procurement Guide and SAFE AI Technical Assurance return to this point at the procurement and assurance stages.

Section 4: Document the decision

The architecture decision is too consequential to leave undocumented.

Before the use case moves into procurement, record the following in the SAFE AI Transparency Card (Section 1, system snapshot, and Section 3, data and model summary):

- The architecture pathway you chose, and the alternatives you considered and ruled out.
- The reasoning behind the pathway choice, in plain language a non-technical reader can follow.
- The stack-layer decisions at each layer, with the trade-offs you accepted.
- The data dimensions you have addressed, and the data risks you have identified.
- The evaluation evidence approach, including where access to evaluation data is constrained by the architecture and what that means for assurance.
- Any cross-cutting risks (technical, operational, ethical and social, political and security, environmental, economic) that are likely to be material for this use case.
- Any findings to surface at the Stage 2 Decision Gate, including unresolved trade-offs or constraints that the Impact Assessment will need to address.

A documented decision that was deliberately made and consciously trade-off tested is more defensible than an undocumented choice that worked out.

A documented decision that did not proceed, where the architecture analysis showed no viable pathway, is itself a SAFE AI outcome

Internal and external governance alignment and oversight by layer

Architecture decisions sit alongside the governance arrangements that hold the deployed system to account.

The table below sets out which governance mechanisms apply at which architecture layer.

Governance layer	Governance mechanisms	Humanitarian implementation
Multi-stakeholder oversight	Co-design with affected populations, collaborate with civil society and academics	Shared AI ethics/oversight dashboards/ committees, peer review
Infrastructure	Sustainable procurement, cloud ethics	Local hosting, energy audits
Data	Data responsibility frameworks, informed consent, datasheets	ICRC Data Protection Policy, GDPR alignment
Model	Model cards, system and model card, bias audits, reproducibility	Transparency to field users and affected populations
Operational	Red-teaming, field evaluation, monitoring	Field pilots, continuous evaluation
Organisational	Ethical review boards, accountability lines	Integration with 'HQ' and cluster/local governance

Where the SAFE AI Architecture and Tech Stack Decision Guide connects

- **Upstream**: The Stage 1 SAFE AI Impact Assessment, which should have confirmed that AI is the right tool for the problem before architecture work begins.
- **Downstream into procurement**: The SAFE AI Procurement Guide, which secures the contractual mechanisms (right-to-audit, exit rights, model change notification) that flow from the architecture choice.
- **Downstream into assurance**: The SAFE AI Technical Assurance, which tests the system's performance against the architecture and data decisions documented here.
- **Throughout**: The SAFE AI Transparency Card, which carries the architecture decision forward as part of the system's auditable governance record.

The SAFE AI Procurement Guide

For organisations with established procurement and contracting practice, the SAFE AI Procurement Guide is the AI-specific lens applied through existing procurement functions, not a replacement procurement framework.

Existing procurement practice continues to handle the standard contractual and commercial considerations.

The SAFE AI Procurement Guide adds the AI-specific contractual conditions that traditional procurement may not yet cover: right-to-audit, model change notification, data ownership and training data use, vendor incentives during pre-deployment assessment, exit rights including model and data portability, and the right to pause updates pending impact assessment.

Where procurement function is well-developed, these clauses are added to standard practice.

Where it is not, the SAFE AI Procurement Guide can serve as a reference for building procurement capacity proportionate to AI risk.

The questions in the tables below are meant as discussion prompts to help raise awareness of specific factors to consider as part of the AI procurement process. Document your answers in the tables below and have them handy for discussions with potential suppliers.

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.
- 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.

Where AI systems interact with other automated systems – including agentic AI tools that take actions on behalf of users – procurement and technical teams should identify and monitor data flows from both human and AI users within shared infrastructure.

As AI-within-AI deployment increases (for example, targeting or needs assessment systems that receive inputs from automated data pipelines), standard access controls designed for human users may be insufficient.

Require suppliers to document all automated access points and confirm that governance controls apply to machine-generated inputs as well as human ones.

Exit strategy from providers mid-deployment

Plan for provider exit as a distinct operational step, separate from system decommissioning.

If a provider relationship must be terminated mid-deployment – due to breach, acquisition, financial failure, or emerging ethical concerns – the organisation needs a defined protocol that protects affected populations throughout the transition. This protocol should specify: how affected people are notified; how service continuity is maintained or safely wound down; how data is retrieved and secured; and who has authority to initiate exit.

Document this protocol before deployment begins, not after a problem emerges.

Disclaimer

SAFE AI is not legal advice. It is designed to support governance, risk awareness, and informed decision-making. Organisations should seek appropriate legal advice where AI use may trigger regulatory obligations or legal risk.

Section 1: Supplier Evaluation Checklist

What is new for AI procurement: alongside standard supplier evaluation, AI procurement tests the supplier's AI maturity, their track record on humanitarian or comparable deployments, and their willingness to accept the AI-specific contractual conditions (right-to-audit, model change notification, exit and portability) that standard procurement contracts may not include.

Question / check	Yes / no / unclear	Notes / actions needed
Have you drafted tender specifications with outcome-focused criteria, avoiding being too prescriptive with technical details?		
Have you requested and reviewed AI model cards – your SAFE AI Transparency Card can help with this – for the specific model and relevant publications associated with model development and use?		
Are open technical standards and APIs to support interoperability and portability and avoid dependency on a single supplier included?		
Have you secured a right-to-audit clause in the contract, scaled to your SAFE AI tier? At Tier 1, this may be a short clause covering the principle of audit access. At Tier 2 and Tier 3, the clause should specify third-party access, audit cycle frequency, and provisions for sharing findings with sector-level assurance bodies. Audit rights should cover the duration of the contract and a defined period after termination.		

Have you assessed the supplier's AI maturity, expertise, and track record? Ask for specific examples of completed work.		
Have you applied Business and Human Rights (BHR) frameworks as go/no-go criteria in vendor selection?		
Have you checked for the supplier's third-party certifications, audits, or adherence to standards (e.g., ISO/IEC 42001, NIST AI Risk Management Framework)?		
Have you actively sought Global South and locally-based AI suppliers, including those rooted in the operating context, as part of the procurement process? Where Global South capacity does not yet exist for the specific AI capability needed, have you documented why, and considered whether the procurement decision strengthens or undermines local capacity over time?		
Some parts of an AI system you procure will be built by others further up the supply chain. Have you identified which parts you cannot inspect for yourself, and considered what this means for your organisation's accountability if the system causes harm?		
Have you assessed potential risks related to the partnership, including technical, ethical, legal, reputational, and financial ones? Ask yourself whether you are comfortable aligning yourself with the supplier in a public forum or operational context. Ask the supplier to provide a written response regarding any concerns you have about a potential partnership.		
Have you required the supplier to provide risk registers and mitigation strategies relevant to your use case and humanitarian context?		
Have you asked your supplier about the process for running SAFE AI Technical Assurance and which costs are associated with that?		

Note on liability

Where AI systems cause harm in humanitarian contexts, liability often lands with the deploying organisation rather than the upstream developer. Procurement is the moment at which this asymmetry can be addressed contractually.

Section 2: Data Governance, Security, and Privacy Procurement Checklist

What is new for AI procurement: alongside standard data protection requirements, AI procurement tests how training data and evaluation data have been sourced and documented, what access your organisation has to evaluation evidence to verify supplier performance claims, and how data provenance and re-identification risks are managed across the AI lifecycle. Getting clarity from your suppliers on how the AI service/ implementation they are providing accesses data is vital. Where the architecture pathway selected is open-weight or open-source (see Architecture Decision Guide), the governance obligations in this guide remain but the mechanism for fulfilling them changes.

There is no supplier to require a model card from, no contract through which to mandate monitoring or updates, and no vendor to notify if the model needs to be withdrawn.

In these cases:

- Document the model version and source at deployment as you would a supplier's model card.
- Assign internal responsibility for monitoring and updating.
- Establish a clear protocol for what happens if the model is found to be causing harm – including who makes that decision and what withdrawing the model means operationally when there is no vendor exit process to invoke.

Question / check	Yes / no / unclear	Notes / actions needed
Have you confirmed the supplier's policies regarding data retention and deletion, ensuring that these meet your own organisational policies, including humanitarian protection protocols?		
Have you requested details from supplier related to the data used for model training and validation, including documentation on data provenance and usage?		
Have you requested documentation of the evaluation dataset used to test the system, including its provenance, demographic representativeness relative to the population the system will be deployed against, and the basis on which the data was chosen?		
Have you secured access to the evaluation dataset, or to evaluation evidence sufficient for your team or an independent reviewer to verify the supplier's performance claims? Where access is not granted, has the supplier explained why, and have you recorded that constraint as a finding in the SAFE AI Transparency Card?		
Have you asked whether the supplier secured informed consent from the individuals who produced the data?		
Have you ensured the AI service/implementation complies with cybersecurity standards (e.g. the ETSI EN 304 223 standard)?		
Have you required secure data transfer, storage, and processing from the supplier?		
Are privacy-by-design and security-by-design clauses in contracts?		
Is the supplier compliant with data protection (e.g., GDPR) and intellectual property rights regulations in the jurisdictions in which you and they will be operating?		

Section 3: Ensuring Responsible AI Checklist

What is new for AI procurement: alongside standard quality and safety requirements, AI procurement tests bias mitigation, explainability of outputs, human-in-the-loop design at decision points that affect people, and the supplier's commitments to humanitarian principles in operational practice.

Question / check	Yes / no / unclear	Notes / actions needed
Have you asked your supplier what specific steps were taken to mitigate bias and ensure fairness in model output?		
Have you mandated implementation design and transparency that allows for human-in-the-loop where critical decisions are made in supplier's components?		
Have you required the supplier to disclose use of third-party models or training datasets?		
Have you asked the supplier about their confidence level in the AI capability working well for your target audience (e.g. robustness to local dialect)?		
Have you requested the supplier's safeguard details* and contingency plans for when AI fails or underperforms?		
Have you asked the supplier about their approach to maintaining adherence to humanitarian principles and commitments, particularly Do No Harm?		
Have you requested details from the supplier related to how AI outputs can be explained to non-technical stakeholders?		
Have you planned for responsible decommissioning of the AI system (e.g. documenting what worked and what did not), as well as intended use and limitations in case the system is resurrected later?		
Has the supplier confirmed in writing what they will provide to support community-facing functions, including language coverage for the operational languages of the affected population, accessible interfaces, and explanations that can be shared with affected communities in plain language?		
Where community feedback surfaces concerns about the system during deployment, what obligations has the supplier accepted to investigate, respond to, or act on that feedback? Has the contract secured your ability to pause or defer system updates until community-flagged impacts have been assessed?		
Where affected communities are the ones encountering the system or its outputs, has the supplier disclosed what they will and will not allow your organisation to share with those communities about how the system works, including model behaviour, training data sources, and known limitations?		

* Safeguard details (e.g., how safety and reliability have been engineered into the system) could include information like monitoring mechanisms that catch errors, thresholds that trigger alerts or fallback procedures, quality checks at different stages, bias mitigation steps, and audit trails that track how the system has performed.

Section 4: SAFE AI Metrics and Evaluation Checklist

What is new for AI procurement: alongside standard performance metrics, AI procurement tests for model drift, bias drift over time, feedback flow from affected populations, and the supplier's role in ongoing monitoring after deployment.

Question / check	Yes / no / unclear	Notes / actions needed
Have you defined key performance metrics for model/service accuracy and efficiency?		
Have you required regular reporting and audits during deployment from supplier, such as audits for bias, performance and accuracy, compliance with safety standards, security, and data protection audits?		
Have you established appropriate feedback mechanisms for the end user (i.e. the individuals using the AI implementation)?*		
Have you set up processes internally, or with your supplier, for ongoing monitoring of model performance with agreed triggers for when model retraining or model update is needed (see SAFE AI Technical Assurance)?		
Have you required documentation on model development, testing, and validation processes from supplier (see SAFE AI Transparency Card)?		
Have you confirmed that the evaluation data used to validate the system is distinct from the training data, representative of the deployment population, and accessible to your team or to an independent reviewer? (See SAFE AI Technical Assurance.)		
Have you tested the robustness and resilience of models under real-world conditions?		

* For example, between the technical team and other staff for back-office AI systems, and potentially between communities and programme staff for community-facing AI systems.

SAFE AI Technical Assurance

Technical Assurance under SAFE AI is the AI-specific governance assurance conducted at the Decision Gates by the Oversight body.

It draws on the existing assessment practices humanitarian organisations already conduct – security assessments, programme quality monitoring, evaluations, post-distribution monitoring – and adds the AI-specific work that those practices do not yet cover: assessing the system against the seven trustworthiness characteristics, monitoring model performance against deployment baseline, watching for bias drift, and assessing whether the system continues to meet the conditions for its tier classification.

Technical Assurance is integrated into existing assessment and monitoring schedules where they are operating, and added where they are not.

The work is the same governance function organisations already perform on other operational areas; SAFE AI provides the AI-specific lens to apply.

What SAFE AI Technical Assurance does

Following this process allows you to verify that your AI system performs as expected, and continues to do so after deployment.

It draws on the NIST AI Risk Management Framework's trustworthiness attributes and applies them to humanitarian contexts.

What this tool covers: the trustworthiness questions every AI system in humanitarian use should be able to answer, organised by lifecycle stage; how to calibrate those questions to your AI type; and the resourcing implications of doing this work seriously.

What this tool does not cover: a minimum test suite for each SAFE AI tier, named performance thresholds calibrated to humanitarian cost of error, or a library of sector-validated case studies.

These require sustained sector investment and form part of Phase 2's scope.

If you need external technical capacity to answer these questions and lack it, that is itself a finding to record in the SAFE AI Transparency Card.

How to implement SAFE AI Technical Assurance

Start by revisiting your AI Impact Assessment and the answers you provided there and review residual risk.

Describe in the table below how you addressed or mitigated the risks you had already identified for your use case.

Where you identified a risk which you deemed unnecessary to address, document your reasoning.

For each risk, record where it sits after mitigation. This is what allows you, your oversight body, and any external reviewer to see the residual risk position rather than a list of unaddressed risks.

Use a simple scale of Low, Medium, High for both residual likelihood and residual impact. The detail behind the rating belongs in the mitigation column.

Risks that remain High on either dimension after mitigation should be escalated through your AI Governance route before sign-off.

Begin your technical assurance work by returning to the residual risks from your SAFE AI Impact Assessment, and use the table below to record where each sits after mitigation.

Risk	Mitigated or accepted?	Describe the mitigation if mitigated, or the reasoning behind accepting the known risk, if accepted.

The Stage 3 section below sets out discussion prompts to raise awareness of factors to consider before deployment sign-off.

The Stage 4 section operationalises the same trustworthiness characteristics into thresholds, frequencies, and documented responses for ongoing assurance, with additional prompts to surface periodically.

Record findings from both in Section 3 of your SAFE AI Transparency Card.

Performance testing and conformity validation

Trustworthiness characteristics are inextricably linked to humanitarian principles, organisational culture and behaviour, the datasets used in AI systems, the choice of AI models and architectures, those informing decisions of the designed and developed system, and the interactions between users of the system.

Drawing from NIST RMF, there are seven common attributes of trustworthiness of an AI system: valid and reliable, safe, secure and resilient, explainable and interpretable, privacy enhanced, fair with harmful bias managed, and accountable and transparent.

Each of these seven trustworthiness attributes have several potential metrics that can be used and assessed.

The SAFE AI multi-disciplinary governance team should determine the specific metrics related to AI trustworthiness characteristics and the precise threshold values for those metrics that are most applicable to the specific use case.

Humanitarians and the deploying organisation can face challenging decisions in balancing these attributes and determining the associated metrics to employ.

One of the greatest challenges is that there are often trade-offs involved between some AI trustworthiness attributes and associated metrics, and the responsible team needs to make decisions about which attributes and/or metrics are a priority over another in the given crisis context.

For example, in certain scenarios, trade-offs between achieving fairness and optimising for accuracy or achieving privacy and optimising for interpretability.

Hence, determining AI trustworthiness attributes independently will not ensure the AI system's trustworthiness.

It requires a delicate consideration of attributes and priorities to meet diverse stakeholder needs for trust and AI system trustworthiness.

Operationalising this process happens with the Technical Assurance.

Technical Assurance is performed during two separate stages, for two related but separate purposes.

As illustrated in the SAFE AI Journey, the first time you perform technical assurance is at the end of Stage 3: Development when the full end-to-end AI system has been developed.

You complete technical assurance at this stage to confirm whether or not your AI system works as expected.

This is required before you can collectively sign off on the deployment to your target audience.

Y**ou then continually perform Technical Assurance again regularly** in Stage 4: Deployment and Monitoring & Evaluation after you have fully deployed the AI system and you have confirmed some real-world usage with your target audience.

The purpose of doing technical assurance at this post-deployment stage is to confirm system behaviour with your target audience and should be performed regularly, for example every six months depending on your specific use case or as conditions change.

Calibrating technical assurance to your AI type

The trustworthiness questions in the tables below apply to every AI system. The form and depth of the evidence required to answer them will differ by AI type.

Definitions of the main AI types in humanitarian use, indicative use cases, and example risks sit in the **SAFE AI Glossary** accompanying guidance (e.g. LLMs, prediction and classification models, or computer vision).

Read your AI type alongside the questions below. Where evidence sources, test methods, or thresholds are unclear for your type, that uncertainty is itself a finding to record in your SAFE AI Transparency Card and to surface for review at the relevant Decision Gate.

A note on evaluation datasets

Technical Assurance depends on having data to test against.

The dataset you use to evaluate an AI system is itself a governance object, with its own protection, consent, and bias risks.

The choice of evaluation dataset, its provenance, its representativeness of the deployment population, and your right of access to it are decisions that sit upstream in SAFE AI Architecture and Tech Stack Decision Guide and SAFE AI Procurement Guide.

Where those decisions were not made deliberately, or where the supplier controls evaluation evidence and you do not have meaningful access, that is a finding to record in the SAFE AI Transparency Card and to surface at the relevant Decision Gate before sign-off.

An evaluation dataset is the data your AI system is tested against to confirm it performs as expected. It must be different from the data used to train the system.

For prediction and classification models, this is usually a held-out portion separate from the training data. For LLMs, it may be a structured set of test prompts and expected behaviours.

Evaluation datasets typically come from one or more of:

- Vendor supply (treat as a starting point that requires triangulation).
- Adaptation of open benchmarks (few are humanitarian-relevant).
- Internal construction from operational data (requires the same protection and consent assessment as any other use of beneficiary data, see Procurement and Risk and mitigation).
- Collaborative construction with affected communities through the co-design process.

Three things to watch for:

- Evaluation data that does not represent the population the system will be deployed against.
- Evaluation data drawn from the same source as training data, which can mask poor generalisation.
- Evaluation data that does not include the protection-sensitive scenarios the system will encounter.

Stage 3: Pre-deployment technical assurance

Use the questions in the following table to evaluate your AI system behaviour as part of your Technical Assurance during Stage 3: Development.

What is new for AI: Alongside the standard pre-deployment quality testing your organisation already conducts, AI Technical Assurance tests whether the system meets the seven trustworthiness characteristics in your specific use case and context, with documented evidence that supplier performance claims hold under realistic test conditions.

NIST attribute	Trustworthiness principle	Question(s) to consider
Valid and reliable	Accurate	How will we assess the accuracy of what the model has learned... How will we communicate this as needed?
	Reproducible	How will we test whether desirable outputs can be reproduced in different circumstances?
	Efficient	How will we improve efficiency in energy, model size, and memory?
	Verifiable	How will we verify that the system is behaving as expected?
	Reliable	How will we ensure the system performs predictably and as intended?
	Valid	How will we validate the outputs including through external validation?
Safe	Safely interruptible	How will we ensure controls for deactivation and fail-safe shutdown?
	Containment	How can we contain the AI system to prevent safety and security breaches?
	Protection from proxy gaming	How will we test and prevent proxy gaming?
Fair with harmful bias managed	Mitigation of computational bias	How will we assess and mitigate computational bias?
	Non-discrimination	How will we ensure the system is not discriminatory?
Secure and resilient	Built-in defenses	How will the AI system respond to attacks as they occur?
	Robust	How will we protect against cyber and adversarial attacks?
	Resilient	How will we assess ability to handle uncertainty?

Explainable and interpretable	Interpretable Uncertainty	How will we make model uncertainty more interpretable?
Privacy-enhanced	Protection from unwarranted data access	How will we ensure the system cannot give unwarranted access?
Accountable and transparent	System honesty	How will we ensure outputs are accurate and not deceptive?
	Community validation	How have affected communities tested the system in operational conditions, and what did they find?
	Documented community influence	Where community testing surfaced concerns, how have those shaped the system before sign-off?

Stage 4: Ongoing technical assurance

Use the questions in the following table to evaluate your AI system behaviour as part of your Technical Assurance during Stage 4: Deployment and Monitoring &Evaluation.

What is new for AI: alongside the standard programme monitoring, post-distribution monitoring, and evaluation work your organisation already does during operations, AI Technical Assurance tests whether the AI system continues to meet the seven trustworthiness characteristics as conditions change, including model drift, supplier updates, as well as shifts in the deployment population.

Note: All seven trustworthiness characteristics continue to apply through deployment, at every SAFE AI tier.

What changes is the depth and frequency of evidence. Tier 1 use cases require lighter-touch monitoring.

Tier 2 use cases require structured re-evaluation at defined intervals with documented findings.

Tier 3 use cases require independent verification, more frequent re-evaluation, and external review where capacity permits.

Ongoing Stage 4 technical assurance is structured around seven domains, each corresponding to a core trustworthiness requirement.

For each domain, organisations should define:

- What they are assessing.
- How they are assessing it.
- The threshold or baseline against which assessment is measured.
- The documented response required if that threshold is breached.

These parameters should be set at the point of deployment sign-off and recorded in your SAFE AI Transparency Card.

They must be updated at every formal reassessment.

Trustworthiness Dimension	What to assess	How to assess it	Minimum frequency	Response if threshold
1. Valid and Reliable	Whether the system continues to produce valid, accurate outputs within agreed tolerance of the deployment baseline, across all population subgroups.	■ Automated metric tracking (accuracy, recall, precision, F1) against documented baseline ■ Disaggregated performance by subgroup. Qualitative output review. Outcome comparison	■ Continuous metric monitoring ■ Formal review quarterly (Tier 2/3) or biannually (Tier 1) ■ Immediate on trigger	■ Root-cause analysis; notify supplier ■ Interim human review of all outputs ■ Escalate to accountable authority ■ Suspend if unresolved within agreed period
2. Safety safeguards	Whether human oversight mechanisms, override rights, and harm-prevention safeguards are functioning in practice, not only in documentation. Whether human decision-makers retain meaningful, practical control over AI-influenced outcomes including the ability to override, pause, modify, or stop the system – and whether that control is exercised rather than delegated by default to AI outputs.	■ Structured staff survey on override mechanism usability and use. Case review of decisions where outputs were overridden or not followed. Incident log review. Active testing of interrupt mechanisms (agentic systems) ■ Review of override and escalation records: how often are outputs challenged, and with what result? ■ Staff interview or survey on perceived ability to override in practice ■ For agentic systems: action log review and autonomy boundary	■ Formal review quarterly (Tier 2/3) or biannually (Tier 1) ■ Active mechanism testing at each review cycle, or immediate on any evidence of scope creep or autonomy boundary breach	■ Immediate retraining and governance review if safeguards are found to be non-functional or unknown to staff ■ Suspend deployment if safeguards cannot be restored ■ Investigate why control is not being exercised. Address structural barriers (time pressure, authority gaps, training) ■ For agentic systems: immediately verify and restore autonomy boundaries before further operation
3. Secure and resilient	Whether data protection, access controls, audit trails, and supplier security obligations remain intact and unchanged since deployment.	■ Periodic internal security review against procurement contract ■ Supplier confirmation of no material change to data handling or access controls. ■ Audit trail completeness review. Penetration	■ Formal review biannually as minimum ■ Immediate on any security trigger or supplier infrastructure change	■ Notify data protection lead and legal team ■ Require supplier remediation within defined period. Incident response protocol applies immediately if a breach has occurred

4. Fair with harmful bias managed	Whether the system continues to produce equitable outputs across all population subgroups, and whether new bias patterns have emerged since deployment.	■ Disaggregated performance metrics by demographic subgroup ■ Community feedback disaggregated by group ■ Qualitative review by protection, gender, and safeguarding specialists ■ Independent bias audit (Tier 2/3)	■ Possible formal review quarterly (Tier 2/3) or biannually (Tier 1). Independent audit annually for Tier 2/3. Immediate on community feedback trigger ■ These timelines may differ for the type or architecture and contexts.	■ Document specific bias pattern; assess harm extent; notify affected communities ■ Pause outputs that cannot be verified as unbiased ■ Redesign or retrain before resuming
5. Privacy protections	Whether consent conditions, data minimisation, access restrictions, and community expectations around data use remain aligned with actual practice.	■ Review of data flows against consent records and DPIA commitments. Spot-check of data access against agreed scope ■ Community feedback review for data use concerns ■ Supplier confirmation of no change to retention, sharing, or processing	■ Formal review biannually as minimum ■ Immediate if consent conditions change or data incident reported	■ Notify data protection lead immediately. ■ Assess whether affected communities must be informed ■ Consider suspension pending review. Document response in your SAFE AI Transparency Card
6. Explainable and interpretable	Whether the system's outputs can be explained in plain, accessible language to non-technical staff and, where appropriate, to affected communities – and whether those explanations remain accurate as the system evolves.	■ Review community-facing explanations against current system behaviour ■ Test whether frontline staff can explain to communities what the system does, why it produces the outputs it does, and how to challenge them ■ Qualitative community feedback on comprehension	■ Formal review at each scheduled reassessment ■ Immediate reassessment if supplier updates the model or system configuration	■ Update explanations and communications ■ Retrain staff ■ If the system has become genuinely inexplicable following a supplier update, treat as a deployment pause trigger pending review

7. Accountable and transparent	Whether accountability structures – named individuals, documented responsibilities, active governance bodies, and functioning redress pathways – remain clear, current, and operational throughout the deployment lifecycle.	■ Confirm named accountable authority is still in post, aware of responsibilities, and actively engaged ■ Review whether community feedback and grievance mechanisms are operational and reaching decision-makers ■ Review SAFE AI Transparency Card currency against actual system state	■ Formal review at each scheduled reassessment ■ Immediate on any staff change affecting named accountable roles	■ Reassign and document ■ If feedback mechanisms are non-functional: restore before continuing deployment ■ If SAFE AI is materially out of date: update before next operational cycle

In addition to the operational monitoring set out above, Stage 4 Technical Assurance should surface the following discussion prompts for periodic review by the SAFE AI Project Team and oversight body, with findings recorded in your SAFE AI Transparency Card:

- **Usability:** How will we test usability and ensure users continue to understand outputs?
- **Accessibility:** How will we ensure accessibility for all intended users?
- **Shared benefit:** How are the benefits of the system distributed across affected populations?
- **Adversarial testing:** How will we enable red teaming and vulnerability discovery on a defined cycle?
- **Responsible disclosure:** How are system risks and findings safely released and shared with the sector?

Documenting monitoring outcomes

Every assessment against the seven domains must be documented in the SAFE AI Transparency Card, whether or not a threshold was breached.

- A record that shows consistently stable performance and functioning safeguards is evidence of responsible governance.
- A record that shows identified issues and documented responses demonstrates that monitoring is working.
- A record that shows no issues ever identified, across a complex humanitarian deployment at Tier 2 or Tier 3, should itself be treated as a governance signal requiring scrutiny.

Monitoring triggers – what should activate reassessment?

A monitoring trigger is an event or condition that should activate a structured reassessment of the AI system's performance, governance, or continued deployment.

Not all triggers require the same response, but all require a documented decision, and all must be capable of reaching the Senior responsible owner and SAFE AI Project Team.

The table below maps triggers across the three categories relevant to humanitarian AI deployment.

Technical triggers	Socio-technical triggers	Humanitarian triggers
Model drift detected: outputs shift compared to validated baseline without a corresponding change in inputs.	Community feedback indicates confusion, fear, mistrust, or perceived harm attributable to AI outputs.	Significant change in the size, composition, or vulnerability profile of the affected population.
Performance degradation: accuracy, recall, precision, or F1 score falls below the agreed deployment threshold.	Frontline staff report developing workarounds, discarding outputs, or losing confidence in the system.	New displacement event, conflict escalation, or natural disaster materially alters the operational context.
Unexpected or out-of-distribution outputs falling outside anticipated parameters for the validated use case.	Evidence that outputs are being applied to decisions beyond the system's agreed scope or intended use.	Change in the threat context that increases data sensitivity, personal safety risk, or surveillance potential.
Supplier update to the underlying model, training data, system prompt, fine-tuning approach, or architecture.	Changes in community composition alter who the system is serving and how it is experienced in practice.	New or changed legal, regulatory, or donor compliance requirements affecting the system's permissible use.
Infrastructure failure, data pipeline disruption, connectivity loss, or system access control breach.	Loss of community trust or breakdown of feedback channels connecting communities to system designers.	Shift in humanitarian principles alignment due to changed operational partnerships or mandates.
Adversarial manipulation: evidence that users are attempting to game, prompt-inject, or exploit system outputs.	Staff turnover removes people with system knowledge, oversight authority, or accountability responsibilities.	Partner organisation changes data sharing, procurement, or governance arrangements in ways that affect the system.
Data source change: new collection practices, loss of a data stream, data quality decline, or schema change.	Incident report received from any source relating to an AI-influenced decision or AI-caused harm.	Donor requirements change in ways that affect how the system's outputs may be used, shared, or reported.
Audit trail gaps or logging failures that prevent retrospective review of system behaviour.	Community co-design representatives or oversight bodies raise formal concerns about system behaviour.	Regulatory or sector body issues new guidance, standard, or investigation relevant to the deployed system.
Security vulnerability discovered in the model, its hosting environment, or its integration points.	Evidence that the system is reinforcing or amplifying existing power imbalances within the community.	Funding change or programme closure that affects the system's governance or support structure.

Not all triggers will be visible to technical teams. Socio-technical and humanitarian triggers are most often first identified by frontline staff, community engagement officers, or affected communities themselves, not by monitoring technical or quantitative dashboards.

Feedback channels that reliably connect those sources to the people who can act on them are as important as any technical monitoring tool.

A trigger that is identified but does not reach a decision-maker within a defined timeframe is a governance failure, regardless of how sophisticated the monitoring infrastructure is.

Architecture-specific trigger considerations

Some triggers are particular to the AI architecture in use.

The four categories below set out what the trigger table above does not capture by type. Reassessment should pay closer attention to the relevant row for the system you have deployed.

Definitions of each architecture sit in the SAFE AI Glossary.

Architecture	Triggers specific to this type
Generative AI (LLMs)	■ Hallucination patterns reappearing or worsening ■ Prompt injection attempts detected ■ Output drift after a supplier model update; scope creep into rights-affecting decisions the system was not validated for
Agentic AI	■ Autonomy boundary breach ■ Action chains exceeding authorised scope ■ Interrupt or kill-switch mechanism failure on test ■ Action log gaps that prevent retrospective review
Predictive and classification models	■ Model drift against documented baseline ■ Proxy variable invalidation through population movement, conflict escalation, or context change ■ Performance degradation by subgroup
Computer vision	■ Environmental or sensor input change ■ Training data distribution shift ■ Accuracy decline on under-represented groups ■ Misclassification patterns that emerge only at field scale

The incident response process

When an incident is identified, by any member of staff, by community feedback, by a technical alert, or by any other route, the following SAFE AI process applies.

The process is intentionally simple: the governance value of an incident reporting system comes from its consistent use, not its complexity.

Stage	Action required	Who is responsible
1. Identify and log	As soon as an incident is identified, log it. The incident log should capture: ■ Date and time ■ Source of identification (staff, community, technical alert) ■ Description of what occurred ■ The system, version, and configuration involved ■ The population or individuals affected ■ An initial severity assessment Do not wait for certainty before logging. A suspected incident that is not logged cannot be investigated. Log it as a suspected incident and update as investigation proceeds.	Any staff member who identifies the incident. The incident log must be accessible to all staff involved in system operation.
2. Notify the accountable authority	Notify the named accountable post-deployment authority immediately – within the timeframe defined in the incident response protocol documented at deployment sign-off. For Tier 2 and Tier 3 systems, immediate notification means within 24 hours. For Tier 1 systems, within 48 hours. The accountable authority is responsible for all subsequent decisions in the response process.	The staff member who identified the incident, or their line manager.
3. Assess severity and scope	The accountable authority (Senior Responsible Owner or other), with input from technical leads, programme leads, and protection or safeguarding specialists, assesses the severity of the incident and its scope how many people are affected, what harm has occurred or may occur, and whether the incident is contained or ongoing.	Senior Responsible Owner, with technical and programme input.
4. Protect affected people	Where individuals or communities have been or may be harmed, take immediate action to stop further harm before investigation continues. This may include suspending the system, reversing decisions where possible, and providing alternative access to assistance or information for those affected. Protection of affected people takes priority over investigation, supplier notification, or documentation.	Senior responsible owner and programme leads.

5. Investigate	Conduct a structured investigation of the incident: what happened, why it happened, whether it is a symptom of a broader problem, and what the system's contribution to the harm was relative to human decisions. For Tier 2 and Tier 3 incidents assessed as Medium or above, involve an independent reviewer. Document the full investigation record.	Technical leads and programme leads, with independent input for Medium+ incidents.
6. Notify communities and stakeholders	Where the incident has affected communities, notify them through existing channels, in accessible language, and with clarity about what happened, what has been done, and what protection is now in place. Notification is a governance requirement, not a discretionary communication decision. Also notify relevant internal governance bodies, and where required by regulation or donor conditions, external bodies.	Senior Responsible Owner and community engagement leads.
7. Document the incident report	Complete the SAFE AI Incident Report (see below) and update the SAFE AI Transparency Card to reflect the incident, the investigation findings, and the response. The incident report becomes part of the system's permanent governance record.	Senior Responsible Owner with input from technical and programme leads.
8. Determine response and outcome	On the basis of the investigation, determine whether the system should: ■ Resume operation without change ■ Resume with operational modifications ■ Be paused pending supplier remediation ■ Or be decommissioned Document the decision and rationale. Confirm the conditions and timeline for any resumed operation.	Named accountable authority, with input from Oversight body where applicable.

Resourcing your SAFE AI technical assurance

Technical Assurance has a cost. For Tier 2 and Tier 3 deployments, it is not optional. Cost it into the AI use case from the outset, alongside development, deployment, and decommissioning.

Independent estimates from the AI assurance industry place the cost of a single thorough assurance engagement at £100,000 to £150,000 over six weeks for one agency (estimate as of May 2025, and not yet updated for 2026). Most humanitarian organisations cannot absorb this individually.

SAFE AI Phase 2 will explore cost sharing avenues. Until that infrastructure is in place, your options are:

- Build internal Technical Assurance capacity, where the AI use case is part of a sustained programme and the cost amortises over time.

- Procure assurance services from a qualified third party, where the use case is high-stakes and internal capacity is absent. This is the route for Tier 3 systems where independent verification is required.

- Combine a vendor's evaluation evidence with internal sample testing and external peer review by another humanitarian agency or an academic partner, where budget is genuinely constrained. This is the realistic baseline for Tier 2 systems in many organisations.

Whichever route you take, record the resourcing decision and its rationale in the SAFE AI Transparency Card. A documented cost-constrained assurance approach is more defensible than an undocumented one.

The skills needed to lead this work include statistical literacy, AI and machine learning evaluation experience, humanitarian context and protection expertise, and red teaming for Tier 3 systems. Few humanitarian organisations carry all four. Document the gap and the mitigation.

For further information on SAFE AI, please visit **www.cdacnetwork.org/safe-ai**.

The SAFE AI Transparency Card

SAFE AI Transparency Card Guidance Note

What is this?

The SAFE AI Transparency Card is the central authoritative record of your SAFE AI journey. It captures key decisions, risks, safeguards, and evidence at each stage of the process, providing a structured way to document how and why an AI system is designed, developed, and used. The transparency card enables clear governance in resource-constrained settings by capturing who made decisions, why, and on what basis. This is updated based on triggers and is a live audit or provenance tracking of decisions and monitoring of the AI system over time. This reduces reliance on institutional memory and specialist capacity. It supports continuity despite staff turnover, allows rapid oversight by leadership and donors, and ensures decisions can be reviewed and challenged.

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.

Who should be involved?

The accountable programme lead, with input from technical teams, risk and safeguarding specialists, and others involved at each stage of the process. Contributions may come from multiple teams, but a single responsible person should be clearly assigned for maintaining the card.

How and when should I read this?

The Transparency Card is built throughout the SAFE AI journey, starting in Stage 1 and updated at each step as decisions are made, risks are identified, and evidence is gathered. You revisit it as new information arises, or your situation or needs change. It draws on inputs from the SAFE AI Impact Assessment, and every step along your SAFE AI journey will provide you with insights to record within it. It should especially be kept up to date during Monitoring and evaluation. The completed card is the authoritative record of decisions, risks, and accountability. This is the minimum governance requirement for any AI system. In addition, it can be used to support and demonstrate internal governance, external communication, and accountability to affected communities and other stakeholders.

How to use the SAFE AI Transparency Card

Producing a transparency card for an AI implementation – a short, standardised documentation document that accompanies an AI model, outlining its intended use, performance metrics, limitations, and training data – is considered a responsible, transparent, and accountable practice in machine learning and AI (CHAI and OECD).

This is particularly important for AI systems in humanitarian action, as they involve

complex ecosystems of data, models, suppliers, decision-makers, users, and affected communities. The AI Transparency Card builds on this concept of the model card to provide a transparent, auditable record of responsibility, risk awareness, and humanitarian alignment.

The AI Transparency Card operationalises the right to know as a governance requirement. The card is the documentation that allows affected people, donors, and partners to know what an AI system does, who is responsible, and how to challenge it. In humanitarian contexts, this right applies not only to individuals who interact with an AI system directly, but to anyone whose access to assistance, prioritisation, or protection is influenced by it – including people who have no knowledge that automation is in play. Clear disclosure of system function, decision influence, and redress pathways is the minimum standard consistent with accountability principles.

It extends the standard transparency reporting mechanisms of the AI community by embedding system-level, institutional, humanitarian, and ethical dimensions, allowing organisations to demonstrate the actions they have taken towards responsible design, deployment, and oversight of AI systems used in humanitarian contexts, and integrating community participation, which is relevant to good humanitarian practice.

For organisations that already maintain internal AI governance documentation, the SAFE AI Transparency Card is the additional sector-facing record produced through existing governance processes, not a substitute for them. Internal governance documentation continues to serve internal purposes. The Transparency Card serves the sector-level comparability and accountability that no individual organisation's internal framework can deliver alone.

Which version of the SAFE AI Transparency Card you should use

We have developed two versions of the AI Transparency Card: **lean** and **full.** Which one you should use depends on the maturity and risk profile of your AI system. See the table below to understand which card you should use to document your use case.

Version	Purpose	When to use It	Depth and evidence required
Lean	Streamlined record for exploratory, internal, or low-risk AI applications.	For tier 1 AI systems: ■ Pilots, research, or prototypes ■ Advisory or analytic tools not affecting direct assistance ■ Early design stages before roll out ■ Donor review	Summarised text or checkboxes, referencing external documentation where relevant.
Full	Full lifecycle documentation and governance record for complex or sensitive systems and full deployment of 'low risk' systems.	For tier 2-3 AI systems: ■ High-risk or sensitive use cases (e.g. protection data, crisis settings, autonomous and agentic AI systems) ■ When AI informs decisions affecting people's entitlements or safety ■ Multi-partner, externally procured, or donor-funded AI systems	Detailed sections with supporting evidence (DPIA, AIA, model documentation, governance frameworks).

How to fill out the SAFE AI Transparency Card

The following table summarises key components of the SAFE AI Transparency Card and shows who is responsible for filling in each component of the SAFE AI Transparency Card, the purpose, and intended output.

Both the lean and the full version of the SAFE AI Transparency Card have all these sections, what will differ is the level of detail for each section.

The SAFE AI Transparency Card is a living governance record. It is updated throughout the lifecycle of the AI system and captures key decisions, risks, and accountabilities at each stage, providing an auditable record of the life (and afterlife) of the AI system.

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 expertise	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.

Who fills out the SAFE AI Transparency Card?

Role / team	Primary responsibilities in the AI Transparency Card process
Technical team	Populate Sections 1, 3, 4 (data, model, infrastructure, and lifecycle monitoring) and referencing; support other sections as needed in Sections 2 and 5 regarding humanitarian principles checking, risk mitigation and governance.
Programme/humanitarian team	Lead Section 2 (context and alignment), and 5 (community and governance), giving input to all other sections.
Legal/ethical team	Oversee Sections 1 (snapshot), and lead and give strong input to Sections 4, 5 and 6 (AI stack, governance, referencing).
Procurement/partnership lead	Document vendor dependencies (Section 1, 4 and 5). Contribute to any references and documentation in Section 6.
Oversight board	Review completeness, verify residual risk, and endorse final AI Transparency Card before deployment or scaling with particular responsibility of Sections 1 and 6, remaining up to date.

When to transition between versions

The lean card acts as an early-stage ethical and technical screening tool. Information from the lean version feeds into the full card if the system progresses to operational use, scale-up, or external engagement. Both forms ensure continuity of oversight throughout the AI lifecycle.

Lean → Full: When moving from pilot to operational stage, engaging external partners, or impacting affected populations.

Full → Archived: When the AI system is decommissioned or replaced, for record-keeping and learning.

See the table below for examples of triggers that would encourage moving from a lean card to a full card.

SAFE AI journey stage	Lean card use	Trigger for full card
Stage 1 and 2: Use case / design	Document intent, data, expected benefits, and early risks.	Funding approval, data sharing, or prototype deployment.
Stage 3: (step 3.3.) Development / testing	Record deployment context, feedback mechanisms, and local engagement.	When results inform programmatic decisions or affect people's entitlements.
Stage 4: Deployment / ongoing monitoring and evaluation	Lean card archived or appended.	Full card completed for full transparency, due diligence, and donor accountability.

How to build trust with your SAFE AI Transparency Card

Organisations should publish the "snapshot" which provides a high-level summary of their SAFE AI Transparency Card publicly to demonstrate due diligence and build trust

with partners, donors, and affected communities. Publication should happen before deployment, not after.

A SAFE AI Transparency Card published once a system is already operating does not give affected people meaningful opportunity to raise concerns before decisions about them are made. In both the lead and the full versions of the card that follows, you will see a 'publish' tag on the sections you should make public. We recommend publishing on your website.

Publication serves different purposes for different audiences and the format should reflect this.

- For donors and implementing partners, the full published card demonstrates governance due diligence.

- For affected communities, the relevant sections what the system does, how decisions can be challenged, who is accountable should be communicated in accessible formats, in local languages, and through channels communities already use, not only through organisational websites.

Where a system is updated or significantly changed, the Card should be updated and the change communicated to communities through the same channels used at deployment. The sector-wide repository named here is a Phase 2 commitment and in the interim, organisations are encouraged to share their published Cards with SAFE AI at CDAC Network for aggregation and learning.

SAFE AI Phase 2 will include the development of a sector-wide repository of SAFE AI Transparency Cards to facilitate knowledge sharing, learning, facilitate collaboration and minimise resource wastage.

SAFE AI Transparency Card: Lean version

What is this?

This is the Lean SAFE AI Transparency Card. It is the minimum governance record for Tier 1 AI use cases: internal, low-risk, and not affecting access to assistance, protection, or information. It should take less than an hour to complete in full for a typical Tier 1 use case, drawing on the work already done in your SAFE AI Use Case Assessment, Impact Assessment, and Onboarding Readiness Checklist.

What you'll find here

Six short sections covering system snapshot, humanitarian principles red lines, data and model summary, deployment and monitoring, community in the loop and governance, and references and versioning. Each row tells you where in the SAFE AI journey the content comes from. Where any row triggers escalation, the use case is no longer Tier 1 and should move to the Full Card.

For detailed guidance on completing each section, see the SAFE AI Transparency Card Guidance Note. You will generate the information required during your SAFE AI journey.

Make this section publicly available on deployment

Lean Section 1: System snapshot

To be completed by: Programme Lead with input from Technical Lead.

Purpose: State intent, scope, and humanitarian alignment.

Field	Entry	Guidance	SAFE AI journey
System basics	Name: Version: Date: Developer: Model type:	State model name, version, date, developer, and a one-line description of model type (analytical, predictive, decision-support, generative, etc).	▪ Procurement ▪ Use Case Assessment
Purpose		What humanitarian problem does the system address? Why is AI the right tool, and why is it better than the current approach? One paragraph.	▪ Use Case Assessment

Deployment context	**Where used:** **Who could be affected:** **Community-facing or affecting access to assistance, protection, information?** ☐ Yes ☐ No	State deployment context, who could be affected, and whether the system is community-facing or shapes access to assistance, protection, or information. **If Yes, this is not a Tier 1 use case. Escalate to Full Card and reassess tier before proceeding.**	■ Use Case Assessment ■ Impact Assessment
Architecture and stack	**Architecture pathway:** ☐ Locally-built ☐ Frugal / small AI ☐ Open-weight ☐ Commerical foundation model ☐ Hybrid **Supplier(s):** Key dependencies:	State the architecture pathway, supplier(s), and key dependencies (data, cloud, connectivity, other platforms). Note right-to-audit status. Where the pathway is hybrid, describe in one line.	■ AI Architecture and Tech Stack Decision Guide ■ Procurement ■ Onboarding Readiness Checklist
Safeguards and known limitations	**Humanitarian safeguards:** Technical safeguards: Legal and ethical safeguards: Known limitations:	State one line per safeguard category. List known limitations of the system, including accuracy, bias, and any conditions under which outputs should not be relied on.	■ Impact Assessment ■ Risk Mitigation ■ Technical Assurance
Sign-off	**Senior Responsible Owner:** Role: Date: Version:	SRO confirms the snapshot reflects the system as deployed and approves proceeding to monitoring. Update on any material change.	

Lean Section 2: Humanitarian principles red lines screen

To be completed by: Programme Lead with ethics or legal review.

Purpose: Screen the use case against the SAFE AI red lines.

For Tier 1 use cases, any Y or Potentially answer means the use case is not Tier 1 and should escalate to the Full Card, where mitigation and residual risk work is documented.

Transfer answers from your SAFE AI Impact Assessment.

For guidance on how humanitarian principles are interpreted in SAFE AI, see the SAFE AI Framework section on humanitarian principles and AI.

Principle	Red line	Answer	Escalation
Impartiality	**Systemic exclusion:** Could this system systematically exclude or disadvantage specific groups from assistance, protection, or information in ways that cannot be detected and corrected?	☐ Yes ☐ No ☐ Potentially	If Yes or Potentially: escalate to Full Card.
Impartiality	**No meaningful redress:** Are there insufficient or inaccessible mechanisms for affected people to challenge, correct, or appeal AI-influenced decisions?	☐ Yes ☐ No ☐ Potentially	If Yes or Potentially: escalate to Full Card.
Independence	**Loss of control:** Would reliance on suppliers, data providers, or infrastructure undermine organisational independence or create unacceptable conflict-related exposure?	☐ Yes ☐ No ☐ Potentially	If Yes or Potentially: escalate to Full Card.
Humanity	**Harmful outputs at scale:** Could the system plausibly generate or amplify dangerous misinformation or harmful guidance at scale without reliable safeguards?	☐ Yes ☐ No ☐ Potentially	If Yes or Potentially: escalate to Full Card.
Humanity	**Protection and safeguarding risk:** Could the system generate or materially increase protection or safeguarding risks to affected people or staff without effective mitigation or redress mechanisms?	☐ Yes ☐ No ☐ Potentially	If Yes or Potentially: escalate to Full Card.
Neutrality	**Conflict exposure:** Could data sources, vendors, hosting, or deployment create real or perceived alignment with a party to conflict, or enable surveillance or coercion?	☐ Yes ☐ No ☐ Potentially	If Yes or Potentially: escalate to Full Card.

Decision record	Entry
Any escalation triggered above?	☐ Yes ☐ No If Yes, stop the Lean Card here and move to Full Card before proceeding.
Decision authority (role, name, date)	

Lean Section 3: Data and model summary

To be completed by: Technical Lead.

Purpose: Provide minimum transparency on data sources and model choices for a Tier 1 use case.

Field	Entry	SAFE AI journey source
Data source	Origin and brief description of the data the system uses. Include source, type, and approximate scale.	Procurement; Use Case Assessment
Personal data involved?	☐ Yes ☐ No If Yes, include reference to your DPIA. **If Yes: this may not be a Tier 1 use case. Reassess tier before proceeding.**	Impact Assessment; existing DPIA process
Model adaptation	State any fine-tuning, grounding, prompt engineering, or other adaptation applied. If none, state "off the shelf".	AI Architecture and Tech Stack Decision Guide; Procurement
Known limitations	Plain language. Include accuracy concerns, known bias patterns, conditions under which outputs should not be relied on, and any caveats from the supplier.	Procurement; Technical Assurance
Sign-off	**Senior Responsible Owner:** **Date:** SRO confirms data and model choices meet minimum requirements for a Tier 1 use case, including any accepted limitations.	

Lean Section 4: Deployment, monitoring and lifecycle

To be completed by: Programme Lead with technical input.

Purpose: Confirm operational readiness, set a light-touch monitoring approach, and document the decommissioning plan.

Field	Entry	SAFE AI journey source
Operational readiness	**Skills and capacity in place:** ☐ Yes ☐ No **Infrastructure readiness:** ☐ Yes ☐ No **Budget readiness:** ☐ Yes ☐ No	Onboarding Readiness Checklist

Monitoring	Who reviews the system, on what cycle, and what would trigger a reassessment? Light-touch monitoring is acceptable for Tier 1. State the trigger conditions explicitly: model update, supplier change, scope creep, or any incident.	Technical Assurance (Stage 4)
Decommissioning	When and how will the system be stopped or replaced? Who is responsible for that decision? What happens to any data the system holds?	Impact Assessment
Sign-off	**Named owner for monitoring and review:** **Role:** **Date:**	

Lean Section 5: Community in the loop and governance

To be completed by: Programme Lead with legal or ethical review.

Purpose: Confirm accountability and feedback routes. Tier 1 use cases are by definition not community-facing and do not affect access to assistance, protection, or information.

Field	Entry	SAFE AI journey source
Community-facing check	Does this system interact with crisis-affected communities, or shape decisions affecting their access to assistance, protection, or information? Y / N If Y: this is not a Tier 1 use case. Escalate to Full Card and reassess tier before proceeding.	Impact Assessment; How to Co-design with Affected Communities
Accountable lead	Named individual: Role: Contact:	Onboarding Readiness Checklist
Feedback and redress route	How do staff using the system raise concerns? Where does that feedback go, and who is responsible for acting on it? Even in internal use cases, there should be a named route.	Risk Mitigation; Technical Assurance

Lean Section 6: References and versioning

To be completed by: Technical Lead with Programme and oversight input.

Purpose: Provide audit trail, version management, and continuity over time.

Field	Entry
Linked documents	List all references in one place: SAFE AI Use Case Assessment, SAFE AI Impact Assessment, any DPIA, supplier documentation, model cards or datasheets, relevant policies. Free text or bullet list.

Sign-off	This Card is the minimum governance record for this AI use and should be updated if the use changes. Senior Responsible Owner confirms record is complete and current.

Lean version control table

Update on each material change: model update, supplier change, scope change, identified incident, or scheduled review.

Version	Date (DD/MM/YY)	Approver (name and contact)	Reason

REMEMBER!

If it's not captured in the Transparency Card, it is not considered governed by SAFE AI.

SAFE AI Transparency Card: Full version

What is this?

The Full SAFE AI Transparency Card records key information about your AI implementation in a humanitarian context, enabling transparency, accountability, and trust. This is the comprehensive version for more mature or AI systems (Tier 2 and Tier 3), and for any Tier 1 use case where the Lean Card surfaced an escalation.

What you'll find here

For detailed guidance on completing each section, see the SAFE AI Transparency Card Guidance Note. You will generate the information required during your SAFE AI journey.

Full Section 1: SAFE AI system snapshot, overview and intent

Make this section publicly available on deployment

To be completed by: Programme Lead with input from Technical Lead.

Purpose: State intent, scope, and humanitarian objectives, with humanitarian principles alignment recorded as the foundation for every entry in this section.

Field	Entry	Guidance	SAFE AI journey
Model details	Name: Version: Date: Developer:	State model name, version, date, and developer.	▪ Procurement
Model type and architecture	Type: Architecture:	Briefly describe the type of model (analytical, predictive, decision-support, generative, etc.) and architecture (classical machine learning, reinforcement learning, generative adversarial networks).	▪ Procurement; AI Architecture and Tech Stack Decision Guide

Purpose		What humanitarian problem does it address? Why do you believe that an AI implementation will provide value over current approaches? State the alignment with humanitarian principles informing this judgement.	■ Use Case Assessment ■ Impact Assessment
Intended use		Short description of how the AI supports humanitarian operations. State the boundaries of intended use, including what the system will not be used for.	■ Use Case Assessment
Deployment context	**Geographic focus:** **Affected population:** **Humanitarian crisis:**	State the deployment context of the humanitarian crisis including geographic region(s) and affected population. Add any available Sex, Age and Disability Disaggregated Data drawn from existing context analyses, gender analyses, and conflict analyses.	■ Use Case Assessment ■ Impact Assessment
System values, performance and limitations	**Key values:** **Key performance metrics:** **Key limitations:**	State the key values and associated metrics (e.g. accuracy, privacy, fairness). State key limitations of the system. Values should map to humanitarian principles where applicable.	■ Use Case Assessment ■ Co-design ■ Procurement ■ Technical Assurance ■ Impact Assessment
System and infrastructure	**Scope:** **Infrastructure environment:**	State the scope of deployment (e.g., internal/ back-end, public-facing, affected community-facing, partner(s) facing, multiple) and infrastructure environment (on device, cloud, hybrid, etc.).	■ Procurement ■ AI Architecture and Tech Stack Decision Guide

System control and stack management	**System control:** **Architecture pathway (select one or describe hybrid):** ☐ Locally-built ☐ Frugal / small AI ☐ Open-weight ☐ Commerical foundation model ☐ Hybrid (describe) **In-house:** ☐ Yes ☐ No **Vendor:** ☐ Yes ☐ No **Open Source:** ☐ Yes ☐ No **Stack dependencies:**	State the architecture pathway selected and why, the build arrangement, the supplier(s), and stack dependencies (data, cloud, connectivity, other platforms), including which suppliers have access to which data. Document the trade-offs accepted in pathway choice. Independence as a humanitarian principle should be tested explicitly: external control over critical components is a principles question, not only a technical one.	■ AI Architecture and Tech Stack Decision Guide; Procurement ■ Onboarding Readiness Checklist
Right to audit	**Right to audit secured in contract:** ☐ Yes ☐ No **Duration:** **Third-party access permitted:** ☐ Yes ☐ No **Findings shareable with sector-level assurance bodies:** ☐ Yes ☐ No **If N to right-to-audit: record reason and escalate to oversight body before proceeding.**	Where the pathway involves a vendor or external supplier, record whether the contract secures right-to-audit, including duration, third-party access, and whether audit findings may be shared with sector-level assurance bodies. Tier 3 systems require third-party access and findings shareable with sector-level bodies.	■ Procurement
Ecosystem	**Other humanitarian organisations involved:** **Other stakeholders involved:** **Delivery partners:**	State participating stakeholders and partners including vendor(s).	■ Onboarding Readiness Checklist ■ Procurement

Governance and community record	**Humanitarian and community:** Technical: Legal and ethical:	State agreed safeguard mechanisms to ensure system values for privacy, misuse prevention and redress. Safeguards should be traceable to humanitarian principles where applicable.	■ Impact Assessment ■ Risk Mitigation ■ Co-design; Technical Assurance
Decision record		SRO confirms purpose, scope, and boundaries of use, including what the system will not be used for, and approval to proceed to design stage.	

Full Section 2: Humanitarian principles alignment and context

To be completed by: Programme Lead with ethics and legal review.

Purpose: Record how the four humanitarian principles have been applied in the design of this AI system, including the red line risks tested in the SAFE AI Impact Assessment. Transfer your relevant answers from the AI Impact Assessment into the table below. A Yes or Potentially answer to any red line is a stop condition unless mitigations bring the residual risk to a level the Decision Gate can accept. For guidance on how humanitarian principles are interpreted in SAFE AI, see the SAFE AI Framework section on humanitarian principles and AI.

Difference from Lean: Record key community engagement concerns, decisions and design objectives. Apply full assessment with mitigation and residual risk columns. Include broader principles questions beyond the red lines.

2.1 Context and community engagement

Field	Entry	SAFE AI journey source
Purpose	State the humanitarian problem the AI system is addressing. State how this AI system complements, replaces or provides a new solution to the problem (take from your Impact Assessment).	Impact Assessment
Deployment context	**Geographic focus:** **Affected population:** **Humanitarian crisis:**	Impact Assessment; Use Case Assessment

2.2 Red lines: full assessment with mitigations

Principle	Red line	Assessment	Escalation
Impartiality	**Systemic exclusion:** Could this system systematically exclude or disadvantage specific groups from assistance, protection, or information in ways that cannot be detected and corrected?	☐ Yes ☐ No ☐ Why or how	Residual risk: Low / Medium / High
Impartiality	**No meaningful redress:** Are there insufficient or inaccessible mechanisms for affected people to challenge, correct, or appeal AI-influenced decisions?	☐ Yes ☐ No ☐ Why or how	Residual risk: Low / Medium / High
Independence	**Loss of control:** Would reliance on suppliers, data providers, or infrastructure undermine organisational independence or create unacceptable conflict-related exposure?	☐ Yes ☐ No ☐ Why or how	Residual risk: Low / Medium / High
Humanity	**Harmful outputs:** Could the system plausibly generate or amplify dangerous misinformation or harmful guidance at scale without reliable safeguards?	☐ Yes ☐ No ☐ Why or how	Residual risk: Low / Medium / High
Humanity	**Protection and safeguarding risk:** Could the system generate or materially increase protection or safeguarding risks to affected people or staff without effective mitigation or redress mechanisms?	☐ Yes ☐ No ☐ Why or how	Residual risk: Low / Medium / High
Neutrality	**Conflict exposure:** Could data sources, vendors, hosting, or deployment create perceived or real alignment with a party to conflict?	☐ Yes ☐ No ☐ Why or how	Residual risk: Low / Medium / High

2.3 Broader principles alignment

These questions test principles alignment beyond the red lines. Each should be answered for Tier 2 and Tier 3 use cases.

Principle	Question	Response
Humanity	Is there a significant cost of error if outputs are wrong or misleading, and have you identified who could be harmed?	
Humanity	Do you need to implement safeguards to prevent affected people being exposed to harmful outputs (e.g., unsafe advice, dehumanising categorisation)?	
Humanity	Do you need to mandate a minimum level of human oversight before outputs are acted upon?	
Impartiality	Should you establish clear mechanisms for affected people to challenge or correct AI-influenced decisions?	

Impartiality	Should you identify how disparate performance or failure modes across relevant groups will be detected and addressed?	
Impartiality	Should you analyse and document which groups may be excluded due to language, connectivity, disability, gender norms, displacement, or documentation status?	
Independence	Does external control over critical system components (model, hosting, data pipelines) compromise your independence?	
Independence	Can the organisation pause, modify, or decommission the system without external permission?	
Neutrality	Could outputs or data be intercepted and exploited for surveillance, coercion, or political inference?	
Neutrality	If trade-offs exist between neutrality and delivering assistance, are the rationale explicit and are safeguards in place?	

2.4 Summary of humanitarian principles decision

Field	Entry
Overall humanitarian principles risk	Low / Medium / High
Are risks proportionate to expected humanitarian value?	Y / N, with rationale
Residual risks accepted	Y / N
Decision record	Summary of key risks and red lines identified, with decision to proceed, redesign, or not proceed, including rationale
Decision authority (role, name, date)	

Full Section 3: Data and model, training and evaluation summary

To be completed by: Technical Lead.

Purpose: Maintain datasheets, data statements, and model card-level transparency on core technical components during training and evaluation stage.

Difference from Lean: For higher risked and tiered systems, the Full version requires more detailed documentation and understanding of the data curation and model training and evaluation processes, metrics and benchmarks.

3.1 Data curation (complete for each training dataset)

Field	Entry	SAFE AI journey source
Name and version	State name, date, and dataset version.	■ Procurement
Representation (who or what) and bias	Who or what is represented? Are specific populations over or under-represented?	■ Procurement ■ Impact Assessment
Type and size	Describe type, size, and scope of dataset (e.g., survey of x households, satellite imagery of Amazon forest).	■ Procurement
Coverage	Describe the coverage and completeness of dataset (e.g., time period and geographic location). State any gaps as it relates to the humanitarian problem.	■ Procurement ■ Impact Assessment
Quality, accuracy, and reliability	Provide quality (e.g., resolution of images), accuracy and reliability measures. What QA or verification was performed?	■ Procurement ■ Technical Assurance
Data collection methodology and assumptions	How was data collected, by whom and what process?	■ Procurement
Consent and privacy protections	How was consent obtained? What privacy mechanisms were used? Is data proprietary?	■ Procurement ■ Impact Assessment
Source of data	Provide links or references to dataset.	■ Procurement
Community role in data validation	What role did affected communities have in validating data quality, representation, or appropriateness, including any data repurposed from service delivery?	■ How to Co-design with Affected Communities ■ Impact Assessment

3.2 Model selection

Field	Entry	SAFE AI journey source
Model features	Describe key features (variables) included in the model.	AI Architecture and Tech Stack Decision Guide
Model assumptions and approximations	State all assumptions and any approximations (numerical, optimisation). Include use and risk of proxies.	AI Architecture and Tech Stack Decision Guide; Impact Assessment
Model architecture decisions	State or provide link to decision and trade-off analysis of architecture choice.	AI Architecture and Tech Stack Decision Guide
Model training	State approach to model training including grounding and adaptation techniques.	Procurement

Principles tensions	Identify tensions between model or data choices and the four humanitarian principles (neutrality, independence, impartiality, humanity). State mitigations.	Impact Assessment; AI Architecture and Tech Stack Decision Guide

3.3 Data curation (complete for each evaluation dataset)

Field	Entry	SAFE AI journey source
Name and version	State name, date, and dataset version.	Technical Assurance
Representation (who or what) and bias	Who or what is represented? Are specific populations over or under-represented?	Technical Assurance
Type and size	Describe type, size, and scope of dataset.	Technical Assurance
Coverage	Describe coverage and completeness. Include protection-sensitive scenarios the system will encounter.	Technical Assurance
Quality, accuracy, and reliability	Provide quality, accuracy and reliability measures. Confirm evaluation data is distinct from training data unless explicitly designed that way.	Technical Assurance
Data collection methodology and assumptions	How was data collected, by whom and what process?	Technical Assurance
Consent and privacy protections	How was consent obtained? What privacy mechanisms were used? Is data proprietary?	Technical Assurance; Impact Assessment
Source of data	Provide links or references to dataset.	Technical Assurance

3.4 Metrics, analysis, and assurance

Field	Entry	SAFE AI journey source
Technical objectives	Describe the technical objectives to be met for the model to achieve the seven trustworthiness measures.	Technical Assurance
Bias and fairness	Describe which model bias and fairness measures are used.	Technical Assurance; Impact Assessment

Key performance metrics	Accuracy: Explainability: Robustness: Privacy-enhanced: Other technical metrics relating to architecture type: Describe how model performance will be measured, ensuring it meets and reflects system trustworthiness values and objectives.	Technical Assurance
Model technical risk and mitigation strategies	State if the AI model is at risk of drift, hallucinations, or other AI risks. State what considerations, mitigations and safeguards have been put in place, whether technical or procedural, to mitigate and manage such identified issues.	Risk Mitigation; Technical Assurance
Decision thresholds	State decision thresholds. Consider decision thresholds and considerations given that outcomes of AI models are probabilities.	Technical Assurance
Comparison with other model performance	List if there are similar models and how they perform in comparison (generally and in context, if known). State any known limitations of the model.	AI Architecture and Tech Stack Decision Guide
Other benchmarks for evaluation	State if there are any other benchmarks to evaluate performance.	Technical Assurance
Conditions impacting model performance	State what conditions may impact the performance of the model.	Technical Assurance
Relevant peer review publications, technical notes	List and provide links to any relevant peer review publications, technical notes, etc.	Procurement; Technical Assurance

3.5 Sign-off

Field	Entry
Decision record	SRO sign-off that data and model choices meet minimum requirements for safety, fairness, and appropriateness in context, including any accepted limitations.

Full Section 4: Deployment and lifecycle

Purpose: Confirm organisational preparedness, stack, monitoring approach, and decommissioning plan.

4.1 Organisational readiness and deployment environment

Completed by Programme Lead.

Field	Entry	SAFE AI journey source
Operational readiness	Skills and capacity in place: Y / N Infrastructure readiness: Y / N Budget readiness: Y / N	Onboarding Readiness Checklist
Deployment environment	Describe the environment in which the AI system will be deployed (e.g., stable Wi-Fi, active war, displacement camp infrastructure).	Use Case Assessment; Impact Assessment

4.2 System stack and infrastructure

Completed by Technical Lead.

Field	Entry	SAFE AI journey source
Software dependencies	What are software components and dependencies?	AI Architecture and Tech Stack Decision Guide
Cloud / hosting / compute providers	List cloud, hosting, and compute providers.	Procurement
Data flow and integration	Provide data flow and integration diagram.	Procurement; Impact Assessment
Interoperability	State any interoperability with humanitarian information systems.	AI Architecture and Tech Stack Decision Guide
Supply chain or vendor dependencies	Provide a supply chain list or diagram with vendor dependencies. Identify any conflict-relevant or surveillance-relevant exposure as a neutrality and independence test.	Procurement; Impact Assessment
AI humanitarian stack	See the humanitarian AI stack guidance in the SAFE AI Architecture and Tech Stack Decision Guide.	AI Architecture and Tech Stack Decision Guide

4.3 Monitoring and principles alignment over time

Field	Entry	SAFE AI journey source
Monitoring approach	State who reviews the system, on what cycle, and how findings are documented. Cross-reference the SAFE AI Technical Assurance for the seven domains and frequencies.	Technical Assurance

Principles alignment over time	What triggers a reassessment of humanitarian principles alignment? Context changes, partner changes, supplier shifts, data source changes, conflict escalation. Principles alignment is not assessed once at deployment and assumed thereafter.	Technical Assurance; Impact Assessment
Community feedback integration	How does community feedback reach the monitoring process? Who is responsible for acting on it?	How to Co-design with Affected Communities; Technical Assurance
Incident response readiness	Confirm incident response protocol is in place and known to staff. Reference the SAFE AI Technical Assurance incident response process.	Technical Assurance

4.4 Decommissioning plan

Completed by Programme Lead with legal and technical inputs.

Field	Entry	SAFE AI journey source
Duration	What is the expected duration of the AI system implementation?	Procurement
System requirement and consideration	Describe your decommissioning system requirement and considerations.	Procurement; Impact Assessment
Long-term impact	Describe potential post-decommissioning long-term algorithmic imprints on use of the system and on communities served.	Impact Assessment
Exit and continuity	Where the system serves a continuing function, describe what replaces it on decommissioning. Affected communities must not be left without access to the function the system performed.	Procurement; How to Co-design with Affected Communities
Decision record	Confirm this reflects how system performance, risks, and unintended impacts will be monitored in practice, including thresholds or triggers for review, escalation, or decommissioning.	

Full Section 5: Community in the loop and governance

To be completed by: Programme Lead and Legal/Ethical Team.

Purpose: Ensure operational readiness for oversight and affected-community participation in governance and accountability, with named routes to redress.

5.1 Community and participation

Field	Entry	SAFE AI journey source
Community engagement	How were affected people involved in design, testing, and what authority did they have in changes in AI system design, development, monitoring and evaluation? (100 words maximum.) If not appropriate, state why.	How to Co-design with Affected Communities
Participatory practices	Which mechanisms or practices have been fostered to empower affected communities regarding data, outcomes, design, development, or implementation?	How to Co-design with Affected Communities
Wider stakeholder engagement	Which other stakeholders or actors (humanitarian organisations, state or other entities, tech partners) are engaged in the design, development, deployment, and decision-making process?	Onboarding Readiness Checklist
Stakeholder ecosystem access	List humanitarian organisations, state or other entities, and tech partners, and their access to model inputs or outputs.	Procurement

5.2 Accountability and redress

Field	Entry	SAFE AI journey source
Accountable leads	Humanitarian: name, role, contact Technical: name, role, contact Legal and ethical: name, role, contact Senior Responsible Owner: name, role, contact	Onboarding Readiness Checklist
Oversight mechanisms	Which governance body oversees this system, on what cycle, and through what process? How are findings escalated?	Onboarding Readiness Checklist; Technical Assurance
Redress and grievance mechanisms	How can affected communities raise concerns, challenge decisions, or seek remedy? Are these mechanisms operational at the point of deployment?	How to Co-design with Affected Communities; Impact Assessment
Notification approach	How are affected people informed that the system is in use, what it does, what it does not do, and how decisions can be challenged? In which languages and through which channels?	How to Co-design with Affected Communities

5.3 Sign-off

Field	Entry
Decision record	Who holds final accountability for decisions to deploy, pause, redesign, or decommission the system, and how affected people can raise concerns, challenge outcomes, or seek redress.

Full Section 6: References and versioning

Completed by: Technical Lead with Programme and Oversight Body input.

Purpose: Provide audit trail, version management, and continuity over time.

6.1 Relevant documentation and references

Field	Entry
Datasets	Links or references to datasets.
Benchmarks	Links or references to benchmarks.
Transparency documents	Links to related datasheets, dataset statements, model cards, systems cards, or other AI transparency documents.
Humanitarian crisis	Links or references and background information on the humanitarian crisis.
Other relevant AI system or governance resources	Links or references to other relevant AI system or governance resources (e.g., NIST RMF, legislation, policies, etc.) referenced or used.
Linked assessments	Links to completed assessments (DPIA, AIA, Security Review, SAFE AI Impact Assessment, SAFE AI Technical Assurance findings, etc.).
Decision record	This card serves as the authoritative record of decisions, risks, and accountability for this AI system and should be updated throughout its lifecycle. Key governance decisions should be supported by reviewable documentation or evidence where necessary.

Version control table

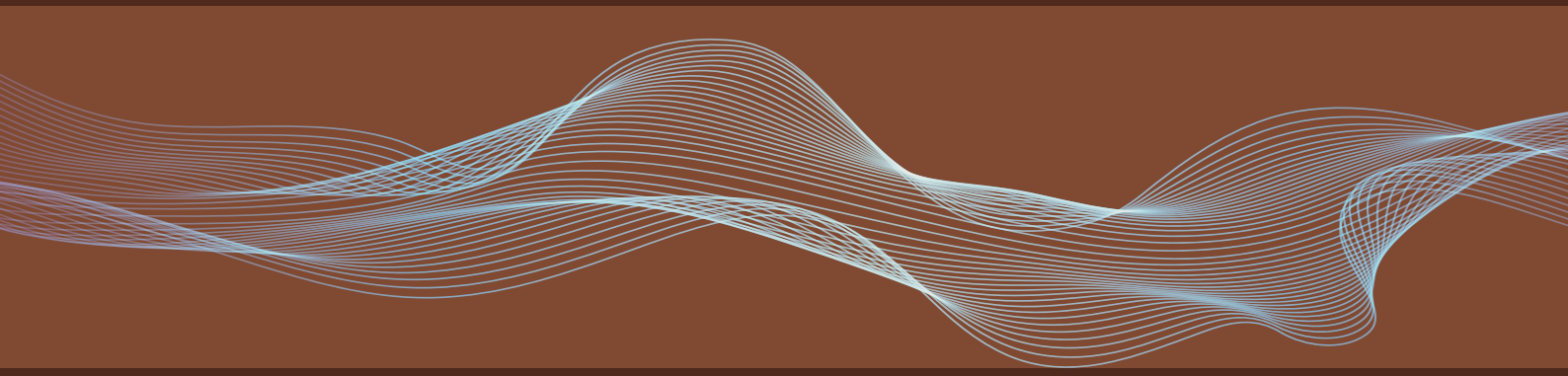
Card should be updated at key trigger points (e.g., model drift, change in access to affected persons, data provider changes, supplier updates, identified incidents).

Version	Date (DD/MM/YY)	Approver (name and contact)	Reason

REMEMBER!

If it's not captured in the Transparency Card, it is not considered governed by SAFE AI.

You can find more information and guidance about the SAFE AI Framework at cdacnetwork.org/safe-ai



SAFE AI

cdacnetwork.org/safe-ai