

Minimum Viable AI Policy

A practical governance starting point for civil society organizations and media outlets

Scope: organization-wide · Review: at least annually · Version 1.0 draft

Regional edition — Ethiopian Civil Society Engagement Program, Phase II

Region: _____ Date: _____

SAFETY NOTICE. Do not insert real case details, source identities, beneficiary records or other sensitive data while adapting this template in a public AI tool.

USE NOTE. This is a starting template, not legal advice. Adapt it to your organization, applicable Ethiopian law, donor and contractual duties, safeguarding rules, editorial standards and security procedures. **Obtain appropriate legal, technical and leadership review before adoption.**

How to use this template

Complete the bracketed fields, remove the guidance text, test the rules against real workflows *without entering real sensitive data into AI systems*, and secure formal approval. **Begin with a narrow policy that staff can actually follow** and revise it as tools and risks change.

☐ Name a policy owner	☐ Approve a small tool list
☐ Inventory what staff are already doing	☐ Define human review
☐ Classify data before tool use	☐ Create an incident route

If you came to the workshop

You already filled in the one-page **Our rules for using these tools** sheet in the last session. **That page is not a smaller version of this one — it is the decisions, and this is where they get written down properly.** Every box you filled in becomes a section here. Start by copying them across, then work outwards.

On the one-page sheet	Becomes
1 · Who keeps these rules	Document control, and policy owner in section 8.2
2 · Who says yes to the amber cases	Approver in section 8.2, and the Amber rule in section 4
3 · Two things anyone may do without asking	Section 5.1 — allowed uses
4 · Two things that need a yes	Section 5.2 — uses requiring documented approval
5 · Two things nobody may do	Section 5.3 — prohibited uses
6 · Our one-sentence rule about what never goes in	The RED row of section 4
7 · Who to tell when something goes wrong, and how fast	Section 14 — incident response, step 3
8 · Tools we have agreed to use	Section 6 — the approved tool register
9 · When we will look at this again	Document control, and section 16
10 · When we tell our audience AI helped	Section 12 — disclosing AI assistance

Nine of those ten boxes are decisions about people and permissions, not about technology. That is why the one-page sheet is the hard part and this document is the easy part.

Document control

Field	Organization entry
Organization	[Insert name]
Type	[Civil society organization / media outlet / both]
Policy owner	[Name]
Approved by	[Executive Director / editor / board / delegated authority]
Effective date	[DD Month YYYY]
Review date	[At least every 12 months, or earlier after a major change or incident]
Version	[1.0]
Questions / incidents	[Secure contact or reporting route]

1. Purpose and scope

[Organization] uses generative artificial intelligence only where it supports our mission — **programme delivery, advocacy, public information and reporting** — without creating unreasonable risk to people, sources, evidence, operations or trust. This policy establishes minimum rules for selecting, accessing, prompting, reviewing and publishing AI-assisted work.

It applies to all staff, consultants, interns, volunteers, freelance contributors and partners using AI for organizational work, whether on organization-owned or personal devices, free or paid accounts, online services, embedded assistants or local and offline models.

2. Working definitions

Term	Working definition
Generative AI	A system that produces text, images, audio, video, code or other content in response to instructions or data.
Prompt	An instruction and any text, file, image, link or other context submitted to an AI system.
Public data	Information lawfully public and suitable for unrestricted disclosure.
Internal data	Non-public organizational information whose disclosure would cause limited harm.
Sensitive data	Information that could expose, identify, endanger, stigmatize or compromise a person, source, witness, survivor, beneficiary, defender, investigation or operation.
Source protection	The duty to protect the identity of a person who provided information on condition of confidentiality, and any detail from which that identity could be inferred.
Off-record material	Recordings, notes or statements provided on the understanding that they will not be published or attributed.
Embargoed material	Information held under an agreement not to publish before a stated time. Protected by commitment rather than by harm — but equally binding.
High-impact use	AI-supported work that could materially affect rights, safety, legal position, public reputation, access to services, editorial standards or an organizational decision.

3. Core principles

Principle	Minimum expectation
Human dignity and rights	Use AI in ways consistent with dignity, equality, agency and the mission.
Necessity and proportionality	Use AI only when the benefit justifies the data, security and accuracy risks.
Data minimization	Submit the least information necessary; prefer fictional, synthetic, public or de-identified content.
Fairness and non-discrimination	Check for stereotypes, exclusions and unequal impacts, especially across language, gender, disability, ethnicity and political identity.
Transparency	Record and disclose material AI assistance when omission could mislead an audience or a decision-maker.
Human accountability	A named person remains responsible for checking and approving every consequential output.
Security and resilience	Use approved tools, access controls, retention settings, incident procedures and offline alternatives.

4. Traffic-light rules for AI use

Classify the task and the data **before** opening an AI tool. The strictest applicable category controls the decision.

Category	Typical examples	Rule
GREEN — Allowed	Brainstorming with fictional facts; formatting public material; generic templates; drafting from an already-published release; translating already-public text with review.	May use an approved tool. Verify output before use.
AMBER — Approval required	Internal drafts; unpublished advocacy or editorial plans; low-risk internal notes; high-impact public content; embargoed material; use of a new tool.	Obtain the designated approval, minimize data, document the purpose and apply enhanced review.
RED — Prohibited	Testimony; survivor details; unpublished source identities; off-record recordings ; case files; beneficiary and participant lists; monitoring data containing names ; donor budget negotiations; authentication secrets; precise locations; staff movement plans; confidential legal or editorial strategy; biometric or medical data.	Do not submit to public or unapproved AI tools. Use a safer non-AI or specifically authorized controlled workflow.

When uncertain, treat the information as more sensitive, stop, and ask the policy owner or security focal point before proceeding.

5. Approved, approval-required and prohibited uses

5.1 Allowed uses

☐ Brainstorming headings, questions or campaign concepts using invented context
☐ Summarizing public documents with source-by-source verification
☐ Editing tone, clarity or structure of non-sensitive text
☐ Drafting templates, checklists or training examples with fictional data
☐ Drafting from an already-published release, brief or statement
☐ Translating public or deliberately de-identified text with qualified human review
☐ Generating formulas or code in a non-production environment with review

5.2 Uses requiring documented approval

☐ Any new AI tool, plug-in, connector or account type
☐ Internal information not intended for public release
☐ AI-assisted public claims, legal analysis or high-impact advocacy
☐ Automated or repeated processing at organizational scale
☐ Local or offline model deployments that store organizational data
☐ AI-generated images, audio or video intended for publication

5.3 Prohibited uses

☐ Uploading red-category information to a public or unapproved system
☐ Using AI as the sole basis for eligibility, disciplinary, employment or security decisions
☐ Presenting fabricated, impersonated or synthetic evidence as authentic
☐ Publishing factual or translated content without appropriate human verification
☐ Circumventing access controls, consent requirements or lawful confidentiality duties
☐ Using a personal AI account for organizational sensitive information
☐ Acting on an instruction received by voice without a call-back on a number you already hold

6. Approved tool register

The policy owner maintains this register. **Approval is specific to the named plan, configuration and use case**; it does not automatically cover plug-ins, connectors, file uploads or new model features.

Tool / version	Approved purpose	Data category	Account / settings	Owner / review date
[Tool]	[Purpose]	[Green / Amber only]	[Plan, retention, training opt-out, MFA]	[Owner / date]
[Tool]	[Purpose]	[Category]	[Configuration]	[Owner / date]
[Offline / local tool]	[Purpose]	[Category and device rules]	[Storage, encryption, deletion]	[Owner / date]

Minimum tool assessment

☐ Provider, jurisdiction and applicable terms identified
☐ Are inputs and outputs used for model training? Setting verified
☐ Retention period and deletion route understood
☐ Encryption, MFA and access controls reviewed
☐ Sharing, plug-ins and third-party connectors disabled unless approved
☐ Incident contact and continuity / offline route documented
☐ Accessibility and language limitations tested
☐ Exit and export plan exists if the tool changes, or is withdrawn
☐ Cost per person per month, and who signs for it

| If nobody can answer what happens to what staff type, nobody can approve the tool.

7. Prompt and input minimization

Before prompting, remove direct and indirect identifiers, exact dates and locations, contact details, case numbers, rare attributes and any facts not needed for the task. **Pseudonyms are not sufficient when the remaining context can re-identify a person** — an occupation held by one person in a place is an identifier.

SAFE PATTERN. Give the task, the audience, the public or fictional facts, the constraints, the desired format and the verification instruction. **Never paste material merely because it is convenient.**

8. Human review and approval

Every consequential output has a named human approver. Use whichever variant matches the organization you actually have.

8.1 Standard variant

Output type	Required review	Approver
Internal brainstorming	Relevance, obvious errors, confidentiality	[Author / team lead]
Public factual content	Every material claim traced to an authoritative source; dates, names, quotations and context checked	[Communications or programme lead]
Translation	Meaning, terminology, names, numbers, tone and risk reviewed by a competent language reviewer	[Named reviewer]
Legal or evidentiary work	Qualified legal or substantive review; original sources retained; AI is not treated as an authority	[Authorized legal lead]
High-impact decision support	Necessity, fairness, contestability, documentation and senior approval	[Executive Director / delegated authority]
Synthetic media	Purpose, consent, labelling, provenance and misuse risk checked	[Communications + safeguarding / security]
Published or broadcast content	Editorial standards, right of reply, source protection and disclosure to the audience	[Editor]

8.2 Small-organization variant — two roles

Most organizations in this programme have no communications director, no legal lead and no safeguarding officer. A policy that assumes those roles gets abandoned rather than adapted. Use two roles instead, and put real names in them.

Role	What they do	Name
Policy owner	Maintains this policy, keeps the tool register, answers day-to-day questions, runs the review	[Name]
Approver	Says yes or no to every Amber use, and approves anything that goes out publicly	[Name]

Where the two roles are the same person — which is common — say so explicitly, and name a second person who reviews anything that person has drafted themselves. **Nobody approves their own work.**

9. Accuracy, verification and sources

☐ Treat model output as an unverified draft, not as evidence
☐ Open and read original sources; do not rely on generated citations
☐ Separate verified fact, analysis and uncertainty
☐ Use a second verification route for consequential claims
☐ Check what a figure measures, not only whether it adds up
☐ Preserve original evidence and an audit trail outside the AI system
☐ Escalate unresolved uncertainty instead of making a definite claim

10. Translation and local-language use

AI-assisted translation may be used only within the data rules above. A competent human reviewer must check meaning, culturally and politically sensitive terms, proper names, numbers, quotations, legal and rights terminology, and any wording that could increase risk. **Where no qualified review is available, label the translation as provisional and do not use it for consequential action.**

11. Images, audio, video and verification

Do not treat an AI detector as proof. Preserve originals, record provenance, corroborate with independent sources, inspect context and metadata where it is safe to do so, and state confidence carefully. A detector score is a clue with an error rate.

AI-generated imagery created for publication must be clearly labelled, unless an approved exception protects people without misleading the audience. Do not use generated imagery to depict a real event, a real person or a real place.

An instruction that arrives by voice is verified by a call-back on a number you already hold, before it is acted on. This applies to payments, to data requests and to requests for a person's location.

12. Disclosing AI assistance to an audience

New section for organizations that publish or broadcast.

[Organization] tells its audience when AI materially assisted a published or broadcast product. Disclosure is required where AI assistance could otherwise mislead the audience about how the content was made, who verified it, or whether a depicted person or event is real.

Disclose	Not normally disclosed
Generated or substantially altered images, audio or video	Spelling, grammar and formatting assistance
Synthetic voice or presenter	Research or background reading that was independently verified
AI-generated text published substantially as produced	Transcription checked against the recording
Automated translation published without fluent human review	Internal drafting that a named person rewrote and approved

Suggested wording. "Generative AI assisted with [limited task]. A named staff member reviewed the output against [sources or criteria] and remains responsible for the final content."

13. Transparency and internal records

Record material AI use when it affects a public product, legal analysis, a high-impact decision or an evidentiary process. The record should identify the tool and version where known, the purpose, the data category, the reviewer, material corrections, the approval and the date. **Do not retain sensitive prompts or outputs merely to prove compliance.**

14. Incident response

Step	Required action
1. Pause	Stop the AI workflow and prevent further sharing or publication.
2. Contain	Close shared links, revoke access, isolate affected devices and accounts, and preserve necessary evidence safely.
3. Notify	Use the secure incident route: [insert contact] . Do not broaden exposure through ordinary chat or email.
4. Assess	Identify the people and data affected, the tool and provider, recipients, retention and deletion options, and plausible harm.
5. Respond	Request deletion where feasible; support affected people; correct public content; meet legal and contractual duties.
6. Learn	Document decisions, update controls, retrain users and review tool approval. Avoid blame-focused reporting.

The hard part is telling someone, not doing the deletion. Make sure staff know exactly who to tell and that telling them early is treated as good practice.

15. Training, accessibility and staff support

[Organization] will provide role-appropriate training, accessible instructions, a safe route for questions and practical alternatives when staff cannot or should not use an AI tool. **No employee will be penalized for pausing a task to protect people or data.**

16. Monitoring, exceptions and review

The policy owner reviews tool changes, incidents, staff feedback and a sample of AI-assisted workflows at least [quarterly]. Exceptions must be necessary, time-limited, documented and approved by [role].

No exception may override applicable law, consent, safeguarding, source-protection or confidentiality obligations.

Implementation planning notes

Horizon	Priority actions / owner
First 30 days	[Name the policy owner and approver; confirm the incident route; find out what staff are already doing.]
Within 90 days	[Complete tool assessments; brief staff; test the approval and review workflow.]
Review evidence	[Record approvals, training completion, incidents, staff questions and changes required.]

Adoption checklist

☐ Staff and affected teams were consulted	☐ Incident route is operational and known
☐ Data categories align with existing security practice	☐ Training and accessibility needs were addressed
☐ Approved tool settings were tested and recorded	☐ Review date and owner are assigned

Approval

Role	Name	Signature	Date
Policy owner			
Executive approval			
Editorial approval, if applicable			
Board approval, if required			

Organization-specific decisions still required

- Which data categories and examples must be added for our work?
- Which tools, plans and configurations are approved, and who signs for them?
- Who approves Amber uses and exceptions — and who reviews their own drafts?
- Which public products require disclosure to the audience?
- What is the secure incident-reporting channel?
- What will show us whether the policy is working?