Network AI Policy

A Rights-Based Framework for Responsible AI Use

Effective: June 1, 2026 | Next Review: January 1, 2027

This policy includes:

Annex A: Use Case Classification Matrix
Annex B: Data Classification Reference Guide
Annex C: Commercial AI Tools Settings Guide

1. Introduction

Artificial Intelligence (AI) presents transformative opportunities for governance, accountability, and civic engagement. For Accountability Lab, which operates across fragile states, conflict-affected environments, and countries with restricted civic space, AI can expand our reach, improve our research, and lower barriers to community participation.

At the same time, AI introduces serious risks that are not abstract in our contexts. Misuse of AI can expose program participants to surveillance and retaliation, reproduce and amplify social biases, undermine the integrity of our research and communications, and erode the trust of the communities we serve. Using AI for content generation without checks and balances can also pose serious reputation risks.

This policy establishes a comprehensive framework for the responsible, rights-based, and operationally sound use of AI across Accountability Lab's global operations. It is grounded in a do-no-harm commitment and reflects our obligation to protect program participants, partners, sources, and staff from AI-related harms, including those arising from the inadvertent or negligent use of commercial tools.

2. Purpose

This policy aims to:

- Align AI use with our values, and with our obligations under international human rights frameworks.
- Establish clear obligations for the protection of program participants, team members, partners, and others whose data or identity may be at risk.
- Promote transparency and accountability in all AI-related activities, internally and in our external communications.
- Provide clear, practical guidance for staff and partners on the ethical and secure use of AI tools across all operating contexts.
- Establish a responsive governance structure for AI oversight that can adapt to a rapidly evolving landscape.
- Foster a culture of responsible experimentation and contribute to the broader civil society knowledge commons on AI governance.

3. Definitions

For the purposes of this policy, the following terms apply:

Term	Definition
AI Tool	Any software application, model, or platform using machine learning, large language models, or automated decision-making to generate content, analyse data, make predictions, or take actions. Includes generative AI tools, agentic AI systems, AI-assisted features embedded in existing software, and automated workflow tools.

Agentic AI	AI systems that can take autonomous or semi-autonomous actions across multiple steps without direct human input at each step — including browsing the web, executing code, sending emails, reading files, or calling external services. Distinct from standard prompt-response interactions.
Embedded AI	AI features integrated into software platforms already in use by AL staff, including AI-generated summaries in Google Workspace, AI meeting assistants in Zoom or Google Meet, and AI writing assistance in Notion. These may be enabled by default and require explicit configuration and governance.
Program Participant	Any individual who participates in or is the subject of AL's programs, including civic monitors, community members, research informants, activists, beneficiaries, and local partners. This category receives the highest level of data protection under this policy.
Source	Any individual who provides information to AL in confidence, including whistleblowers and individuals disclosing sensitive governance or accountability information. Sources are entitled to the same protections as program participants and often require additional protection given active threat environments.
Sensitive Data	Data classified at Tier 3 (Confidential) or Tier 4 (Protected) under this policy. See Section 6 and Annex B for the full classification framework.
Public AI Tool	Any AI tool or platform whose default or consumer configuration does not include contractual data protection commitments equivalent to Tier 3 requirements, or where the provider may use inputs to train or improve models. Free consumer tiers of major AI tools are treated as Public AI Tools unless explicitly confirmed otherwise.
Enterprise AI Tool	A version or tier of an AI tool that includes contractual data protection commitments, organizational controls over data retention and training opt-outs, and administrator-level privacy configuration. Enterprise or Teams tiers from major providers generally qualify; this must be verified against Annex C and the Approved Tools List.
Do-No-Harm	The principle that AL's activities must not directly or indirectly increase risks to the safety, security, dignity, or rights of program participants, staff, partners, or communities — including through negligent or inadvertent use of technology. Harm prevention takes precedence over efficiency gains.

4. Scope

This policy applies to:

- All employees, contractors, consultants, volunteers, fellows, and partners who use or develop AI tools on behalf of Accountability Lab or in connection with AL-funded activities.
- All organizational activities involving AI tools, including internal operations, communications, content development, data analysis, monitoring and evaluation, and external engagement.

- All AI tools used in the course of AL's work, regardless of whether procured by AL or brought in by individual staff or partners.
- All geographies and country programs. Local legal requirements, threat environments, and community norms may require more restrictive application. See Section 12 on Country-Level Addenda.

Where the line between personal and professional use is unclear, staff should apply this policy.

5. Guiding Principles

The following principles underpin all AI-related decisions and activity at Accountability Lab. They are grounded in international human rights standards, including the UN Guiding Principles on Business and Human Rights, the OHCHR guidance on AI and human rights (A/HRC/51/26), and the data protection rights recognized in regional frameworks applicable in our operating countries.

Do No Harm

We will not use AI in ways that increase risk of harm to program participants, sources, partners, staff, or communities — including through inadvertent data exposure, discriminatory outputs, or loss of community trust. Where a potential AI use raises harm concerns that cannot be adequately mitigated, we will not proceed regardless of efficiency benefits.

Rights-Based Approach

AI use must be consistent with the rights of the communities we serve, including rights to privacy, non-discrimination, dignity, and participation. We apply heightened scrutiny to any AI use that could affect these rights, and we centre the perspectives of affected communities in evaluating AI tools.

Mission Alignment

AI must advance our strategic goals in governance, accountability, and civic empowerment. We do not adopt AI tools solely for efficiency if their use conflicts with our mission or values.

Transparency

We disclose AI use in public-facing outputs. We document AI use internally. We are honest with program participants and partners about whether and how AI tools may interact with their data or work.

Human Accountability

Final responsibility for all outputs and decisions rests with human staff. AI tools assist human judgment — they do not replace it. We maintain meaningful human oversight of all AI-assisted work and do not make consequential decisions about individuals solely on the basis of AI outputs.

Equity and Non-Discrimination

AI tools must not reproduce, amplify, or entrench social biases based on gender, race, ethnicity, religion, nationality, disability, sexual orientation, or other protected characteristics. We assess AI tools for bias risks prior to approval and monitor for discriminatory outputs in use.

Privacy and Security

We protect the privacy of individuals whose data we hold. Sensitive data, particularly data relating to program participants and partners must not be processed through AI tools that lack adequate data protection guarantees. This applies regardless of apparent convenience.

Proportionality

AI use should be proportionate to the task. More powerful tools that carry higher data risks should not be used where simpler, lower-risk approaches are adequate.

Sustainability

We prefer energy- and cost-efficient tools. We consider the environmental impact of AI use as part of our broader sustainability commitments.

Learning and Adaptation

We learn from experimentation and share findings with the civil society community. This policy is a living document updated in response to developments in AI technology, governance, and field experience.

6. Data Classification

Not all information that AL staff handle carries the same level of risk. Our data classification system allows staff to make consistent, proportionate decisions about which AI tools can be used with which types of information. Annex B provides a detailed reference guide with examples drawn from AL's operational contexts.

Tier	Label	Description	AI Tool Restriction
1	**PUBLIC**	Information intended for public distribution or with no reasonable expectation of confidentiality.	Any approved AI tool, including consumer-tier tools.
2	**INTERNAL**	For AL staff and approved partners only. Not public but not personally sensitive.	Approved tools with training opt-out confirmed. Enterprise tools strongly preferred.
3	**CONFIDENTIAL**	Disclosure could cause material harm to AL operations, reputation, or partner obligations. Includes staff personal data, financial data, and unreleased research.	Enterprise-tier tools only, with training opt-out confirmed and organisational access controls in place.
4	**PROTECTED**	Disclosure could directly endanger individuals or cause serious rights violations. Includes all program participant data and location data in high-risk contexts.	Must NOT be input into any commercial AI tool. Default position: no AI processing.

> *Tier 4 (Protected) data must never be entered into any commercial AI tool regardless of privacy settings, or provider assurances. When working with mixed-tier data, the highest tier governs tool selection.*

7. Use Classification Framework

Each AI use case is evaluated according to one of three classifications. See Annex A for the full Use Case Matrix.

Classification	Meaning	Governance
ENCOURAGED	Low risk, high benefit. Tools have a demonstrated record of accuracy. Potential for harm or loss of trust is low. *Examples:* ● *Using Google Workspace AI to suggest edits and improve clarity on an internal memo* ● *Running an interview transcript (with no participant-identifiable content) through a summarisation tool to identify emerging themes* ● *Using AI translation to make a public policy brief accessible in French, Nepali, or Spanish etc.* ● *Using built-in Zoom or Google Meet transcription (with training opt-out confirmed) to produce notes from an internal team meeting* ● *Deploying an agentic AI tool to assist with research tasks, where the tool may browse external sources autonomously*	Staff may proceed with responsible oversight. Accountability rests with the role accountable for that work area.
CASE-BY-CASE	Benefits and risks depend on context. Tools may not always be reliable. Potential for harm is low-to-moderate. *Examples:* ● *Using AI to draft a grant proposal that includes program data, strategies, or unpublished findings (Tier 3 content; requires enterprise tool and Country Director/Co-CEO approval)*	Country Director or team lead must approve before use. Approval must be documented.

	• *Using AI translation for a long-form research report before publication, where errors could damage credibility with key stakeholders* • *Using an AI data analysis tool to identify patterns in a large anonymized dataset for program monitoring/review* • *Using AI to assist in screening job or fellowship applications (requires human final decision on all candidates)* • *Generating a realistic voiceover or synthetic media for an external campaign*	
PROHIBITED	Risks clearly outweigh benefits. Moderate-to-high potential for serious harm, liability, or fundamental conflict with our mission. *Examples:* • *Entering a participant, partner or staff member's name, location, or feedback into any commercial AI tool, even to summarise or translate it* • *Allowing a third-party AI note-taking bot (e.g., read.ai, Otter.ai, Fireflies) to join any meeting involving program participants, partners, or for internally sensitive discussions* • *Using AI to generate realistic video, audio, or images depicting real individuals without their explicit consent and public disclosure* • *Using AI to monitor, score, or profile staff productivity or attendance* • *Making a funding, employment, or program eligibility decision based solely on AI-generated output without human review* • *Using facial recognition on photographs of community members, partners, or event attendees* • *Using AI to draft a grant proposal, or translate or summarise text without any human intervention*	Not permitted without explicit written approval from the Country Director or Co-CEOs. This bar is intentionally very high.

8. Operational Guidelines

8.1 Tool Approval and Procurement

Staff must consult the current Approved AI Tools List before using any AI tool for AL-related work. The list is maintained by the AL AI Lead. New tools proposed for use must be submitted for vetting covering:

- Data privacy policy and terms of service, with attention to training data practices, data retention, and third-party sharing.
- Whether enterprise or organizational-tier access is available with adequate privacy controls.
- Compliance with relevant data protection legislation in AL's operating jurisdictions.
- Environmental and cost considerations.
- Fitness for purpose, including suitability for use in low-bandwidth or high-risk field environments.

Vetting decisions are made by the AL AI Lead in consultation with the relevant Country Director and Co-CEOs. Approval is required before use begins.

8.2 Disclosure Requirements

All public content generated or substantially assisted by AI must include a clear disclosure (e.g., "This content was produced with AI assistance and reviewed by [role]"). "Substantially assisted" means AI contributed to the structure, argument, or language of the content, not merely spell-checking or formatting.

Teams must document significant AI usage and decisions in project records. Where AL uses AI tools in direct engagement with program participants or communities, this must be disclosed to those participants in advance, in language they can understand.

8.3 Human-in-the-Loop

AI outputs must be reviewed, fact-checked, and verified by a responsible human staff member before distribution or use in decision-making. The reviewing staff member is accountable for accuracy, fairness, and appropriateness. The use of AI does not transfer or diminish that accountability.

AL does not make consequential decisions about individuals, including decisions on employment, work performance, program eligibility, safety assessments, or grant allocation, based solely or primarily on AI outputs.

8.4 Participant and Source Protection

> *This section reflects AL's heightened obligations to the individuals who trust us with their safety and information. Staff working in program delivery, data collection, or community engagement should treat this as the most operationally important section of this policy.*

Program participant data and source information are Tier 4 (Protected). The following requirements apply absolutely:

- Do not input the name, location, contact information, testimony, or any other identifying information of a program participant or source into any commercial AI tool.
- Do not upload documents, interview transcripts, meeting notes, or case files containing participant or source information to any AI tool, even if you intend to "anonymize" content within the prompt.
- Do not use AI tools to generate, summarise, or analyze content in ways that could allow a third party to identify a program participant or source.
- Be aware that metadata — file names, folder paths, email headers, and phrasing patterns — can be identifying. Anonymizing names does not necessarily make a document safe for AI processing.

Consent for AI processing of participant data is not a substitute for this restriction. Even with consent, commercial AI tools do not provide the security guarantees our duty of care requires.

Any project proposing to use AI tools in direct interaction with program participants requires an explicit data protection impact assessment and approval from the relevant Country Director and the AL AI Lead before deployment. Staff who are uncertain whether specific content is safe to process with an AI tool should ask their Country Director before proceeding.

8.5 Agentic and Embedded AI Tools

Agentic AI tools take autonomous actions across multiple steps and can access, process, and transmit information beyond what a user explicitly enters in a single prompt. Before enabling or using any agentic AI tool for AL work, staff must:

- Understand what data the agent will access (files, emails, calendar, contacts, etc.) and confirm that no Tier 3 or Tier 4 data is within the agent's scope.
- Review and restrict the agent's permissions to the minimum necessary for the task.
- Obtain approval from the AL AI Lead for any agentic AI use beyond basic workflow automation with Tier 1 or Tier 2 data.

- **Embedded AI tools** — AI features integrated into Google Workspace, Zoom AI Companion, Notion AI, and similar platforms — may be enabled by default in organizational accounts. Country Directors and operations leads are responsible for reviewing and configuring embedded AI features in accordance with this policy for their programs. Where embedded AI features cannot be configured to meet this policy's requirements, they must be disabled for that program or context.

Staff should not assume that an embedded AI tool is safer or more private simply because it is integrated into a familiar platform. Data protection guarantees vary significantly. See Annex C.

8.6 Prohibited Practices

The following practices are prohibited under all circumstances:

- Inputting Tier 4 (Protected) data into any commercial AI tool.
- Using AI for surveillance, profiling, or monitoring of AL staff or program participants.
- Generating or distributing deepfake content — synthetic realistic images, video, or audio of real individuals without their explicit consent and public disclosure.
- Using AI to generate content that impersonates real individuals or organisations.

- Automated decision-making that directly affects employment, benefits, program eligibility, or the safety of individuals without human review and accountability.
- Using AI tools to process communications of program participants or sources without explicit approval from the AL AI Lead.
- Using AI tools that do not appear on the current Approved Tools List without prior vetting.
- Granting AI tools access to Tier 3 or Tier 4 data without explicit written approval.
- Using free consumer tiers of AI tools for anything classified Tier 2 or above unless training opt-out has been confirmed in writing for that specific tool.

> *Facial recognition on images of program participants, sources, or community members is prohibited in all forms. Use of facial recognition for internal purposes (e.g., photo organisation) is strongly discouraged and must never involve participant photographs.*

9. Governance Structure

Effective AI governance requires clear accountability, not aspirational committee structures. The following roles and responsibilities apply.

AL AI Lead

Responsibilities include: maintaining and updating the Approved AI Tools List; reviewing and approving new tool requests; conducting semi-annual policy reviews; reviewing and adjudicating incident reports; coordinating AL's civil society AI governance contributions; and approving exceptions to this policy in consultation with Country Directors and Co-CEOs.

Country Directors and Co-CEOs

Responsibilities include: ensuring all staff and contractors in their programs understand and apply this policy; approving Case-by-Case AI uses in their program area; reviewing and configuring embedded AI tool settings for their accounts; escalating incident reports to the AL AI Lead; and developing Country-Level Addenda where local conditions require additional restrictions.

All Staff

Responsibilities include: completing required AI policy training within 60 days of joining AL and annually thereafter; consulting the Approved Tools List and this policy before AI use; classifying data correctly; reporting AI-related incidents promptly; and raising concerns with their Country Director rather than making unilateral decisions.

Policy Review

This policy is reviewed semi-annually by the AL AI Lead in consultation with Country Directors. Reviews consider changes in the AI tool landscape, lessons from incident reports, staff and community feedback, and changes in legal and threat environments. Urgent revisions can be initiated at any time by the AI Lead with Co-CEO approval.

10. Incident Reporting and Response

An AI-related incident is any event in which:

- Tier 3 or Tier 4 data is input into an AI tool without appropriate authorization.
- An AI tool generates or distributes harmful, inaccurate, or discriminatory content in AL's name.
- An AI-assisted decision or output causes or is credibly alleged to cause harm to a program participant, partner, or staff member.
- A prohibited practice (Section 8.6) is carried out, regardless of intent.
- A staff member suspects an AI tool has transmitted or exposed AL data in an unauthorized manner.

Reporting

Incidents should be reported to the relevant Country Director and the AL AI Lead within 24 hours of discovery. Reports should describe: what happened, what data or content was involved, the approximate time and context, and any immediate steps already taken. Staff will not face adverse consequences for good-faith incident reporting.

Response

On receipt of an incident report, the AL AI Lead will: assess potential harm to affected individuals and escalate to Co-CEOs if serious harm is possible; coordinate with the Country Director on immediate protective measures; determine whether notification to affected program participants or partners is required; document the incident and any remedial actions; and incorporate lessons into the next policy review.

> *Where an incident involves a credible risk to the safety of a program participant or source, the response must prioritise their protection over organisational reputation management.*

11. Staff Training and Accountability

Required Training

All staff must complete an AI policy orientation within 60 days of joining AL, covering the key principles of this policy, the data classification system, how to use the Approved Tools List and configure tools correctly, and incident reporting procedures.

All staff must complete an annual AI policy refresher incorporating updates to the policy, new tool guidance, and lessons from the past year. Country Directors and program leads must complete supplementary training on AI governance and participant data protection.

Accountability

Adherence to this policy is expected of all staff and contractors. Deliberate violations — particularly those involving Tier 4 data or prohibited practices — will be treated as serious disciplinary matters in accordance with AL's HR policies.

Staff who have concerns about this policy's application, or who believe they are being pressured to violate it, should escalate to the AL AI Lead or, if necessary, to AL's Co-CEOs.

12. Country-Level Addenda

This global policy establishes minimum standards. Country Directors may develop Country-Level Addenda imposing additional restrictions appropriate to local legal environments, threat landscapes, and community norms. Addenda may never reduce the protections of this policy; they may only strengthen them.

Addenda should address at minimum:

- Applicable local data protection legislation and how it interacts with this policy — including Nigeria's NDPR, South Africa's POPIA, Zimbabwe's Data Protection Act, Nepal's Privacy Act, Pakistan's Personal Data Protection Act, and Mexico's LGPDPPSO.
- Specific threat actors or surveillance environments requiring heightened restrictions.
- Community norms or partner agreements limiting certain AI use cases.
- Local incident escalation contacts and procedures.

Country-Level Addenda must be approved by the AL AI Lead and Co-CEOs before taking effect and reviewed at the same time as this global policy.

13. Sharing and Collaboration

Accountability Lab is committed to contributing to the civil society knowledge commons on responsible AI governance. We invite peer organisations to adapt, translate, and build on this policy, and ask only that adaptations maintain its rights-based and do-no-harm foundations.

For inquiries, contributions, or to share lessons from your own implementation: info@accountabilitylab.org

ANNEX A

Use Case Classification Matrix

This matrix provides classification guidance for specific AI applications. Classifications assume use with Tier 1 or Tier 2 data. Use with Tier 3 or Tier 4 data requires escalation regardless of the classification shown here.

Text Editing

Application	Benefits	Risks	Classification
Spelling and grammar check	Increases efficiency and equity for non-native speakers; high accuracy	Minimal	**ENCOURAGED**
Editing for content, organisation, and style	Improves quality and consistency	Minimal for Tier 1–2 content	**ENCOURAGED**
Brand voice consistency (requires feeding content samples)	Improves brand consistency	Requires feeding proprietary content into AI; enterprise tools only	**CASE-BY-CASE**

> *Use AI editing features already integrated into approved platforms (e.g., Google Workspace AI) with training opt-out confirmed. Enterprise tools are preferred for any content Tier 2 and above. Free consumer tiers should not be used for editing documents that are not already publicly available.*

Translation

Application	Benefits	Risks	Classification
Live translation in video calls	Increases accessibility; lowers costs	Variable accuracy by language	**ENCOURAGED**
Short-form text (emails, messages)	High accuracy for widely spoken languages	Error-prone for less widely spoken languages	**ENCOURAGED**
Long-form documents for external publication	Significantly lowers translation costs	Accuracy decreases with length and complexity; risk of loss of trust	**CASE-BY-CASE**

> *Experiment with multiple tools and test accuracy across all languages used. Disclose AI translation in externally facing communications. Do not use AI translation for documents containing Tier 3 or Tier 4 content without enterprise tool approval.*

Transcription

Application	Benefits	Risks	Classification
Internal transcription using built-in platform tools with training opt-out confirmed	Increases accessibility and efficiency; generally high accuracy	Errors can reduce trust; data stored on provider servers	**ENCOURAGED**
Third-party AI note-taking bots joining meetings	Convenience	Cannot monitor accuracy or manage distribution; significant privacy and security risks	**PROHIBITED**
Any transcription of meetings involving program participants or sources	Documentation	High risk of exposing Tier 4 data	**PROHIBITED**

> *Third-party AI note-taking bots must be removed from all non-public meetings. Inform participants that AL does not permit them. Designate a human notetaker or use internal transcription features with training opt-out confirmed. Built-in transcription tools (e.g., Zoom AI Companion) must also be configured correctly — see Annex C.*

Content Summarisation

Application	Benefits	Risks	Classification
Summarising internal meeting notes or email threads (no Tier 4 content)	High efficiency; can surface themes	May lose nuance; occasional errors	**ENCOURAGED**
Summarising published or publicly available content	High efficiency; generally reliable	Not always reproducible	**ENCOURAGED**
Summarising complex research, field reports, or partner content for external publication	Efficiency; theme identification	Variable accuracy; risk of losing critical nuance; trust loss if errors appear publicly	**CASE-BY-CASE**

> *For summaries that will be shared publicly, validate outputs by attempting to reproduce them and/or spot-checking against source material. Always disclose AI use. Never summarise documents containing Tier 4 data using commercial tools.*

Novel Text Generation

Application	Benefits	Risks	Classification
Internal drafts, brainstorming, emails (Tier 1–2 content)	Efficiency; equity for non-native speakers	Inaccuracies; over-reliance risk	**ENCOURAGED**
External short-form content (social media, newsletters)	Efficiency; reach	Brand voice risk; transparency required	**ENCOURAGED**
Long-form external content (blog posts, reports, grant proposals)	Drafting efficiency	Accuracy; plagiarism; trust risk if undisclosed	**CASE-BY-CASE**

> *AI drafts require significant human review and editing before publication. Disclose AI assistance in all public-facing content. Grant proposals require Country Director or Co-CEO sign-off before AI-assisted versions are submitted.*

Workflow Automation

Application	Benefits	Risks	Classification
Database updates, scheduling, simple email routing (Tier 1–2 data)	Efficiency; staff capacity	Errors increase with task complexity; requires regular monitoring	**ENCOURAGED**
Complex multi-step automation involving Tier 2–3 data	Significant capacity gains	Error propagation; requires documented monitoring protocols	**CASE-BY-CASE**

Agentic AI

Application	Benefits	Risks	Classification
Agentic tools with strictly scoped access to Tier 1–2 data (e.g., research assistant	Significant productivity gains	Unexpected data access; hallucinated actions	**CASE-BY-CASE**

browsing public sources)			
Agentic tools with broad system access (email, file systems, contacts)	High automation potential	Risk of accessing and transmitting Tier 3–4 data without staff awareness	**PROHIBITED**

> *Agentic AI tools must never be granted access to email systems, file storage, or communication platforms that contain Tier 3 or Tier 4 data without explicit written approval from the AL AI Lead.*

Data Analysis

Application	Benefits	Risks	Classification
Pattern identification on anonymised or public data	Efficiency; pattern recognition	Variable accuracy; results may not be reproducible	**ENCOURAGED**
Sentiment analysis, image analysis (non-participant content)	Efficiency; scale	Bias risk; accuracy variation	**CASE-BY-CASE**
Predictive analysis informing program or resource decisions	Evidence base	High error rate in complex applications; bias-driven decision risk	**CASE-BY-CASE**
Any analysis involving Tier 4 (program participant) data	N/A	Unacceptable data protection risk	**PROHIBITED**

> *Human verification of analytical outputs is required before they inform program or publication decisions. Predictive outputs must not drive consequential decisions without significant corroborating evidence and human review.*

Image, Media, and Video Editing and Generation

Application	Benefits	Risks	Classification
Minor photo editing — no participant images (retouching, colour/lighting, background removal)	Efficiency	Minimal if no participant images involved	**ENCOURAGED**

Generating non-realistic visual content (logos, illustrations, abstract art)	Lowers design barriers	Unintentional IP use; bias in generated content	**ENCOURAGED**
Generating realistic images or voiceovers for campaigns	Production quality	Trust risk; potential for misrepresentation	**CASE-BY-CASE**
Generating deepfakes or realistic synthetic media depicting real individuals	N/A	Severe misrepresentation risk; violates dignity rights	**PROHIBITED**
Facial recognition on images of participants, sources, or community members	N/A	Severe rights violation risk	**PROHIBITED**

IT, Compliance, and Assessment

Application	Benefits	Risks	Classification
Spam filters, virus scanning, network security (trusted platforms)	Efficiency; high accuracy	Minimal with trusted providers	**ENCOURAGED**
Legal or compliance monitoring of contracts or website content	Early compliance issue detection	Not a substitute for human legal review; over-reliance risk	**CASE-BY-CASE**
AI screening of job or fellowship applications	Efficiency	Bias risk; qualified candidates rejected; trust risk	**CASE-BY-CASE**
Monitoring or evaluating staff activity or productivity	N/A	Severe trust violation; contradicts our values and rights-based commitments	**PROHIBITED**

> *No third-party software should be granted access to staff email data. AI may assist compliance screening but must not replace human final decision-making for employment, grant, or membership decisions.*

ANNEX B

Data Classification Reference Guide

This guide helps AL staff correctly classify information before using any AI tool. When in doubt, classify at the higher tier and seek guidance from your Country Director or the AL AI Lead.

Tier 1 — PUBLIC

> **AI restriction: Any approved AI tool, including consumer-tier tools. Standard care applies.**

Tier 1 information is intended for public distribution or carries no reasonable expectation of confidentiality. Its disclosure causes no harm to any individual or to AL's operations.

Examples:

- Published reports, blog posts, policy briefs, and press releases.
- AL website content and social media posts.
- Job postings and public event announcements.
- Information that has already been publicly reported or disclosed.
- General research and analysis on public datasets.

Tier 2 — INTERNAL

> **AI restriction: Approved tools with training opt-out confirmed. Enterprise tools strongly preferred. Free consumer tiers only if training opt-out is confirmed in writing.**

Tier 2 information is for AL staff and approved partners only. It is not public, but its disclosure does not directly harm individuals. It may contain strategic, operational, or reputational information AL needs to protect.

Examples:

- Internal staff communications and meeting notes (where no participant information is discussed).
- Draft documents and work plans not yet approved for external sharing.
- Internal project updates and operational reports.
- General partner correspondence not containing sensitive information.
- Staff directories and internal contact lists.
- Non-sensitive financial and administrative records.

> *When internal documents reference program participants, sources, or field security contexts — even in passing — those portions must be treated as Tier 4. The whole document does not need to be re-classified, but those portions must not be fed into AI tools.*

Tier 3 — CONFIDENTIAL

> **AI restriction: Enterprise-tier AI tools only, with confirmed training opt-out and organisational access controls. Verify settings before use. See Annex C.**

Tier 3 information could cause material harm to AL's operations, reputation, or obligations to partners if disclosed without authorisation. It includes sensitive organisational information and personal data about staff.

Examples:

- Staff personal data: HR files, performance records, health information, salary data.
- Financial data: budget details, donor information, unpublished financial reports, payment records.
- Unreleased research findings and draft investigative reports.
- Confidential partner agreements, MoUs, and grant proposals under negotiation.
- Legal correspondence and case files not relating to program participants.
- Security and risk assessments for AL programs.
- Donor relationship information not intended for public attribution.

Tier 4 — PROTECTED

> **AI restriction: MUST NOT be input into any commercial AI tool. No exceptions without Co-CEO written approval and use of specifically approved locally deployed tools.**

Tier 4 information could directly endanger individuals, enable targeting by hostile actors, result in serious rights violations, or cause irreversible harm if disclosed. This category exists because of the specific nature of AL's work — in our operating contexts, the exposure of an individual's identity, location, or testimony can have life-altering or life-threatening consequences.

Examples:

- Names, contact details, or biographical information of program participants, civic monitors, or community members.
- Identities, contact information, or testimonies of sources, whistleblowers, or confidential informants.
- Location data, travel patterns, or activity logs of individuals in high-risk environments.
- Interview transcripts, testimonies, or case notes containing participant-identifiable information.
- Sensitive legal case information involving program participants.
- Security incident reports naming individuals at risk.
- Information about AL staff operating in hostile or restricted environments where identification could create risk.
- Communications or documents shared by partners under a strict confidentiality expectation relating to their own at-risk participants.

> *The prohibition on commercial AI tool use with Tier 4 data is absolute. Even enterprise versions with training opt-out do not provide the security guarantees AL's duty of care requires. The default position is: no AI processing of Tier 4 data.*

Where AI-assisted analysis of Tier 4 data is genuinely necessary, the required process is:

1. Prepare a written rationale explaining why AI processing is necessary and what alternatives have been considered.
2. Obtain written approval from the relevant Country Director and the AL AI Lead.
3. Use only locally deployed or self-hosted AI tools that have been specifically approved for this purpose, in an environment that does not transmit data to any external server.
4. Document the decision, tools used, and outputs in the project record.

Quick Reference: What Can I Use?

Tier	Free Consumer	Paid Consumer	Enterprise	Local / Self-hosted	Notes
1 — PUBLIC	**YES**	**YES**	**YES**	**YES**	Verify tool is on Approved List.
2 — INTERNAL	Only if training opt-out confirmed	If training opt-out enabled	**YES**	**YES**	Enterprise strongly preferred. Confirm settings per Annex C.
3 — CONFIDENTIAL	**NO**	**NO**	YES — with confirmed org controls	**YES**	Verify enterprise settings. Document approval.
4 — PROTECTED	**NO**	**NO**	**NO**	Only with Co-CEO written approval	Default: NO AI processing.

ANNEX C

Commercial AI Tools Settings Guide

This annex provides practical guidance on configuring commercial AI tools to meet AL's data protection requirements. It covers the most commonly used tools as of mid-2025. Tool configurations and policies change frequently — always check the current Approved Tools List maintained by the AL AI Lead.

> *This guide addresses what "safer" looks like for Tier 1–3 content on properly configured enterprise tools. No commercial AI tool configuration makes it appropriate for Tier 4 (Protected) data. That restriction is absolute.*

General Principles for All Commercial AI Tools

Before using any AI tool with AL content, confirm all of the following:

- The tool appears on the current Approved AI Tools List.
- You are using an organisational account, not a personal account. Personal accounts cannot be configured at the organisational level and must not be used for AL work above Tier 1.
- Training on user data is disabled — confirmed through account settings or contractually guaranteed. Free consumer tiers of most tools do not offer this by default.
- You understand the data retention policy. Disabling training does not necessarily mean data is not stored. Providers may retain inputs for safety review or other purposes. Understand the retention period.
- The tool's permissions are limited to what you are actively working on. You have not granted the tool access to your file system, email, calendar, or contacts unless that access has been specifically approved.

If you cannot confirm any of the above, do not use the tool for anything above Tier 1 content, and consult your Country Director or the AL AI Lead.

Claude — Anthropic

Setting / Consideration	Guidance for AL
Required tier	Claude Pro (individual) or Claude Teams/Enterprise. The free tier (claude.ai) is not approved for Tier 2 or above content.
Training opt-out — Pro	Go to Settings > Privacy. Find "Improve Claude for everyone" and set to OFF. Confirm the setting is saved. This must be done for each individual user account.

Training opt-out — Teams/Enterprise	Disabled by default for all users in the organisation. Verify with your account administrator that this setting has not been changed.
Data retention	Anthropic retains conversation data for a period even with training opt-out enabled. Data may be reviewed by Anthropic staff for safety purposes. Do not treat Claude as a secure vault.
Claude Projects	Projects allow uploading documents for use across conversations. Documents uploaded to a Project are stored on Anthropic servers for the Project's duration. Only Tier 1–2 documents should be added to Projects. Never upload Tier 3+ content.
Integrations and agents	Claude's browser/computer use features and third-party integrations may access data beyond what you type in the prompt. Review and restrict permissions before enabling. These features require AL AI Lead approval for use with Tier 2+ content.

> *AL's approved Claude configuration: Teams or Enterprise tier, training opt-out confirmed at admin level, no Project uploads above Tier 2, no browser or computer-use features enabled without AI Lead approval.*

ChatGPT — OpenAI

Setting / Consideration	Guidance for AL
Required tier	ChatGPT Plus (individual) or ChatGPT Team/Enterprise. The free tier is not approved for Tier 2 or above content.
Training opt-out — Plus	Go to Settings (bottom left) > Data Controls. Turn off "Improve the model for everyone." This applies to new conversations; it does not retroactively remove previously collected data.
Training opt-out — Team/Enterprise	Disabled by default. Enterprise contracts include Data Processing Agreements (DPAs) with additional contractual protections. Verify your organisation's contract includes a DPA.
Memory feature	ChatGPT's memory feature stores information across conversations and must be carefully managed. Go to Settings > Personalization > Memory to review and clear stored content. Ensure no Tier 3+ data has been retained in memory.
GPTs and custom apps	Third-party GPTs (custom AI apps built on ChatGPT) may have different data practices from OpenAI's core product. Do not use third-party GPTs with any AL content without vetting through the standard tool approval process.

File uploads	Files uploaded to ChatGPT are processed on OpenAI's servers and may be retained. Only upload Tier 1–2 documents using an account with training opt-out confirmed.

Google Gemini and Google Workspace AI Features

Setting / Consideration	Guidance for AL
Consumer Gemini (gemini.google.com)	The consumer version is linked to personal Google accounts and is not approved for AL content above Tier 1. Google may use consumer Gemini conversations to improve its products by default.
Gemini for Workspace	If AL uses Google Workspace, AI features are controlled by the organisation's admin account. The admin can enable or disable Gemini features for all users. Country Directors responsible for Workspace administration must review and configure AI feature settings.
Disabling training on Workspace data	In the Google Workspace Admin Console: go to Apps > Additional Google Services > Generative AI additional terms. Disable "Allow Google to use account data for AI improvements" for your organisation's domain. This controls whether Workspace content trains Google's AI models.
Gemini in Gmail, Docs, Meet, Drive	These embedded features may be enabled by default and can access the content of emails, documents, and meeting recordings. Confirm admin settings have restricted training data use before using these features with Tier 2+ content.
NotebookLM	Google's NotebookLM allows uploading documents for AI analysis. Only Tier 1–2 documents may be uploaded. Verify that Workspace admin settings are appropriately configured.

Microsoft Copilot

Setting / Consideration	Guidance for AL
Consumer vs. M365 Copilot	Consumer Copilot (copilot.microsoft.com with a personal account) is not approved for Tier 2+ content. Microsoft 365 Copilot (integrated into Word, Excel, Outlook, Teams) is covered by Microsoft's enterprise data protection commitments when used with an M365 business or enterprise account.
M365 Copilot data protection	Microsoft commits that M365 Copilot does not use organisational data to train foundation models, and that data stays within the organisation's M365 compliance boundary. Verify that your organisation has an appropriate M365 licensing agreement that includes these commitments.

Copilot in Teams and Outlook	Copilot can read emails, meetings, and Teams messages within the organisational account — representing broad potential access to Tier 2–3 content. Ensure your organisation has reviewed and configured M365 Copilot permissions to the minimum necessary scope.
Copilot Studio / custom agents	Custom AI agents built with Copilot Studio may access organisational data sources depending on configuration. These require AI Lead approval before deployment and use.

Zoom AI Companion

Setting / Consideration	Guidance for AL
Training on meeting data	By default, Zoom may use meeting audio, transcripts, and chat to train AI models. This must be disabled at the account admin level: go to Account Settings > AI Companion and disable "Allow Zoom to use account data for AI model training."
Enabling/disabling per meeting	Account admins can control whether AI Companion features are available and whether users can enable them for individual meetings. AL's account admin must configure these settings before AI features are used.
AI Companion in meetings with participants	If AI Companion is active in a meeting (generating summaries, transcripts, or smart chapters), participants must be notified. For any meeting involving program participants, AI Companion must be disabled entirely.

Quick Reference: Required Settings by Data Tier

Tier	Claude	ChatGPT	Google Workspace AI	M365 Copilot	Zoom AI
1PUBLIC	Any tier; free OK	Any tier; free OK	Consumer OK	Any tier	Any
2INTERNAL	Pro+ — training opt-out OFF	Plus+ — training opt-out OFF	Workspace — admin training opt-out OFF	M365 business account	Admin training opt-out OFF
3CONFIDENTIAL	Teams/Enterprise only	Team/Enterprise with DPA	Workspace Enterprise with confirmed DPA	M365 Enterprise; verify compliance settings	Not recommended; use secure transcription alternative
4PROTECTED	**NO**	**NO**	**NO**	**NO**	**NO**

> *Tool configurations and provider policies change frequently. This table reflects settings as of mid-2025. The AL AI Lead reviews and updates this guidance semi-annually. Before using any tool with Tier 2+ content, check the current Approved Tools List for the latest configuration requirements.*

— END OF DOCUMENT —

Questions and contributions: cheri@accountabilitylab.org

Acknowledgment and Signature Page

By signing below, I confirm that I have read and understood the Accountability Lab Network AI Policy, including the data classification system, the use classification framework, and the requirements for participant and source protection. I agree to comply with this policy in all use of AL's paid AI subscriptions and any other AI tools used in connection with AL's work.

This acknowledgment is required of every staff member, contractor, consultant, volunteer, or fellow who is granted access to an AL-paid AI subscription. Re-acknowledgment is required on an annual basis, and again whenever this policy is significantly revised, using the version noted below.

Policy Version / Effective Date Acknowledged	Reason for Acknowledgment
	☐ New staff onboarding ☐ Annual renewal ☐ Significant policy revision

Printed Name	Role / Country Program	Signature	Date