Responsible AI: Ethics and Governance for NGOs

Executive Summary

NGOs exist to advance justice, protect rights, and serve communities — often the most marginalized and vulnerable. This mission creates a special responsibility when adopting AI: the same tools that can amplify an NGO's impact can, if deployed carelessly, reproduce discrimination, violate privacy, erode trust, and cause real harm to the very communities the organization serves. Responsible AI adoption is not a technical afterthought — it is a mission-critical practice.

This guide provides NGO practitioners with a practical framework for ethical AI governance: the policies, processes, habits, and values that ensure AI use remains aligned with organizational mission, respects the dignity of beneficiaries, and remains accountable to communities and stakeholders. It covers the major ethical risks NGOs face when using AI, how to build an internal governance structure appropriate to organizational scale, how to communicate AI use to communities and donors, and how to stay current as the AI landscape continues to evolve rapidly.

The framing throughout is practical rather than theoretical. Ethical AI governance does not require a dedicated legal team or an AI ethics committee (though both help if you have them). It requires clear organizational commitments, simple but enforced policies, staff who are empowered to raise concerns, and a culture of ongoing reflection. Many of the principles in this guide are extensions of values your organization almost certainly already holds: transparency, accountability, inclusion, and "do no harm."

This guide draws on emerging best-practice frameworks from organizations including the Partnership on AI, the AI Now Institute, NetHope, the Digital Civil Society Lab at Stanford, and the principles published by Amnesty International and Oxfam. It is designed to be practical for organizations of any size, from a small local NGO to a large international federation.

Evidence Table

Key Finding	Strength	NGO Implications
AI systems can amplify existing societal biases, producing outcomes that disproportionately harm marginalized groups.	High (extensive research)	NGOs must actively audit AI use for disparate impacts, especially in work affecting individuals.
Communities most affected by AI decisions are rarely involved in designing those systems.	High (participatory design research)	Involve beneficiary communities in AI governance decisions, not just technical staff.
Transparency about AI use increases, not decreases, public trust in organizations.	Moderate (emerging evidence)	Proactive disclosure of AI use is a trust-building strategy, not a liability.

Key Finding	Strength	NGO Implications
Data minimization — collecting only the data you need — is the most effective privacy protection.	High (data protection best practice)	Review what data you collect and share with AI tools; collect less by default.
Governance structures make organizations more agile, not less — because clear rules enable faster decisions.	Moderate (organizational research)	Invest in governance up front; it saves time and crises later.
Staff who understand the ethical risks of AI are better at catching problems early.	High (practitioner evidence)	Ethics training is a practical investment, not just a values exercise.
AI governance that is only top-down fails; front-line staff need agency to raise concerns.	High (case evidence)	Create accessible, low-barrier channels for staff to report AI concerns.

Step-by-Step Framework

Step 1: Adopt a Set of Organizational AI Principles

Before writing policies, articulate the values that will guide all AI decisions in your organization. These principles should be brief, memorable, and connected to your existing mission and values.

A suggested framework for NGOs — adapt language to your own context:

1. **Human dignity**: We will only use AI in ways that respect the inherent dignity and rights of every person we serve. We will not use AI to surveil, manipulate, or harm individuals.
2. **Transparency**: We will be open about when and how we use AI, with our staff, beneficiaries, donors, and the public.
3. **Accountability**: A human being is always responsible for AI-assisted decisions. AI does not absolve us of accountability for outcomes.
4. **Inclusion**: We will work to ensure our AI use does not discriminate against or exclude any group, and we will proactively seek input from those most affected.
5. **Privacy**: We will collect and share only the minimum data necessary. We will protect sensitive information with the highest standards.
6. **Reversibility**: We will design AI processes so they can be easily stopped, revised, or overridden by humans when problems arise.

Adopt these principles through a formal board or leadership resolution. This signals organizational commitment and gives staff a clear reference point when difficult questions arise.

Step 2: Build a Tiered Risk Classification

Not all AI uses carry the same risk. A tiered risk classification allows your organization to apply appropriate scrutiny without creating bureaucratic overload for low-stakes applications.

Tier 1 – Low Risk: AI is used for internal productivity tasks with no direct impact on beneficiaries. Examples: drafting internal meeting notes, generating first drafts of internal reports, summarizing staff documents.
Governance: Team-level self-review; documented in agent registry; basic training required.

Tier 2 – Moderate Risk: AI is used in external communications, grant reporting, or program analysis — outputs may reach funders, media, or the public. Examples: donor communications, advocacy materials, media monitoring.
Governance: Designated reviewer approval required before publication; periodic audit of AI outputs; disclosure policy applies.

Tier 3 – High Risk: AI is used in processes that affect individual beneficiaries — determining who receives services, prioritizing cases, or processing sensitive personal data. Examples: eligibility screening, case management support, data analysis involving vulnerable populations.
Governance: Leadership approval required before deployment; full ethical impact assessment; legal review; beneficiary community consultation; ongoing monitoring and annual review.

Tier 4 – Prohibited: Uses that are incompatible with organizational values regardless of technical capability. Examples: facial recognition of beneficiaries without explicit consent, predictive risk scoring of individuals in criminal justice contexts, generating deceptive content, surveillance of staff or beneficiaries.

Define these tiers explicitly in your AI policy and require staff to classify any new AI use case before deployment.

Step 3: Conduct an Ethical Impact Assessment

For any Tier 3 use case — and recommended for Tier 2 — conduct a brief Ethical Impact Assessment (EIA) before deployment. An EIA is not a bureaucratic form; it is a structured conversation that surfaces risks before they become problems.

An NGO Ethical Impact Assessment covers:

1. **Purpose and necessity:** What problem does this AI use solve? Is AI the best solution, or would a non-AI approach be simpler, safer, and more appropriate?
2. **Affected populations:** Who will be affected by this AI system's outputs or decisions? Are any affected groups vulnerable, marginalized, or historically subject to discrimination?
3. **Data assessment:** What data does this system use? Is that data accurate, representative, and ethically obtained? Does it include personal information about beneficiaries?
4. **Bias and fairness:** Could this system produce different outcomes for different demographic groups? How will you check for and address disparate impacts?

5. **Privacy**: What personal data is collected, processed, or shared? Is this proportionate? Are data subjects informed and consenting?
6. **Transparency**: Will the people affected by this system know it is being used? If not, why not? Is that consistent with your principles?
7. **Accountability**: Who is responsible if this system causes harm? What is the remedy process?
8. **Exit strategy**: How will you stop using this system if it is not working or causes harm?

Document the EIA and keep it on file. Review it if the system's use case changes significantly.

Step 4: Establish Data Governance for AI

AI agents are only as trustworthy as the data they handle. NGOs often work with sensitive data — personal information about beneficiaries, case records, health data, migration status, survivor testimonies. This data requires special protection in the context of AI.

Key data governance rules for NGO AI use:

- **Never input sensitive personal data into public AI tools** without explicit consent, anonymization, and legal review. "Public AI tools" include ChatGPT, Claude, and similar consumer-facing services where your inputs may be used for model training.
- **Anonymize before analysis**: Where AI analysis of beneficiary data is necessary, anonymize or aggregate data before inputting. Remove names, locations, and other identifying information.
- **Use enterprise-grade tools for sensitive data**: Enterprise plans of AI tools typically include data privacy agreements that prevent inputs from being used for training. This is not a guarantee of perfect privacy, but it is a meaningful additional protection.
- **Know where your data goes**: Review the privacy policy and data processing agreement of every AI tool you use. Ask: Where is data stored? Who can access it? How long is it retained? Is it transferred internationally? Does this comply with applicable data protection law (GDPR, PIPEDA, etc.)?
- **Apply the principle of data minimization**: Collect, store, and share only the data that is strictly necessary for the task. AI makes it easy to process large datasets; that does not mean you should.

Designate a data steward — someone responsible for maintaining an inventory of what data is used with which AI tools, and for reviewing new uses.

Step 5: Create Staff Training and Accountability

An AI governance policy is only as effective as the staff who implement it. Build a practical training program that equips everyone who uses AI tools with the knowledge to use them responsibly.

A minimum training program covers:

- What AI agents are and how they work (the "black box" is not that mysterious)

- Your organization's AI principles and policy
- The tiered risk framework and how to classify a new use case
- Data privacy rules and what never to input into AI
- How to recognize and report a problem (hallucination, biased output, privacy risk)
- The human review process and why it matters

Training should be role-differentiated: frontline staff need the basics; managers need to be able to evaluate AI use in their teams; senior leaders need to understand governance and accountability.

Create a simple, accessible reporting channel — an email address, a form, or a designated conversation in your team communication platform — where staff can flag AI concerns without fear of judgment. Psychological safety is essential; staff who feel that raising concerns will be unwelcome will stay silent when it matters most.

Step 6: Communicate AI Use Externally

As AI becomes a normal part of organizational operations, proactive communication with external stakeholders — donors, beneficiaries, media, regulators — becomes important. Early, honest communication is a trust-building strategy.

With beneficiaries: If AI is used in any process that affects individuals (case management, service delivery, communications), people have a right to know. Develop simple, plain-language disclosure statements. Make opt-out options available where feasible. Engage beneficiary communities in feedback on AI use — their input is invaluable and demonstrates respect.

With donors: Many institutional donors now ask about AI governance in grant applications and reporting. Be prepared to describe your AI policies clearly. Frame AI adoption as a responsible stewardship story — AI helps stretch donor resources further, with robust safeguards in place.

With the public: Consider adding a brief "How We Use AI" page to your website. This does not need to be comprehensive — a clear, simple statement of your principles and main uses is enough. It demonstrates accountability and builds trust.

In communications: When content is substantially drafted by AI (with human review and editing), consider whether to disclose this. Practice in the sector is still evolving, but a simple note ("Prepared with AI assistance and human review") is increasingly common and appreciated.

Step 7: Build a Culture of Ongoing Reflection

The AI landscape is changing faster than any policy can keep up with. The most important governance asset is a culture of ongoing, honest reflection about AI use — one where new tools are examined critically, where problems are surfaced and addressed quickly, and where the mission remains the constant guide.

Practical habits to embed:

- **Quarterly AI review:** Dedicate one leadership meeting per quarter to reviewing AI use, emerging risks, and any problems that arose.
- **Annual AI policy review:** Update your AI policy annually, or whenever a significant new tool or use case is being considered.
- **Learning from the sector:** Stay connected to organizations doing serious AI governance work — TechSoup, NetHope, the Digital Civil Society Lab, the Partnership on AI. Share what you learn.
- **Mission test:** When a new AI opportunity appears, ask: "Does this make us more or less effective at our mission? Does it align with our values? Would we be comfortable if our beneficiaries knew about it?" If the answers are uncertain, slow down.

Tools & Templates

AI Policy Template for NGOs
A modular, adaptable AI policy document covering: Scope | AI Principles | Tiered Risk Classification | Prohibited Uses | Data Governance Rules | Staff Responsibilities | Review Process. Available at AGENTSFORGOOD.NGO.

Ethical Impact Assessment (EIA) Form
A one-to-two page structured form for Tier 2 and Tier 3 use cases, covering the eight dimensions described in Step 3. Google Docs format for easy collaboration.

Data Inventory for AI (Template)
A spreadsheet listing each AI tool your organization uses, with columns for: Tool Name | Use Case | Data Types Processed | Personal Data Involved (Y/N) | Enterprise Plan (Y/N) | Data Processing Agreement in Place (Y/N) | Review Date.

AI Disclosure Statement Templates
Ready-to-adapt templates for: (a) Beneficiary disclosure, (b) Donor disclosure, (c) Website "How We Use AI" page, (d) Content disclosure footnote.

Staff AI Training Modules
A set of short (20–30 minute) self-paced training modules aligned to roles: Module 1 (All Staff) – AI Basics and Our Policy; Module 2 (Program Staff) – Data Privacy and Sensitive Information; Module 3 (Leaders) – Governance, Risk, and Accountability. Available in PowerPoint and Google Slides formats.

Concern Reporting Template
A simple Google Form or email template for staff to report AI concerns: What tool? What happened? Who was potentially affected? What is the urgency? This normalizes concern-raising and creates a record.

"Should We Use AI Here?" Decision Tree
A one-page flowchart guiding staff through: Is this task suitable for AI? → What tier is it? →

Does it involve personal data? → Have we done an EIA? → Is a human reviewer assigned? → Proceed / Pause / Escalate.

Case Vignettes

Case Vignette 1: Building Trust Through Transparency — A Refugee Support Organization

A refugee support organization began using AI to help case workers draft service referral letters and interview summaries. Early on, a volunteer raised a concern: clients were not told their information was being processed using an AI tool. The concern was taken seriously by leadership, and the organization paused to assess.

The review revealed that the AI tool was a general consumer product (not enterprise), meaning that case inputs — which contained sensitive personal information including nationalities, legal status, and family details — were potentially being used for model training by the provider. This was a significant privacy risk for a population already vulnerable to surveillance and discrimination.

The organization took the following steps: they switched to an enterprise plan with explicit data processing protections, anonymized all inputs to remove personally identifying information, developed a simple disclosure statement translated into five languages that was given to all clients, and established a quarterly data governance review. They also consulted with their client advisory group — a body of current and former clients — who provided feedback on the disclosure statement and the use of AI generally.

The result: enhanced trust rather than eroded trust. Several client advisors said they appreciated being consulted and were comfortable with the AI use, provided the privacy safeguards remained in place. The organization published a brief case study on the experience, which became a reference for other NGOs in similar situations.

Case Vignette 2: When an AI Agent Gets It Wrong — Catching Bias in Practice

A social services NGO deployed an AI agent to help triage incoming service requests, flagging those with the highest urgency for immediate follow-up. The system worked well in testing, but after three months of live use, a program manager noticed a pattern: requests from non-native English speakers and from people using mobile interfaces (often lower-income users) were being consistently rated as lower urgency than comparable requests from native English speakers.

Investigation revealed two causes: (1) the training data used to develop the urgency model over-represented clear, formally structured requests; and (2) the AI tended to rate shorter, less formally written requests as lower priority regardless of their content. The result was a systematic bias disadvantaging users who were already marginalized — the opposite of the organization's intent.

The organization took immediate corrective action: the automated triage was suspended pending redesign, a manual review of the previous three months' triage decisions was conducted, and the affected individuals were re-contacted where possible. The redesigned system included explicit anti-bias testing as a pre-deployment requirement and increased the weight given to specific urgency keywords regardless of writing style.

Key lessons: (1) Bias audits are not one-time events — they must be ongoing. (2) Disparate impact can be invisible without deliberate monitoring. (3) Having a clear escalation and correction process — as this organization did — limits harm and restores trust. (4) The bias was caught because a front-line manager was paying attention; this underscores the value of empowered, vigilant staff.

Metrics & KPIs

Metric / KPI	What It Measures	How to Measure
% of AI use cases with completed EIA	Governance coverage	Track in agent registry
Staff AI training completion rate	Readiness and accountability	HR training system
Concern reports received and resolved	Culture and responsiveness	Concern reporting log
Data privacy incidents involving AI	Risk management	Security/privacy incident log
% of Tier 3 use cases with beneficiary consultation	Community inclusion	EIA documentation
Bias audit completion rate	Fairness assurance	Quarterly governance report
Policy review completion (annual)	Governance health	Policy version history
External disclosure statements in place	Transparency	Communications audit

Risks & Mitigations

Risk: AI governance perceived as bureaucratic overhead, causing staff to circumvent it.
Mitigation: Keep governance lightweight and proportionate — Tier 1 tools need minimal oversight; only Tier 3 requires significant process. Explain the "why" behind every requirement. Celebrate cases where governance prevented a real problem.

Risk: Governance frameworks becoming outdated as AI capabilities evolve rapidly.
Mitigation: Build in an annual policy review as a non-negotiable. Subscribe to updates from key organizations (TechSoup, NetHope, Partnership on AI). Appoint a staff member to track AI policy developments in the sector.

Risk: Communities most affected by AI systems having no voice in governance.
Mitigation: Explicitly include beneficiary or community representatives in AI governance reviews for any Tier 3 use case. If a formal advisory body does not exist, create lightweight consultation mechanisms (brief surveys, focus groups, advisory conversations).

Risk: Over-reliance on vendor promises about privacy and ethics.
Mitigation: Read data processing agreements yourself (or have legal counsel review). Ask vendors specific questions about data storage, training use, and sub-processors. Require contractual protections for sensitive data.

Risk: "Ethics washing" — adopting principles without substantive practice.
Mitigation: Ground ethics in concrete policies, documented decisions, and accountability mechanisms. Ask regularly: "Can we point to a specific case where our AI principles changed a decision?" If the answer is always "no," the principles are decorative.

Implementation Checklist

- [ ] Organizational AI principles adopted and communicated to all staff
- [ ] Tiered risk classification in place and integrated into AI workflow
- [ ] Prohibited uses list defined and shared with all staff
- [ ] Data inventory created for all current AI tool use
- [ ] Data processing agreements reviewed for all current tools
- [ ] EIA template in place and completed for all Tier 2–3 use cases
- [ ] Staff training program designed and launched
- [ ] Concern reporting channel established and publicized
- [ ] External disclosure statements drafted (beneficiary, donor, website)
- [ ] Annual AI policy review scheduled in organizational calendar
- [ ] Community/beneficiary consultation process defined for Tier 3 uses

Glossary

Algorithmic Bias: The tendency of AI systems to produce outcomes that systematically favor or disadvantage particular groups, often reflecting biases in training data or system design.

Disparate Impact: A legal and ethical concept referring to policies or practices that appear neutral but disproportionately harm members of a particular group. Applies to AI systems that produce unequal outcomes by demographic.

Data Processing Agreement (DPA): A legal contract between an organization and a vendor specifying how personal data will be handled, stored, and protected. Required under GDPR and other data protection regulations.

Data Minimization: The principle of collecting and processing only the minimum personal data necessary for a specific purpose. A core principle of privacy-by-design and most data protection laws.

Ethical Impact Assessment (EIA): A structured process for identifying and addressing the ethical risks of a proposed AI use before deployment. Similar to a privacy impact assessment but broader in scope.

GDPR (General Data Protection Regulation): The European Union's data protection regulation, widely considered the global standard for data privacy law. Applies to any organization processing data of EU residents.

Partnership on AI: A multi-stakeholder organization of AI developers, civil society groups, and academic institutions working on responsible AI practices. Publishes principles and resources relevant to NGOs.

Prohibited Uses: AI applications that an organization has determined are incompatible with its values and will never deploy, regardless of technical capability or potential efficiency gains.

Tiered Risk Classification: A governance system that categorizes AI uses by their potential for harm, applying proportionate scrutiny to higher-risk applications.

Transparency: In AI governance, the practice of being open with stakeholders — beneficiaries, donors, staff, the public — about when, how, and why AI is used.

References

1. Partnership on AI. *Responsible Practices for Synthetic Media*. Partnership on AI, 2023.
2. AI Now Institute. *Algorithmic Accountability Policy Toolkit*. AI Now Institute, New York University, 2023.
3. NetHope. *Responsible AI in the Humanitarian Sector: A Practical Framework*. NetHope, 2023.
4. Digital Civil Society Lab, Stanford PACS. *Ethical Use of AI in Civil Society*. Stanford PACS, 2023.
5. Oxfam. *Responsible Data for Civil Society*. Oxfam International, 2021.
6. Amnesty International. *Principles for the Responsible Use of AI in Human Rights Work*. Amnesty Tech, 2023.
7. European Commission. *Ethics Guidelines for Trustworthy AI*. High-Level Expert Group on AI, 2019.
8. Reisman, Dillon et al. *Algorithmic Impact Assessments: A Practical Framework*. AI Now Institute, 2018.
9. UNICEF. *Policy Guidance on AI for Children*. UNICEF, 2020.
10. TechSoup. *Data Privacy and AI: A Guide for Nonprofits*. TechSoup, 2024.
11. Information Commissioner's Office (ICO). *Explaining Decisions Made with AI*. ICO, UK, 2020.
12. Gebru, Timnit et al. *Datasheets for Datasets*. Communications of the ACM, 2021.