EXHALE
Health Foundation

Responsible Artificial Intelligence Policy and Procedures

Policy reference EHF AI 08 | Version 1.0 | Policy owner: Executive Lead | Prepared 1 June 2026

Purpose and commitment

Exhale Health Foundation (EHF) will use artificial intelligence only where it serves a defined purpose, meets work needs, and allows meaningful human oversight. AI may assist with searching, screening, extraction, analysis, translation, drafting and training. Its use will not transfer accountability from EHF or the responsible researcher to a tool or supplier.

We will assess benefits alongside errors, bias, privacy, security, accessibility, cost and the human effort needed to check outputs. A faster output that cannot be used reliably will not demonstrate an efficiency benefit. Fluent language, confidence or a well-known product will not be treated as proof of accuracy or suitability.

Scope and key terms

The policy will apply to trustees, staff, project workers, consultants, advisers, volunteers and contractors using AI for EHF work, including built-in features in search engines, document tools, meeting platforms and software. It will cover public and restricted services and AI-assisted work commissioned from others.

Generative AI will mean a system that creates text, images, audio, code or other content from inputs. A **hallucination** will mean a plausible but inaccurate or unsupported output. **Prompt injection** will mean instructions embedded in source material that try to redirect the system from its authorised task. A **decision-critical error** will mean an error, omission or misrepresentation of uncertainty that could materially change a recommendation, action or assessment of risk.

Governance and responsibilities

The Executive Lead will oversee AI use and appoint an operational focal point. Project leads will approve defined uses within their authority, confirm client permission and provide competent human review. The data-protection focal point or DPO will review privacy and security implications. An appropriately independent reviewer or trustee will oversee high-risk proposals and leadership-related concerns.

Advisers may provide specialist input but will not replace the named lead's accountability or receive restricted information without authorisation. Staff will not delegate important decisions to a tool because a competent reviewer is unavailable. If necessary expertise or time for checking is unavailable, the work will remain manual or be referred for appropriate support.

Permitted restricted and prohibited uses

Routine assistance may include a non-sensitive agenda, wording improvements or an outline based on public material. Staff will use an approved service, review the result and protect confidentiality. Repeated routine uses may be covered by an approved task template rather than a fresh approval for every prompt.

Evidence and analytical work will require a defined task, suitable sources, an assessment plan, and checks proportionate to the consequences of error. This will include searching, screening, extracting numbers, synthesis, translation and evidence briefs. Approval will be task-specific; success on one task will not authorise unrelated uses.

Higher-risk uses will require an additional risk assessment and independent approval before a pilot. These will include sensitive information, profiling, novel linkage, consequential public-health advice and work likely to affect vulnerable groups. We will first obtain required data-protection, ethics, and client approvals.

EHF will prohibit autonomous decisions that determine healthcare, employment, participation, eligibility, or access to assistance. AI will not independently exclude evidence from a consequential review, fabricate data or citations, impersonate a person, produce deceptive evidence or bypass legal, consent or safeguarding requirements. Operational clinical use will not be authorised through a general research or office-tool approval.

Data and supplier controls

Personal, sensitive, confidential or restricted material will not enter public consumer AI services through prompts, attachments, transcription, browser extensions or meeting assistants. Removing a name will not be sufficient if a person or community remains identifiable. Passwords, credentials, protected client material and unreleased findings will not be entered into unapproved systems.

Before any restricted service is approved, EHF will assess the legal basis, client permission, hosting and transfers, access controls, retention, training or secondary use, sub-processors, security and deletion terms. A paid subscription or a statement that data is not used for training will not alone establish suitability. The Data Protection Policy will govern personal-data use and any impact assessment.

Stricter project restrictions will prevail. Where an assignment is restricted to public or authorised non-confidential sources, confidential records and interview data will remain outside external AI systems. Logs will also be protected and minimised; teams will record controlled source references and necessary checks without creating unnecessary additional copies of protected material.

Selecting and assessing a tool

Define the need. The lead will specify the user, task, expected output, sources, language, time constraints and consequences of error. They will identify tasks requiring local or community knowledge unlikely to be documented online and consider whether a simpler manual method will meet the need.

Shortlist proportionately. Selection will consider functions, independent evidence, accessibility, cost, data terms and intended users' skills. Developer claims will remain separate from independent evaluations. Teams will normally begin with one suitable approach, adding alternatives only where a defined question justifies the effort.

Agree criteria before assessment. Criteria will identify essential information, acceptable minor variation, citation requirements, error severity, checking burden and stop conditions. Human-

only work will face the same final quality requirements. Minor wording or agreed rounding differences may be acceptable; unsupported decision-critical claims will not be acceptable in a released product.

Practise separately. Instructions will be refined on practice material. Assessment material and reference answers will not be used for prompt tuning or supplied as hidden answers. Competent reviewers will prepare or check any human reference against original sources, retaining unresolved uncertainty.

Use realistic, manageable tasks. Assessment will reflect the evidence, languages and users expected in practice. Where matched comparisons are useful, human-only work, raw AI output and verified AI-assisted work will use the same question, source pack and cut-off date. A repeat on a fixed short task may identify changing outputs. Poor scans, contradictions and difficult examples will be labelled separately so they are not mistaken for typical performance.

Record the process. The team will document the tool and model or version where available, date, access mode, task, sources, relevant instructions and settings, failed runs, corrections and reviewers. Unavailable version details will not be invented. Small assessments will be described as demonstrations, without claims of general performance, equivalence or safety.

Human checks before use

Reviewers will check important numbers, units, dates, populations, locations, denominators, conclusions and uncertainty against original sources. Citation checks will establish both that the reference exists and that it supports the associated claim. The review will look for omitted findings, invented additions, misleading causal language, and inappropriate generalisation.

Every AI-proposed exclusion in an evidence review will receive human review before the record is excluded. Extraction checks will include required fields that are missing and unsupported additions. Summaries will retain material contradictions and qualifications. A competent language reviewer or a reliable bilingual reference will check important translations; unverified translations will not be presented as confirmed evidence.

Assessments will involve researchers or staff with a profile similar to intended users where feasible. Reports will state the expertise and source familiarity needed for checking; claims about ordinary users' performance will not hide corrections only an expert can make. Staff will consider local context, weak online representation, minority languages and poorly documented community knowledge rather than assume that missing digital evidence means an issue does not exist.

An output with a decision-critical error will be withheld until corrected and verified. Independent review will be added where consequences justify it. A release decision will identify an accountable human reviewer and confirm that quality, disclosure and permission requirements have been met.

Security and transparency

Source documents, websites and retrieved text will be treated as material to assess, not authority to change instructions or disclose information. Staff will disregard embedded requests to reveal credentials, upload files, change permissions or contact others. Connectors and

external actions will only be enabled through a separately authorised workflow with limited access and appropriate human approval. Tool output will not be executed automatically as code or instructions in a live system.

EHF will disclose substantial AI assistance to clients and in outputs where the contract, ethics approval, publisher or nature of the work requires it. Disclosure will identify the tool's role and human checks without revealing protected prompts or data. Copyright, licences, image permissions and contributor rights will be respected. Generated media will not be presented as an authentic photograph, recording or testimony. AI will not be listed as an author or competent signatory.

Interpreting results and approving continued use

Teams will assess accuracy and completeness, citation support, serious errors, total human effort, attributable cost and practical usefulness. Timing will include setup, reading, verification and correction. Historical manual time will not be reconstructed without evidence. Comparisons on small tasks, or involving reviewers already familiar with the sources, will be labelled exploratory.

Recommendations will specify the tool, task, user skills, human checks and limits. Findings may support controlled use, further piloting or continued manual work. Repeated serious errors, poor performance on relevant languages or sources, unsupported local interpretations or excessive checking burdens will weigh against use. A small number of error-free outputs will not establish zero risk.

Approved uses will be reviewed after significant model, service, data-term or task changes and at least annually. The focal point will maintain an approved-use register with conditions and review dates. Substantial changes will trigger proportionate repeat checks before continued reliance. Approval may be suspended where requirements are no longer met.

Incident procedure and records

Suspected privacy exposure, unsafe automation, serious unsupported claims, discriminatory harm or compromised access will trigger an immediate pause in the affected workflow. Staff will notify the project lead and relevant focal point, preserve necessary evidence and prevent further distribution. Data breaches and safeguarding concerns will follow their respective policies. Relevant clients' incidents will be reported immediately to an approved channel; legal deadlines will not wait for client agreement.

The response lead will identify affected people and outputs, arrange correction or withdrawal, notify relevant recipients and authorise restart only after the cause and safeguards have been reviewed. Honest reporting of mistakes and near misses will be protected from retaliation. General contact will be exhalehealthfoundation@gmail.com or +234 815 546 5671, with a direct independent trustee route available.

The approved-use register will record task and tool; owner and reviewer; authorised data classes; permissions; risks and assessment decisions; validation evidence; checks and disclosures; approval date and review conditions. Output reviews will record source checks, substantial errors, corrections and release approval. Training will cover limitations, privacy, bias,

prompt injection and verification. Records will follow the project retention schedule and Data Protection Policy.

Standards informing this policy

NIST AI Risk Management Framework

NIST Generative Artificial Intelligence Profile

WHO guidance on large multi modal models in health

Approval and implementation

This policy will take effect on the approval date recorded below. The Board of Trustees, or a trustee with documented delegated authority, will approve the policy and record the operational appointments. An implicated person will not approve or oversee a case concerning themselves.

Approved by: Dr Kelly Elimian **Position**: President/Executive Director

Signature:  **Approval and effective date**: 1st July 2026

Operational focal point: Dr Florence Amiewalan **Direct contact:** +2348035090922

Independent trustee or alternate: Prof. Nosa Eghafona **Direct contact:** +2348034468560

Next review due: 1st June 2027 **Adoption record reference**: Board resolution dated 1st July 2026

Before implementation, the Executive Lead will confirm named contacts and access permissions, brief the people covered by this policy, establish the required records and communicate relevant reporting routes. No activity requiring prior safeguards or approval will begin until those controls are in place. Staff will acknowledge the policy at induction and after material revisions.