Ethical Guidance for the Use of AI in FCDO-Funded Research, Evaluation, and Evidence Synthesis

Lucas Sempé

2026-02-01

Table of contents

1 Purpose, Scope, and How to Use This Guidance

1.1 Purpose

This guidance sets out the ethical standards and practical requirements for using artificial intelligence (AI) tools in research, evaluation, and evidence synthesis funded by the Research Comissioning Centre (RCC) funded by the UK Foreign, Commonwealth and Development Office (FCDO). It is designed to ensure that the use of AI tools in RCC-funded projects upholds the highest ethical standards, protects research participants, maintains scientific integrity, and promotes responsible innovation. The guidance provides a comprehensive framework for assessing the appropriateness of AI tools, managing risks, and ensuring accountability throughout the research process.

AI tools — including large language models (LLMs), automated coding and classification systems, machine-assisted evidence synthesis, and predictive analytics — create new capabilities and new risks. Existing research ethics frameworks were developed before these tools became widely available and do not address AI-specific concerns such as algorithmic bias, data processing through third-party APIs, the opacity of model outputs, or the challenges of validating machine-generated analysis. This guidance aims to fill this gap.

Companion document. This guidance is accompanied by a shorter evidence brief, *Operationalising Ethical AI in Development Research* (Sempé, 2026), which presents the research evidence and analytical framework underpinning these standards. The brief establishes *why* ethical operationalisation is needed; this guidance specifies *how* to do it.

1.2 Scope

This guidance applies to:

- All research, evaluation, and evidence synthesis activities funded wholly or in part by RCC, including projects commissioned through open calls, partnerships, and technical assistance agreements
- All partners, suppliers, and subcontractors involved in such activities
- All uses of AI tools at any stage of the research process — from design through data collection, analysis, reporting, and follow-up

"AI tools" includes, but is not limited to: large language models (e.g., ChatGPT, Claude, Gemini), automated text classification and coding systems, machine learning models for prediction or causal inference, AI-assisted systematic review tools (e.g., automated screening, data extraction), natural language processing tools for survey analysis or translation, and AI-powered data collection or monitoring systems.

This guidance does not cover the use of AI as the *subject* of research (e.g., studies evaluating AI governance policies). It covers AI as a **tool** used in the conduct of research.

1.3 How to use this guidance

The guidance is structured around the **four stages of the project cycle**, mirroring the FCDO's existing ethics guidance:

1. **Commissioning, planning, and design** — before AI tools are deployed
2. **Data collection and analysis** — while AI tools are in use
3. **Reporting, dissemination, and use of evidence** — when communicating AI-assisted findings
4. **Monitoring, follow-up, and data management** — after the research is complete

At each stage, the guidance provides:

- **Standards** — requirements that must be met
- **Decision prompts** — questions to guide ethical assessment
- **Checklists** — verifiable actions to complete

Partners should work through each stage sequentially. Not all standards will apply to every project — the level of ethical scrutiny should be proportionate to the risk. Section 2 provides a risk classification to help determine the level of review required.

1.4 Normative foundations

This guidance draws on convergent international standards for the ethical governance of AI. A cross-national analysis of 2,216 AI policy documents across 193 countries (Sempé, 2026) confirmed that multiple independently developed frameworks agree on the same core ethical architecture:

- **The UNESCO Recommendation on the Ethics of Artificial Intelligence** (2021) — adopted by 193 member states — provides the overarching structure: 4 values, 10 principles, and 11 policy action areas (UNESCO, 2021)
- **The OECD AI Principles** (2019) — adopted by 46 countries — establish five principles for trustworthy AI and five policy recommendations (OECD, 2019)
- **The EU AI Act** (2024) — the first binding AI regulation — operationalises a risk-based approach with enforceable requirements (European Parliament and Council, 2024)
- **The AI ethics literature** — systematic mapping of 84+ AI guidelines confirms convergence on transparency, fairness, non-maleficence, responsibility, and privacy (Floridi et al., 2018; Jobin et al., 2019)

The cross-national analysis (Sempé, 2026) found that these frameworks all fail in the same way: **values are endorsed, principles are articulated, but operational requirements are not implemented.** Across 2,216 national AI policies, coverage rates decline systematically from the abstract to the concrete: 55% for values, 53% for principles, but just 41% for policy action areas — a 14-percentage-point gap between what countries declare and what they operationalise.

This gradient is not random. The most aspirational items — “peaceful, just and interconnected societies” (83%), “ethical governance and stewardship” (82%), “responsibility and accountability” (73%) — appear in the majority of policies. The most operationally demanding items are the least addressed: “proportionality and do no harm” (16%), “human oversight and determination” (36%), and “ethical impact assessment” (28%). Gender-specific protections appear in fewer than 10% of national AI policies worldwide. In short, the global policy landscape tells countries *what to believe* about AI but not *what to do* about it.

The pattern holds across income groups. High-income and developing countries alike cover values more readily than policy actions; the implementation gap is a structural feature of the global governance architecture, not a capacity deficit confined to lower-income contexts.

For research organisations, this matters directly. The same dynamic — endorsing principles, failing to operationalise them — can play out at institutional level. An organisation can commit to “responsible AI use” while lacking any protocol for validating AI-generated outputs, reviewing data governance, or reporting AI-specific limitations. This guidance is designed to close that gap for FCDO-funded research by translating convergent international norms into concrete, stage-by-stage requirements that research teams can follow and funders can verify.

2 Ethical Principles and Standards

This guidance adopts what we call a **UNESCO+ architecture** — the UNESCO Recommendation's three-tier structure (values → principles → policy actions) as the organising skeleton, enriched by two additional analytical traditions that address what UNESCO alone cannot.

The cross-national analysis of 2,216 AI policies (Sempé, 2026) employed three complementary lenses, each grounding a different aspect of this guidance:

1. **UNESCO's Recommendation on the Ethics of AI** (2021) provides the *normative structure* — 4 values, 10 principles, and 11 policy action areas. This is the most comprehensive international AI ethics framework, adopted by 193 member states. But Book 3 of the analysis found that even this framework fails at its own implementation: coverage rates drop from 55% (values) to 41% (policy actions), and entire domains — gender (9.6%), culture (9.1%), human rights as a foundational frame (22.9%) — are systematically neglected.
2. A systematic analysis of ethics governance depth across the same 2,216 policies — drawing on Jobin et al.'s (2019) mapping of 84 AI guidelines, Floridi et al.'s (2018) AI4People framework, the OECD AI Principles (2019), and the EU AI Act (2024) — identified five ethics governance dimensions (E1–E5) that capture what UNESCO's principles name but do not operationalise: ethical framework depth, rights protection specifics, governance mechanisms (ethics boards, impact assessments, auditing), operationalisation requirements, and inclusion and participation structures. These dimensions enrich the principles tier with the institutional architecture that UNESCO's principles lack.
3. **Implementation science** provides *capacity conditions.* Drawing on Mazmanian and Sabatier (1983), Lipsky (1980), Grindle (1996), and Fukuyama (2013), the analysis identified five implementation capacity dimensions (C1–C5) that determine whether any policy — including an AI ethics policy — can actually be executed: clarity and specificity of objectives, dedicated resources and budget, legal authority and enforcement mechanisms, accountability and monitoring systems, and cross-agency coherence and coordination. These conditions shape the policy actions tier: every standard in Section 3 is designed not only to reflect an ethical principle but to be *clear* (C1), *resourced* (C2), *enforceable* (C3), *monitorable* (C4), and *coordinated* (C5).

The resulting architecture integrates all three traditions:

Table 1: The UNESCO+ architecture

Tier	UNESCO provides	The "+" adds
Values	4 foundational values	Convergence with OECD, EU AI Act, Floridi; plus scientific integrity from the research ethics tradition
Principles	10 ethical principles	Institutional specificity from E1–E5 ethics governance dimensions (governance mechanisms, rights protection architecture, operationalisation requirements)

Tier	UNESCO provides	The "+" adds
Policy actions	11 thematic areas	Implementation capacity conditions (C1–C5) ensuring each standard is clear, resourced, enforceable, monitorable, and coordinated; organised by FCDO project cycle stage; calibrated by EU AI Act risk tiers

2.1 Beyond risk management

The standards in this guidance address risks that can be identified, classified, and mitigated — bias, privacy breaches, invalid outputs. But AI tools pose a deeper challenge that risk frameworks alone cannot capture: they raise a question about what kind of activity research actually is, and what role human intelligence plays in it that cannot be replicated by computation.

AI systems process data, identify statistical patterns, and generate text that is often indistinguishable from human writing. But they do not *understand* what they produce. Human intelligence — the kind that produces credible research — is fundamentally different. It is embodied: a researcher who has spent months in a field site, who has sat with respondents, who has felt the heat and heard the ambient noise of the interview, brings to analysis a contextual understanding that no dataset encodes. It is relational: insight emerges from dialogue with colleagues, from the give-and-take of peer review, from relationships of trust with participants and partners. It is interpretive: where AI returns a pattern, a researcher asks *what it means* — drawing on disciplinary knowledge, lived experience, moral sensibility, and the kind of contextual judgment that develops only through years of practice. And it is fallibly self-aware: a researcher can recognise that they do not know something, that their analysis may be wrong, that the evidence does not support the conclusion they hoped to reach — a form of intellectual honesty that has no analogue in systems optimised to produce confident outputs.

When organisations treat AI-generated outputs as functionally equivalent to human analysis — "the AI coded the transcripts," "the model identified the themes," "the system found no significant bias" — they are implicitly adopting a view of research in which what matters is the product, not the process. This is a category error with practical consequences. A research team that is valued only for what it produces can be progressively replaced by tools that produce faster. But if what makes research credible is the *quality of human engagement* with evidence — the interpretive effort, the contextual sensitivity, the willingness to be surprised by data — then no efficiency gain from automation compensates for the erosion of those capacities. The risk is not that AI will produce bad outputs (though it may). The risk is that organisations will come to evaluate researchers the way they evaluate machines: by throughput and consistency rather than by depth of understanding.

This distinction matters practically, not only philosophically. It is the foundation on which the rest of this guidance rests: the reason human oversight (P4) requires *genuine analytical engagement* rather than cursory approval, the reason validation (S2.2) demands that human reviewers actually re-engage with source material rather than spot-check for surface plausibility, and the reason training (P9) must cultivate — not merely presuppose — the interpretive skills that make human review meaningful.

The distinction also sharpens a specific operational concern: the deskilling paradox. Research is not an assembly line in which AI automates discrete tasks more efficiently. The acts of reading sources closely, struggling with ambiguous evidence, iterating through failed analytical approaches, and writing one's way toward clarity are not inefficiencies to be optimised away — they are the processes through which researchers develop judgment, build expertise, and produce genuinely new understanding. When these cognitive processes are routinely delegated to machines, the concern is not only that specific outputs may be invalid (a quality assurance problem addressed in Section 3) but that researchers may gradually lose the capacities that make their work credible and their judgment trustworthy. The labour economics literature calls this *deskilling*: the paradox in which a tool designed to augment human capability ends up eroding it.

The concern is already visible in educational contexts. UNESCO's *Guidance for Generative AI in Education and Research* (2023) documents how students who rely on AI for writing and analysis fail to develop the critical thinking skills that education is meant to cultivate — producing adequate outputs while bypassing the intellectual effort that gives those outputs meaning. The parallel for research organisations is direct: a team that uses AI to draft literature reviews, code qualitative data, and generate analytical summaries may deliver outputs on time while the researchers themselves become progressively less capable of performing these tasks independently.

This is not an argument against AI use. It is a reason to be deliberate about *how* AI is integrated into research workflows. The deskilling risk has practical implications throughout this guidance: AI literacy training (P9) must address not only how to use AI tools but how to preserve the skills AI threatens to atrophy; human validation (S2.2) must involve genuine analytical engagement rather than cursory sign-off; and research teams should periodically perform key analytical tasks without AI assistance, as a form of professional maintenance. The goal is augmentation that strengthens human judgment, not substitution that quietly hollows it out.

2.2 Values

Four core values underpin all AI use in FCDO-funded research. They are adapted from the UNESCO Recommendation's foundational values for research contexts, enriched by the convergent AI ethics literature. Two map directly to UNESCO values: V1 (human rights and dignity) and V2 (equity and inclusiveness). V4 (responsibility and accountability) elevates what UNESCO treats as a principle into a foundational value for research — reflecting the cross-national finding that accountability (C4) is the weakest implementation dimension globally (0.48/4.0). V3 (scientific integrity) is the UNESCO+ addition specific to research: not among UNESCO's four values, but the non-negotiable foundation for any organisation whose outputs must meet evidential standards. Together, these four values converge with norms identified independently across 84 AI guidelines (Jobin et al., 2019): Floridi et al. (2018)'s beneficence, autonomy, and justice; the OECD's human-centred values (2019); and the EU AI Act's fundamental rights grounding (2024).

V1 — Human rights and dignity. Research participants have a right to privacy, autonomy, informed consent, and protection from harm — including harm caused by algorithmic systems. AI tools must not be used in ways that violate these rights. *Convergence: UNESCO Value 1; Floridi et al. (2018) autonomy and non-maleficence; OECD human-centred values; EU AI Act fundamental rights; E2 (rights protection) in the ethics governance analysis.*

V2 — Equity and inclusiveness. AI-assisted research must not systematically disadvantage or marginalise populations it claims to serve. This includes ensuring that AI tools perform equitably across languages, cultural contexts, and population groups, and that the efficiency gains from automation do not come at the cost of participatory research practices. *Convergence: UNESCO Value 3 (diversity and inclusiveness); Floridi et al. (2018) justice; E5 (inclusion and participation); Gwagwa et al. (2020) on context-specific AI deployment.*

V3 — Scientific integrity. AI tools must support, not undermine, the transparency, reproducibility, and rigour that underpin credible research. Automated outputs require the same evidential standards as human-generated analysis. *UNESCO+ addition: not among UNESCO's four values, but foundational for credible research. Operationalised through E1 (ethical framework depth) and C1 (clarity and specificity of objectives).*

V4 — Responsibility and accountability. Named individuals must be responsible for AI-assisted research outputs. Automated processes do not diminish human accountability for research quality and ethics. *Convergence: UNESCO Principle 3 elevated to value status; Floridi et al. (2018) justice; OECD accountability; C4 (accountability and monitoring) — the dimension found weakest globally at 0.48/4.0.*

2.3 Principles

Ten operational principles translate these values into ethical commitments. These are UNESCO's ten principles, but the convergent literature confirms that each is independently endorsed across major frameworks — Jobin et al. (2019) found transparency, fairness, non-maleficence, responsibility, and privacy in at least 60% of 84 AI guidelines analysed. The five ethics governance dimensions from the cross-national analysis (E1–E5) provide the institutional architecture that gives these principles operational teeth:

- **E1 — Framework depth** shapes how thoroughly principles are articulated (P1, P7, P9).
- **E2 — Rights protection** enriches the specific protections that P2, P4, and P5 require — non-discrimination testing, human oversight architecture, data governance protocols.
- **E3 — Governance mechanisms** provides the institutional structures — ethics boards, impact assessments, auditing — that make P3 and P6 enforceable rather than aspirational.
- **E4 — Operationalisation** is the bridge between principles and the policy actions in Section 3: it captures whether requirements are specific, measurable, and monitored.
- **E5 — Inclusion and participation** grounds P10 in concrete participation structures — not just "consult stakeholders" but specified mechanisms for community engagement.

Each principle below states a requirement and indicates how it applies to research practice.

Table 2: Ethical principles for AI use in research

3 No	**4 Principle**	Requirement for research | ==
P2	Fairness and non-discrimination	Test AI outputs for differential performance | across languages, populations, and contexts. | Do not | rely on AI outputs | that show systematic | bias. |
P1	Transparency and explicability	Disclose all AI tools used. Document models, versions, prompts, parameters. Make methods | reproducible.
P3	Accountability	Designate a named person responsible for AI | ethics in each project. Establish oversight and | review processes. |
P4	Human oversight	All AI outputs affecting research conclusions | must be subject to meaningful human review. No AI output is final without human validation. |
P5	Privacy and data governance	Verify that AI tool providers' data handling complies with ethics requirements. Protect participant data processed through AI systems. |
P6	Safety and reliability	Validate AI outputs before they inform research | conclusions. Establish quality thresholds and failure protocols.

P7	Proportionality	Use AI tools only when appropriate for the research question and context. Do not default to | AI when established methods are adequate. |
P8	Sustainability	Consider the computational and environmental costs of AI tool use, particularly for large-scale | processing.
P9	Awareness and literacy	Ensure researchers using AI tools have practical training in capabilities, limitations, and ethical | implications.
P10	Participation and inclusiveness	Include affected communities in decisions about | AI-mediated research. Assess whether AI tools | risk marginalising local knowledge.

A note on anthropomorphization. The language used to describe AI in research practice — AI "assists," "reviews," "identifies," "decides" — can subtly reshape how researchers relate to these tools. When AI is described in agentive terms, there is a natural psychological tendency to treat it as a collaborator rather than an instrument: to defer to its outputs, to feel that "the AI found this" rather than "the algorithm returned this pattern," and to lower the critical scrutiny applied to results that appear to come from an intelligent source. This tendency is amplified by AI systems designed to produce fluent, confident, and contextually appropriate outputs — systems that, by design, present statistical compilations as though they were considered judgments. For research practice, the implication is straightforward: AI tools should be described and treated as tools — powerful, useful, and entirely lacking in understanding, judgment, or intent. Where project documentation, team communications, or research outputs use language that implies AI agency, this should be identified and corrected. Anthropomorphization is not merely a semantic issue; it is a pathway to reduced critical engagement, which is precisely the opposite of what P4 (human oversight) requires.

4.1 Risk classification

Not all uses of AI in research carry the same ethical risk. A study that uses an LLM to tidy a bibliography is fundamentally different from one that uses predictive analytics to identify vulnerable populations for targeting. Ethical scrutiny should be proportionate to the potential for harm — a principle enshrined in the UNESCO Recommendation (Principle 3: proportionality and do no harm) and operationalised most concretely in the EU AI Act's four-tier risk classification (European Parliament and Council, 2024; UNESCO, 2021).

The EU AI Act distinguishes unacceptable, high, limited, and minimal risk, with binding obligations that escalate accordingly. For research contexts, however, the commercial and regulatory categories do not map directly — research rarely involves "social scoring" or "real-time biometric identification," but it routinely involves activities that create serious risks through other pathways: processing sensitive data from conflict-affected populations, generating evidence that shapes policy and funding decisions, or applying automated analysis to contexts where errors could cause material harm to participants or communities.

This section adapts the risk-based approach for RCC-funded research. The classification is organised around **two dimensions**: (1) the *type of data and population* involved, and (2) the *consequentiality of the AI output* — how directly the AI-generated result influences research conclusions, policy recommendations, or decisions affecting human subjects. The higher the sensitivity along either dimension, the higher the risk classification.

4.1.1 How to classify risk

Assess the highest applicable level. If any single criterion in a higher tier is met, that tier applies — even if other aspects of the AI use fall into a lower category.

Table 3: Risk classification framework for AI use in research

<table>
<tr><th>Tier</th><th>Data &amp; population</th><th>Consequentiality of output</th><th>Required actions</th></tr>
<tr><td>High</td><td>Personal, sensitive, or identifiable data; vulnerable or conflict-affected populations; data from minors or without explicit AI-processing consent</td><td>AI output directly informs policy recommendations, funding decisions, or conclusions about human subjects; AI replaces (not merely assists) human analytical judgment</td><td>Full ethical review; written AI impact assessment (S1.2); enhanced human oversight; bias testing across all relevant subgroups; named ethics oversight person</td></tr>
<tr><td colspan="4">Medium| Non-personal but contextually| AI output contributes to | Documented review; standard | sensitive data (e.g., policy | findings but is subject to | human validation (10–15% | | documents from fragile | substantial human review; | subsample); audit trail; | | states); aggregated or | AI augments established | data governance check | | anonymised participant data | methods rather than | | | | replacing them | |</td></tr>
<tr><td>Low</td><td>Publicly available data; no participant data; no sensitive contexts</td><td>AI output is an intermediate input fully reviewed by humans before contributing to any finding</td><td>Documentation and transparency; standard procedures apply</td></tr>
</table>

4.1.2 Illustrative examples by tier

High risk — requires full ethical review, AI impact assessment, and enhanced oversight:

- AI tools processing personal or sensitive data (e.g., health records, GPS coordinates, interview transcripts from conflict-affected populations)
- AI-generated outputs that directly inform policy recommendations or funding decisions without intermediate human analytical steps
- AI tools used in research involving vulnerable populations (refugees, minors, persons with disabilities) where AI errors could lead to misidentification, exclusion, or stigmatisation
- Use of AI for predictive analytics affecting human subjects (e.g., targeting, risk scoring, needs assessment)
- AI-assisted analysis where errors could cause material harm — including to the credibility of evidence that informs programme design
- Automated translation or transcription of sensitive qualitative data (e.g., interviews about gender-based violence, political dissent) processed through third-party APIs

Medium risk — requires documented review and standard oversight:

- AI-assisted systematic review screening or data extraction, where human reviewers verify a subsample and resolve disagreements
- LLM-assisted qualitative coding or thematic analysis of non-sensitive documents (e.g., published programme evaluations, grey literature)
- AI-assisted survey instrument translation or cultural adaptation, with back-translation and expert review

- Automated text classification of policy documents, legislation, or programme reports
- AI-assisted meta-analysis or evidence gap mapping, where AI identifies and categorises studies but humans verify inclusion decisions

Low risk — requires documentation and transparency; standard procedures apply:

- AI-assisted literature searching or bibliography management
- AI tools for drafting, editing, or formatting research outputs (where human review is integral to the workflow)
- AI-assisted data visualisation or presentation design
- Computational tasks (e.g., statistical code generation, data cleaning scripts) with full human verification before execution
- AI-assisted project management (e.g., scheduling, task tracking) with no research content implications

4.1.3 Reclassification

Risk levels are not permanent. A task initially classified as low risk may require reclassification if circumstances change — for example, if a literature review using AI screening unexpectedly identifies studies involving vulnerable populations, or if a funder requests that AI-assisted findings directly inform a funding decision. Project leads should reassess risk classification at each project stage and document any changes.

5 AI in the Research Cycle

This section is the operational core of the guidance — UNESCO's Tier 3 (policy actions), reorganised by the FCDO project cycle, calibrated by risk, and designed to satisfy the implementation capacity conditions (C1–C5) that determine whether any policy can actually be executed.

Where Section 2 established *what* values and principles govern AI use in research, this section specifies *what to do, when, and to what standard.* The requirements are organised around the four stages of the research project cycle, following the structure established in the FCDO's existing *Ethical Guidance for Research, Evaluation and Monitoring Activities* (2019). This reflects a deliberate choice: rather than listing policy actions abstractly (as UNESCO does with 11 thematic areas like "data policy" or "education and research"), this guidance integrates them into the workflow that research teams already follow — so that ethical requirements are encountered at the moment they are actionable.

At each stage, three instruments translate principles into practice:

- **Standards** (numbered S1.1–S4.5) are the binding requirements. Each standard operationalises one or more of the values (V1–V4) and principles (P1–P10) from Section 2. Critically, each standard is also designed to satisfy the five implementation capacity conditions from the cross-national analysis (C1–C5): every requirement has *clear and specific objectives* (C1), identifies the *resources* needed for compliance (C2), specifies who has *authority and responsibility* (C3), includes *accountability and monitoring* mechanisms (C4), and is *coordinated* with related requirements across stages (C5). This is the practical import of the UNESCO+ architecture: standards that are ethically grounded *and* implementable. The level of obligation depends on the risk classification (Section 4.1): all standards apply to high-risk projects; a subset applies to medium- and low-risk projects as indicated.
- **Decision prompts** are the self-assessment questions a project lead should be able to answer affirmatively before proceeding to the next stage.
- **Checklists** are the verifiable actions, designed to be extractable into project management documents and completed as the research progresses.

Each stage addresses a different set of risks:

- **Stage 1 — Commissioning, planning, and design.** The decisions made before AI tools are deployed determine whether ethical risks can be managed at all. This is where appropriateness is assessed, impact assessments are conducted, data governance is reviewed, informed consent is evaluated, and AI methods are pre-registered. Most ethical failures in AI-assisted research are preventable at this stage.

- **Stage 2 — Data collection and analysis.** Once AI tools are in use, the central risks are opacity (what did the AI do?), validity (did it do it correctly?), and equity (did it perform equally across populations?). This stage establishes documentation, human validation, bias testing, audit trails, and failure protocols.

- **Stage 3 — Reporting, dissemination, and use of evidence.** AI-assisted findings carry specific disclosure obligations. This stage covers what must be reported about AI tool use, how validation results should be presented, and how claims should be calibrated to the quality of AI-assisted evidence.

- **Stage 4 — Monitoring, follow-up, and data management.** After the research is complete, AI-specific obligations continue: post-project review, incident reporting, data retention for AI-processed outputs, protocol updates, and knowledge sharing.

For projects classified as **low risk** (Section 4.1), the requirements are lighter: documentation and transparency at each stage, with standard human review. For **medium-** and **high-risk** projects, the requirements escalate — including formal impact assessments, enhanced validation protocols, bias testing, and named oversight responsibilities.

5.1 Stage 1: Commissioning, Planning, and Design

5.1.1 Standards

S1.1 — AI appropriateness assessment. Before adopting AI tools, assess whether they are appropriate for the research question, context, and population. Consider: Does the research question require the interpretive nuance that AI tools may lack? Are validated non-AI methods available? Does the population or context create specific risks (e.g., low-resource languages, sensitive topics)?

S1.2 — AI impact assessment. For medium- and high-risk projects (see Section 4.1), conduct a written AI impact assessment before deployment. The assessment should identify: what AI tools will be used and for what purpose; what data will be processed through AI systems; risks to participant privacy, data protection, and potential bias; mitigation measures for each identified risk; and how AI outputs will be validated.

S1.3 — Data governance review. Verify that the terms of service and data retention policies of all AI tool providers are compatible with: research ethics requirements (informed consent, confidentiality); applicable data protection regulations (UK GDPR, local data protection laws); and funder requirements for data handling. Where AI tools process data through third-party APIs, document what data is transmitted, where it is stored, how long it is retained, and whether it may be used to train future models. Additionally, assess vendor concentration risk: the AI tools most widely used in research are controlled by a small number of companies whose business models, data practices, and service terms can change without notice. Where a research project depends on a single commercial AI provider, evaluate what happens if that provider changes its terms, discontinues the service, or modifies the model mid-project — any of which can compromise reproducibility, data governance, or the viability of the analytical approach. Where feasible, identify alternative tools and avoid architectures that create single-vendor dependency for critical analytical steps.

S1.4 — Informed consent. Where AI tools will process data from or about human participants, assess whether existing consent protocols adequately cover AI-assisted processing. If participants consented to “analysis by the research team,” does that extend to processing by commercial AI systems? Where consent is inadequate, obtain additional consent or anonymise data before AI processing.

S1.5 — Pre-registration of AI methods. Register AI-assisted methods in the research protocol, pre-analysis plan, or terms of reference. Specify: which AI tools will be used; at which stages of the research process; what validation methods will be applied; what quality thresholds must be met; and what alternative methods will be used if AI tools fail quality checks.

S1.6 — Competency requirements. Ensure that research team members using AI tools have adequate training in: the capabilities and limitations of the specific tools; prompt engineering and parameter setting; output validation and bias detection; and the ethical requirements in this guidance. Training should also address the deskilling risk identified in Section 2.1: maintaining analytical and interpretive capabilities alongside AI use, including periodic practice of key tasks without AI assistance. Where capacity gaps exist, include training in the project plan and budget.

5.1.2 Decision prompts

Before proceeding to Stage 2, the project lead should be able to answer “yes” to each of the following:

☐ Have we assessed whether AI tools are appropriate for this research?
☐ Have we completed an AI impact assessment (for medium/high-risk projects)?
☐ Have we verified that all AI tool providers’ data policies meet our requirements?
☐ Have we reviewed informed consent protocols for adequacy?
☐ Have we documented AI methods in the research protocol?
☐ Does the team have adequate training to use AI tools ethically and effectively?

5.1.3 Commissioning checklist

Table 4: Stage 1 checklist

Item	Completed	Notes
AI appropriateness assessment completed		
Risk level classified (high / medium / low)		
AI impact assessment completed (if required)		
AI tool provider data policies reviewed		
Informed consent adequacy assessed		
AI methods registered in protocol / PAP / ToR		
Team competency assessed; training planned if needed		
AI ethics oversight responsibility assigned to named person		
Budget includes AI tool costs and validation activities		

5.2 Stage 2: Data Collection and Analysis

5.2.1 Standards

S2.1 — Documentation. Document all AI tool use contemporaneously. For each AI-assisted task, record: the tool name, model version, and access date; the specific prompts, instructions, or parameters used; any pre- or post-processing of inputs and outputs; and the validation method applied. Use the AI Documentation Template in Section 7.

S2.2 — Human validation. All AI-generated outputs that contribute to research findings must be validated by human review. Minimum requirements:

- **Quantitative threshold**: independently human-code a random subsample of at least 10–15% of AI-processed items. Report inter-rater agreement (e.g., Cohen's kappa, percentage agreement).
- **Qualitative review**: for AI-assisted qualitative analysis (coding, thematic analysis), a researcher must review all AI-generated codes and themes against original data. AI codes are hypotheses, not findings.
- **Spot-checking**: for large-scale processing where full subsample validation is impractical, implement systematic spot-checks with documented sampling strategy.

S2.3 — Bias testing. Test AI outputs for differential performance across relevant subgroups. At minimum: compare AI performance across languages used in the study; assess whether AI accuracy varies across population groups, regions, or data types; and document any systematic patterns of over- or under-performance.

S2.4 — Audit trails. Maintain reproducible records of all AI-assisted analytical steps. Audit trails should enable a third party to: identify which outputs were AI-generated versus human-generated; reproduce AI-assisted steps using the documented prompts and parameters; and verify that validation protocols were followed. Store audit trails alongside research data according to the project's data management plan.

S2.5 — Failure protocols. Apply pre-specified procedures when AI outputs fail quality checks. Before beginning AI-assisted analysis, define: the quality thresholds that AI outputs must meet (e.g., minimum agreement with human coding); the procedures to follow when thresholds are not met (e.g., re-prompting, additional human review, abandoning AI-assisted approach); and the criteria for concluding that AI tools are not suitable for the task. Do not adjust quality thresholds after observing results.

S2.6 — Data security during processing. When transmitting research data to AI tools via APIs or cloud services: do not transmit personally identifiable information unless the data governance review (S1.3) has confirmed this is permissible; use anonymised or pseudonymised data wherever possible; maintain a log of what data was transmitted, to which service, and when; and comply with any data localisation requirements applicable to the research context.

5.2.2 Decision prompts

During data collection and analysis, the project lead should regularly verify:

☐ Is all AI tool use being documented in real time?
☐ Have human validation subsamples been drawn and coded?
☐ Have we tested for differential AI performance across relevant subgroups?
☐ Are audit trails being maintained alongside research data?
☐ Have any AI quality thresholds been breached? If so, have failure protocols been followed?
☐ Is data being transmitted to AI tools in compliance with the data governance review?

5.2.3 Data and analysis checklist

Table 5: Stage 2 checklist

Item	Completed	Notes
AI documentation template completed for each AI-assisted task		
Human validation subsample drawn (min 10–15%)		
Inter-rater agreement calculated and documented		
Bias testing conducted across relevant subgroups		
Audit trail maintained and stored with research data		
Failure protocols applied where quality thresholds breached		
Data transmission log maintained		
No personally identifiable information transmitted without authorisation		

5.3 Stage 3: Reporting, Dissemination, and Use of Evidence

5.3.1 Standards

S3.1 — Disclosure. All research outputs must disclose AI tool use. At minimum, reports should state: which AI tools were used (name, version, provider); at which stages of the research process; for what purpose; and what validation was performed. Disclosure should appear in the methodology section of the report, not buried in footnotes or annexes.

S3.2 — Validation reporting. Report the results of human validation alongside AI-assisted findings. Include: the human-AI agreement rate (kappa, percentage agreement, or other appropriate metric); the size and sampling strategy of the validation subsample; any systematic patterns of AI error or disagreement; and how disagreements between human and AI coding were resolved. Validation results are part of the findings, not an optional appendix.

S3.3 — Limitations. Acknowledge the limitations of AI-assisted methods explicitly. Address: known limitations of the specific AI tools used (e.g., language coverage, training data biases); any contexts where AI performance was weaker; the implications of any validation failures for the reliability of findings; and the degree to which findings depend on AI-generated (versus human-generated) outputs. Do not present AI-assisted findings as equivalent to fully human-generated analysis without evidence of comparable quality.

S3.4 — Reproducibility. Make AI methods reproducible. Where possible: publish the prompts and parameters used; provide sufficient methodological detail for a third party to replicate the AI-assisted steps; and deposit audit trails and validation data in a data repository alongside the research data. Where proprietary tools or closed-source models are used, document version numbers and access dates so that readers understand the tool's state at the time of analysis.

S3.5 — Appropriate claims. Ensure that claims, conclusions, and policy recommendations are warranted by the evidence, including the quality of AI-assisted analysis. Where AI validation revealed limitations (e.g., lower accuracy for certain languages or populations), adjust conclusions accordingly. Do not generalise beyond the populations and contexts where AI tools performed adequately.

5.3.2 Decision prompts

Before finalising any research output, verify:

- □ Is AI tool use disclosed in the methodology?
- □ Are validation results reported alongside AI-assisted findings?
- □ Are AI-specific limitations acknowledged?
- □ Are prompts and parameters available for reproducibility?
- □ Are claims proportionate to the quality of AI-assisted evidence?

5.3.3 Reporting checklist

Table 6: Stage 3 checklist

Item	Completed	Notes
AI tool use disclosed (name, version, provider, purpose)		
Validation results reported (agreement rates, subsample details)		

Item	Completed	Notes
AI-specific limitations acknowledged in text		
Prompts and parameters published or deposited		
Audit trail deposited with research data		
Claims proportionate to AI-assisted evidence quality		

5.4 Stage 4: Monitoring, Follow-Up, and Data Management

5.4.1 Standards

S4.1 — Post-project review. Conduct a review of AI tool performance and ethical issues within three months of project completion. The review should assess: whether AI tools performed as expected; whether any ethical issues arose during the project; what lessons were learned about AI use in this research context; and what changes to protocols or procedures are recommended for future projects. Document the review and share findings with the commissioning team.

S4.2 — Incident reporting. Report any ethical incidents or unexpected issues related to AI tool use to the designated oversight body. Incidents include, but are not limited to: discovery of systematic AI bias that may have affected findings; data security breaches involving AI tool providers; participant complaints about AI-assisted data processing; failure of AI tools that required significant methodological changes; and AI outputs that, upon review, were found to be misleading or fabricated. Maintain an incident log for the project.

S4.3 — Data retention and disposal. Apply data management standards to AI-processed outputs. Specifically: retain AI documentation (prompts, parameters, validation records, audit trails) for the same duration as research data; where AI tools created intermediate outputs (e.g., coded datasets, extracted summaries), retain these as part of the research record; ensure that data transmitted to AI tool providers is handled according to the agreed retention and disposal terms; and where AI providers retain data for model training, document this and assess implications for participant confidentiality.

S4.4 — Protocol updates. Update organisational AI ethics protocols based on lessons learned. AI tools evolve rapidly; protocols established in one year may be inadequate the next. At minimum, review organisational AI ethics protocols annually, incorporating: lessons from post-project reviews; changes in AI tool capabilities, policies, or risks; developments in ethical standards or regulatory requirements; and feedback from researchers, participants, and partners.

S4.5 — Knowledge sharing. Share methodological insights from AI-assisted research with the broader research community, especially FCDO-funded partners. This includes: publishing methodological notes on AI tool performance in specific research contexts; contributing to communities of practice on ethical AI use in development research; and sharing validation data that can help other researchers assess AI tool suitability.

5.4.2 Decision prompts

After project completion, verify:

☐ Has a post-project review of AI tool use been conducted?
☐ Have any ethical incidents been reported and logged?
☐ Are AI documentation and audit trails retained alongside research data?
☐ Have organisational protocols been updated based on lessons learned?
☐ Have methodological insights been shared with the research community?

5.4.3 Monitoring checklist

Table 7: Stage 4 checklist

Item	Completed	Notes
Post-project review conducted (within 3 months)		
Incident log completed (even if no incidents)		

Item	Completed	Notes
AI documentation retained per data management plan		
Data transmitted to AI providers handled per agreed terms		
Organisational protocols reviewed and updated		
Methodological lessons shared with partners / community		

6 Decision Tool: Should I Use AI for This Research Task?

Use this decision tree when considering whether to use AI tools for a specific research task.

Step 1 — Is there a validated non-AI method that would achieve the same purpose?

- *Yes, and it is feasible within the project's time and budget* → Use the established method. AI is not required.
- *Yes, but it is not feasible at the required scale* → Proceed to Step 2.
- *No* → Proceed to Step 2.

Step 2 — Does the task involve personal or sensitive data?

- *Yes* → Classify as high risk (Section 4.1). Complete a full AI impact assessment (S1.2). Verify data governance (S1.3). Review informed consent (S1.4). Proceed to Step 3 only if all requirements are met.
- *No* → Proceed to Step 3.

Step 3 — Can the AI output be validated against human judgment?

- *Yes* → Define validation protocol (subsample size, agreement thresholds, failure criteria) before deploying the AI tool. Proceed to Step 4.
- *No* → Do not use AI for this task. If AI outputs cannot be validated, they cannot be relied upon for research conclusions.

Step 4 — Does the team have the competency to use, validate, and document the AI tool?

- *Yes* → Proceed. Document AI methods in the research protocol (S1.5). Follow Stage 2 standards during implementation.
- *No, but training is available* → Include training in the project plan before deploying AI tools.
- *No, and training is not feasible* → Do not use AI for this task.

Step 5 — Is the AI tool's data handling compatible with ethics requirements?

- *Yes* → Proceed with implementation.
- *No* → Explore alternative AI tools with compliant data handling. If none are available, do not use AI for this task.

7 Annex A: AI Documentation Template

Complete this template for each AI-assisted task in a research project. Retain alongside the research data and audit trail.

7.1 Project information

Field	Entry
Project title	
Project reference / ID	
Responsible researcher	
Date	

7.2 AI tool details

Field	Entry
AI tool name	
Provider	
Model version / identifier	
Access method (API / web interface / local)	
Date(s) of use	

7.3 Task description

Field	Entry
Research task the AI tool was used for	
Stage of research (design / data collection / analysis / reporting)	
Risk classification (high / medium / low)	
Justification for using AI	

7.4 Inputs

Field	Entry
Description of data / text input to AI tool	
Did inputs contain personal or sensitive data?	Yes / No
If yes, was data governance review completed? (ref S1.3)	Yes / No / N/A
Prompt(s) or instruction(s) used (attach separately if lengthy)	
Parameters or settings applied	
Any pre-processing of inputs before submission to AI tool	

7.5 Outputs and validation

Field	Entry
Description of AI output	
Human validation method	
Subsample size (% of total)	
Human-AI agreement rate	
Bias testing performed?	Yes / No
Any systematic patterns of error identified?	
Were quality thresholds met?	Yes / No
If no, what failure protocol was applied?	
Post-processing or editing of AI outputs	

7.6 Data handling

Field	Entry
Was data transmitted via API / cloud service?	Yes / No
Provider's data retention policy	
Data transmission log maintained?	Yes / No

8 Annex B: AI Validation Reporting Template

Include this information in the methodology section of any research output that used AI-assisted methods.

8.1 Minimum reporting requirements

1. **AI tool identification**: Name, version, provider, and access date(s) for each AI tool used.
2. **Purpose**: What the AI tool was used for and at which stage(s) of the research process.
3. **Validation design**: How AI outputs were validated (subsample size, sampling strategy, validation method).
4. **Agreement metrics**: Human-AI agreement rate, reported using an appropriate metric:
   - For categorical data: Cohen's kappa, percentage agreement, or Fleiss' kappa (for multiple raters)
   - For continuous data: ICC, Pearson/Spearman correlation
   - Report both overall agreement and subgroup-level agreement where relevant
5. **Bias assessment**: Results of testing for differential AI performance across languages, populations, regions, or data types.
6. **Error patterns**: Any systematic patterns of AI over- or under-performance (e.g., consistent misclassification of certain categories, lower performance in specific languages).
7. **Failure handling**: Whether any quality thresholds were breached and what actions were taken.
8. **Limitations**: Specific limitations of the AI-assisted methods for this research context.

8.2 Example disclosure paragraph

> This study used [AI tool name, version] to [purpose, e.g., screen 12,000 abstracts for inclusion in the systematic review]. AI-screened results were validated against independent human screening of a random subsample of [N] abstracts ([X]% of total). Human-AI agreement was [metric] = [value], indicating [interpretation]. AI performance was comparable across [languages/subgroups tested], with [any exceptions noted]. [N] abstracts where human and AI screening disagreed were resolved by [method]. Based on validation results, we estimate that AI-assisted screening [correctly identified / may have missed] approximately [X]% of relevant studies.

9 Annex C: Glossary

AI (Artificial Intelligence) — Computer systems that perform tasks typically requiring human intelligence, including learning, reasoning, and pattern recognition. In this guidance, AI refers to tools used in research, not AI as a research subject.

API (Application Programming Interface) — A software interface that allows programmes to communicate. Many AI tools are accessed through APIs, meaning research data may be transmitted to external servers for processing.

Algorithmic bias — Systematic and repeatable errors in AI outputs that produce unfair outcomes for particular groups. In research, this may manifest as differential accuracy across languages, populations, or data types.

Audit trail — A chronological record of AI-assisted research activities, including inputs, prompts, parameters, outputs, and validation results. Audit trails enable third-party verification and reproducibility.

Human-in-the-loop — A design principle requiring meaningful human review and decision-making at critical points in an AI-assisted process. Distinguished from rubber-stamping, where human review is nominal.

Inter-rater reliability — A measure of agreement between different raters (human or AI). Common metrics include Cohen's kappa (two raters), Fleiss' kappa (multiple raters), and ICC (continuous data).

Large Language Model (LLM) — An AI system trained on large text corpora to generate, analyse, and transform text. Examples include GPT-4, Claude, and Gemini. LLMs are the AI tools most commonly used in current research practice.

Prompt — The instruction or query provided to an AI tool. Prompts significantly affect AI outputs; documenting them is essential for reproducibility.

Validation — The process of assessing whether AI outputs meet quality standards, typically by comparing AI outputs against human judgment on a subsample of cases.

10 Annex D: Relationship to Existing Frameworks

This guidance is designed to work alongside, not replace, the following frameworks:

FCDO Ethical Guidance for Research, Evaluation and Monitoring Activities (2019). The foundational ethics guidance for FCDO-funded research. This AI guidance extends the 2019 framework to address AI-specific risks not covered in the original document. Where the 2019 guidance applies, it remains in force; this guidance adds AI-specific standards at each project cycle stage.

The UNESCO Recommendation on the Ethics of AI (2021). This guidance adopts UNESCO's three-tier structure (values, principles, policy actions) as its normative architecture. The ten principles in Table 2 map directly to UNESCO's framework, supplemented by convergent standards from the OECD and EU AI Act.

The OECD AI Principles (2019). The OECD principles for trustworthy AI — inclusive growth, human-centred values, transparency, robustness, and accountability — are reflected in principles P1 through P4 and P6 of this guidance.

The EU AI Act (2024). The risk classification in Section 4.1 adapts the EU AI Act's risk-based approach for research contexts. While the EU AI Act applies primarily to AI providers and deployers in commercial and governmental settings, its risk tiers provide a useful framework for calibrating ethical scrutiny in research.

UK GDPR and the Data Protection Act 2018. Data governance requirements in this guidance (S1.3, S2.6, S4.3) supplement, but do not replace, existing data protection obligations. Where AI tools process personal data, researchers must comply with UK GDPR requirements for lawful processing, data minimisation, storage limitation, and data subject rights.

References

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI act).*

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4Peopleâ an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*, 689–707.

Fukuyama, F. (2013). What is governance? *Governance*, *26*(3), 347–368.

Grindle, M. S. (1996). *Challenging the state: Crisis and innovation in Latin America and Africa.* Cambridge University Press.

Gwagwa, A., Kraemer-Mbula, E., Rizk, N., & Rutenberg, I. (2020). Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions. *The African Journal of Information and Communication*, *26*, 1–28. https://doi.org/10.23962/10539/30361

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Lipsky, M. (1980). *Street-level bureaucracy: Dilemmas of the individual in public services.* Russell Sage Foundation.

Mazmanian, D. A., & Sabatier, P. A. (1983). *Implementation and public policy.* Scott Foresman.

OECD. (2019). *OECD principles on artificial intelligence.*

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.*