EvalCommunity

FIELD MANUAL

Responsible AI for Monitoring and Evaluation: Manual

A practical guide for monitoring, evaluation, research, learning, and accountability professionals who need to use AI responsibly without weakening evidence, ethics, or judgement.

By EvalCommunity

The Ecosystem for Evaluation, AI, and Development Professionals

https://www.evalcommunity.com/

Evidence | Ethics | Context | Governance | Learning

How to Use This Manual

This manual is a practical guide for Monitoring and Evaluation professionals who want to use artificial intelligence without weakening evidence quality, ethics, participation, transparency, or professional judgement. It is designed to be used before, during, and after AI-supported work.

> **The central rule**
> AI may assist the work, but people remain accountable for evidence, interpretation, recommendations, ethics, and final judgement.

What this manual helps you do

- Decide which AI uses are low-risk, which require safeguards, and which should not proceed without approval.
- Protect sensitive data, participant voice, local context, and equity considerations.
- Review AI-supported summaries, codes, drafts, findings, and recommendations before they influence decisions.
- Create documentation that is proportionate to the risk level of the task.
- Explain AI use to clients, funders, stakeholders, and team members in clear professional language.

Contents

1

Core commitments

Responsible AI in Monitoring and Evaluation begins with professional discipline, not technology adoption.

Responsible AI in Monitoring and Evaluation is not about using the newest tool. It is about protecting trust, evidence quality, context, and human accountability while using technology to improve the efficiency and clarity of selected tasks.

The six commitments below should be treated as project operating rules. They apply to evaluators, commissioners, research assistants, analysts, report writers, and technical advisers.

Commitment	Practical meaning
Human accountability	AI may support a task, but responsibility for analysis, interpretation, recommendations, and final outputs remains with qualified people.
Evidence connection	Every important claim should be traceable to source material. A convincing AI-generated paragraph is not evidence.
Proportional safeguards	The more a task affects meaning, judgement, people, resources, or decisions, the stronger the review process should be.
Data care	Personal, confidential, political, commercial, or community-sensitive information requires explicit permission and appropriate protections.
Context awareness	AI outputs should be checked for missing cultural, political, historical, organisational, and local context.
Transparent records	Teams should be able to explain what AI was used for, what was checked, what changed, and who approved the final work.

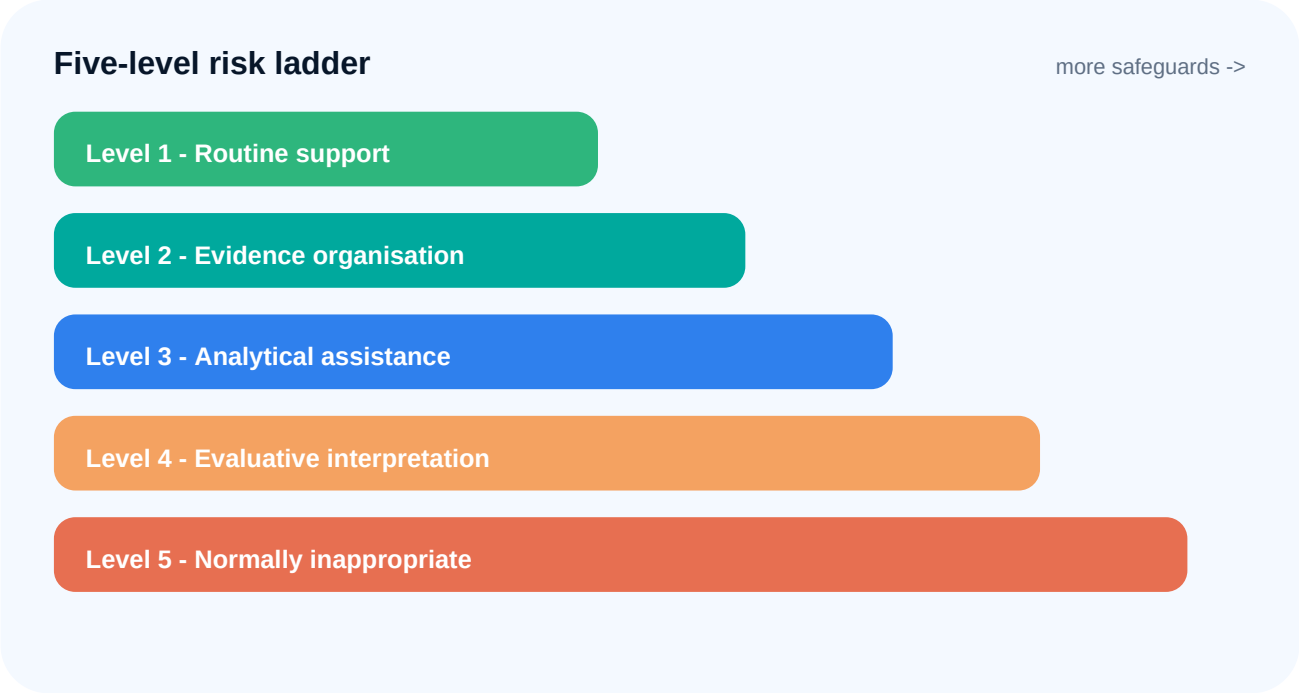
Field rule
Do not let fluency substitute for validity. A well-written output is not automatically an accurate, ethical, or useful evaluation output.

2

The five-level AI risk ladder

A practical framework for deciding what is allowed, what needs review, and what should be avoided.

The risk ladder helps teams classify AI-supported work before it begins. The question is not only what tool is being used, but what task the tool is being asked to perform, what data it is exposed to, and how much influence the output may have on findings or decisions.



Level	Use type	Examples	Minimum safeguard
Level 1	Routine support	Proofreading, formatting, plain-language editing, agenda drafting.	Human review before use.
Level 2	Evidence organisation	Summarising public documents, structuring notes, creating synthesis tables.	Compare with sources and keep a light record.
Level 3	Analytical assistance	Suggesting codes, themes, patterns, comparisons, or categories.	Document prompts, verify outputs, and conduct bias review.
Level 4	Evaluative interpretation	Drafting findings, conclusions, recommendations, or causal explanations.	Senior review, evidence mapping, and formal approval.
Level 5	Normally inappropriate	Replacing professional judgement, simulating stakeholder voice, using sensitive data without approval.	Avoid by default. Proceed only with exceptional approval.

How to use the ladder

- Classify the task before using an AI tool. Do not classify it after the output already looks useful.
- Escalate when the data becomes more sensitive, the audience becomes more consequential, or the output begins shaping analysis.
- Treat mixed tasks at the highest relevant risk level. For example, a summary that includes interpretation is not merely a summary.
- Document why a task was classified at a certain level and who approved that classification for moderate or high-risk uses.

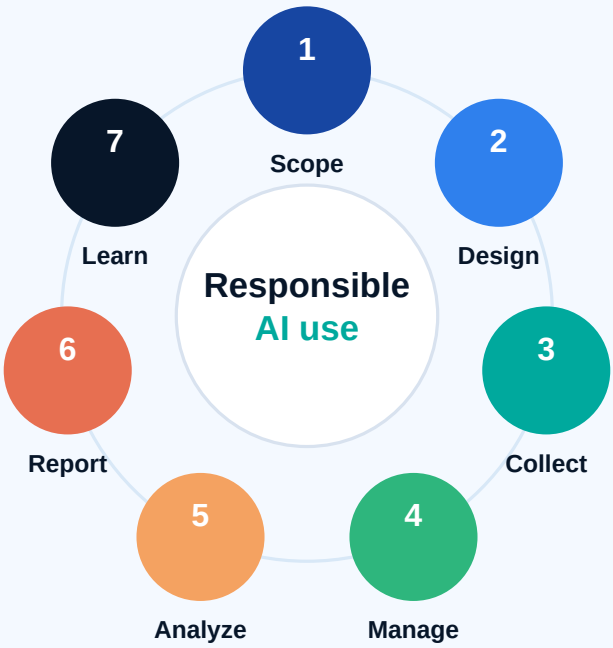
Risk classification practice

Scenario	Situation	Classification response
Scenario A	AI rewrites a public executive summary after the evaluator has approved the findings.	Likely Level 1. Check that meaning has not changed.
Scenario B	AI suggests themes from anonymised focus group notes.	Likely Level 3. Review excerpts, preserve dissenting views, and document prompts.
Scenario C	AI drafts final recommendations from interview notes in a sensitive governance evaluation.	Likely Level 4 or 5. Escalate before use and protect all sensitive data.

3

Responsible AI across the M&E cycle

AI can appear at every stage, but each stage has different opportunities and risks.



Stage	Useful support	Risk to manage
Scoping	Drafting workplans, clarifying learning questions, preparing meeting notes.	AI may shape the evaluation frame before stakeholders have agreed it.
Design	Drafting evaluation questions, logic models, indicators, rubrics, tools.	Generic designs may overlook context, power, feasibility, and local meaning.
Data collection	Improving instruments, translating drafts, preparing interviewer guidance.	Poorly checked AI wording may introduce leading questions or inaccessible language.
Data management	Labelling files, organising excerpts, structuring evidence tables.	Confidentiality, consent, and data minimisation rules must be respected.
Analysis	Suggesting codes, themes, patterns, contradictions, or summaries.	AI may flatten minority views or overstate patterns that are weakly supported.
Reporting	Drafting narrative sections, summaries, visual descriptions, executive briefs.	Fluent text may hide unsupported claims, missing caveats, or unjustified recommendations.
Learning	Preparing workshop prompts, learning questions, and discussion aids.	AI should support facilitation, not replace stakeholder interpretation or validation.

Practical rule
The earlier AI is used in the cycle, the more carefully teams should check whether it is shaping the frame, assumptions, and questions of the evaluation.

Stage-by-stage guardrails

- Before scoping: agree what AI can and cannot be used for in the project.
- Before data collection: confirm that tools, consent language, and translations have been human reviewed.
- Before analysis: decide which data can be used, which must be excluded, and who will verify outputs.
- Before reporting: map findings and recommendations to evidence and document any AI-supported drafting.
- Before dissemination: prepare disclosure language that is clear, honest, and proportionate.

4

Quality assurance and verification

AI-supported work should be reviewed before it shapes evidence, analysis, or decisions.

Quality assurance is the bridge between useful AI support and responsible professional practice. The purpose of review is not to prove that the tool is wrong. It is to confirm whether the output is sufficiently accurate, grounded, balanced, and ethical for the purpose for which it will be used.

Core verification questions

- Is the output supported by the original source material?
- What has been omitted, simplified, overstated, or generalised?
- Are minority, dissenting, unexpected, or politically sensitive views still visible?
- Are causal claims justified by the design, data, and evidence base?
- Could the evaluator explain and defend the final judgement without relying on AI wording?
- Does the output preserve uncertainty, limitation, and context?

Method	What it means	Use when
Source comparison	Compare the AI output against transcripts, documents, notes, datasets, or validated analysis files.	All Level 2-4 uses.
Evidence-to-claim mapping	Link each finding, conclusion, and recommendation to specific supporting evidence.	Reports, briefs, and decision-facing products.
Bias and voice review	Check for erased minority perspectives, stereotypes, cultural assumptions, and overconfident language.	Work involving communities, vulnerability, power, or equity.
Peer challenge	Ask a second person to review the AI-assisted output and the evidence supporting it.	Level 3 and Level 4 uses.
Stakeholder validation	Use appropriate participatory review to test meaning, interpretation, and practical relevance.	Sensitive findings or recommendations.

Red flag
If the team cannot explain how an AI-assisted statement was checked, the statement should not be used as a finding, conclusion, or recommendation.

5

Ethics, equity, and protection of voice

Responsible AI practice must protect people, relationships, context, and power.

Monitoring and Evaluation work often involves people describing services, rights, needs, harms, exclusions, institutions, finances, conflict, health, education, livelihoods, or lived experience. These are not just data points. They are human accounts that require care.

Sensitive information decision guide

Data type	Examples	Guidance
Public information	Published reports, public websites, open policy documents.	Usually suitable for low-risk AI support, with source checks.
Internal information	Project documents, draft reports, operational notes.	Use only if organisational rules allow it; avoid external tools when restricted.
Primary data	Interview notes, survey open text, focus group material.	Use only with explicit permission, minimisation, and review safeguards.
Identifiable or confidential data	Names, locations, cases, protected attributes, sensitive quotations.	Do not enter into AI tools unless explicitly approved and protected.
High-risk contextual data	Political, conflict, protection, safeguarding, whistleblowing, or vulnerability-related data.	Escalate to ethics, data protection, or senior review before any AI use.

Equity and voice checks

- Check whether the output reflects only the dominant or most frequent perspective.
- Look for groups whose views may be under-represented because of sampling, language, access, status, or power.
- Do not use AI to invent participant perspectives, quotations, lived experience, or community priorities.
- Preserve uncertainty and disagreement. Not every pattern needs to become a clean theme.
- Ask whether the people described in the analysis would recognise the interpretation as fair and respectful.

6 Governance, roles, and escalation

Responsible AI is easier when rules are agreed before the work becomes urgent.

Governance does not need to be heavy. It needs to be clear. A small project may only need a simple AI use register and a reviewer. A complex or sensitive evaluation may need formal approval, data protection review, stakeholder validation, and senior sign-off.

Role	Responsibility
Practitioner	Classifies the task, uses the tool within agreed boundaries, checks the output, and keeps records.
Project lead	Approves risk level, review method, documentation expectations, and disclosure approach.
Reviewer	Challenges evidence links, bias, context, interpretation, and recommendations.
Data or ethics lead	Advises on sensitive information, consent, confidentiality, and protection concerns.
Client or commissioner	Confirms expectations for disclosure, allowed uses, deliverable quality, and accountability.

Escalation triggers

- The task involves participant-level, identifiable, confidential, or politically sensitive data.
- The output may influence findings, conclusions, recommendations, funding, policy, or public claims.
- The AI output conflicts with evidence or produces persuasive but unsupported statements.
- Stakeholders could be harmed by misinterpretation, exposure, stereotyping, or loss of voice.
- The team cannot explain what was entered, what was produced, how it was checked, and who approved it.

Governance test
If a client, participant, reviewer, or colleague asked how AI was used, the team should be able to answer clearly and confidently.

7

Field tools and practical instruments

Ready-to-use instruments that turn principles into daily project practice.

These tools can be copied into a workplan, inception report, internal quality assurance process, or project folder. They are designed to be practical rather than bureaucratic.

Tool	Purpose	Best used for
AI use register	Records tool, task, data type, risk level, review method, disclosure decision, and responsible person.	Every project using AI.
Evidence-to-claim map	Links each important finding, conclusion, and recommendation to supporting evidence and review notes.	Reports and decision products.
Sensitive data decision tool	Classifies data and determines whether AI use is allowed, conditional, or prohibited.	Before entering any project data.
Professional judgement check	Confirms that a qualified person can defend the output independently of the AI wording.	Analysis and recommendations.
Equity and voice review	Checks whether small groups, dissenting views, and marginalised perspectives have been preserved.	Equity-focused or participatory evaluations.
Disclosure builder	Creates simple language explaining how AI was used and how human review was maintained.	Reports, proposals, and client communication.

Simple AI use register template

Date	Tool	Task	Data type	Risk	Review method	Owner
[date]	[tool]	[task]	[public/internal/primary/sensitive]	[1-5]	[check used]	[name/role]

8

Risk checker and practitioner checklist

A simple scoring method and a practical acceptance test for AI-supported work.

The risk checker combines three factors: the task type, the information type, and the potential influence of the output. The highest relevant risk should guide safeguards.

Component	Definition
Task value	1 = formatting or proofreading; 2 = organising evidence; 3 = analytical assistance; 4 = evaluative interpretation; 5 = replacing judgement or validation.
Data value	0 = public; 1 = internal non-sensitive; 2 = primary data; 3 = identifiable, confidential, political, or vulnerable-group data.
Impact value	0 = no influence; 1 = affects wording; 2 = affects analysis; 3 = affects recommendations, funding, policy, or public claims.
Scoring formula	Risk level = maximum of task value, data value + 1, and impact value + 1, capped at Level 5.

The responsible AI acceptance test

- [ ] Evidence checked
- [ ] Context preserved
- [ ] Bias challenged
- [ ] Human owner named
- [ ] Record completed

Practitioner acceptance checklist

- I can explain the finding or recommendation in my own words.
- I checked the output against the original evidence.
- I looked for missing, minority, or contradictory perspectives.
- I checked for bias, overconfidence, and loss of context.
- I documented AI use at the level required by the task risk.
- I know whether disclosure is needed for this use.
- I confirmed that consent and confidentiality conditions allow this AI use.
- I recorded who reviewed and approved the AI-assisted work.

9

Documentation standards and templates

Records should be proportionate: simple for low-risk tasks and detailed for consequential tasks.

Risk level	Minimum record	Recommended review	Typical decision
Level 1	Tool and task type.	Human review.	Use with light safeguards.
Level 2	Tool, task, input type, check performed.	Compare with sources.	Use with standard checks.
Level 3	Prompt, output, verification notes, human decision.	Sample checks and bias review.	Use with documented review.
Level 4	Full prompt, output, verification notes, approval.	Senior reviewer and stakeholder validation where appropriate.	Use only after formal approval.
Level 5	Full documentation and ethics review.	External review and documented waiver.	Do not proceed without exceptional approval.

Disclosure statement template

Template

This report used [AI tool name] for [specific task]. All outputs were reviewed by [name/role]. No confidential or participant-identifiable data was entered into the AI system.

Review note template

- AI-supported task completed: [task].
- Risk level assigned: [level].
- Evidence checked against: [sources].
- Bias, equity, and context review completed by: [name/role].
- Changes made after human review: [summary].
- Final owner of the judgement: [name/role].

10

Common pitfalls, examples, and glossary

Use these prompts to discuss real project risks and improve team practice.

Pitfall	What happens	Better practice
Generic findings	The output sounds polished but could apply to any project.	Link every finding to specific evidence, context, and stakeholder experience.
Lost minority views	Small but important perspectives disappear in summaries.	Check outliers, dissenting views, and disaggregated evidence.
Hidden escalation	A simple summary task becomes interpretation or recommendation writing.	Reclassify the task when use changes and apply stronger safeguards.
Invented certainty	The output states patterns more strongly than the evidence supports.	Preserve uncertainty, limitations, and alternative explanations.
Weak disclosure	AI use is hidden or described vaguely.	Explain what AI was used for, what was reviewed, and who remains accountable.

Mini case examples

Example 1 - Low risk: A team uses AI to improve the readability of a public executive summary after the findings have been finalised. The evaluator checks the revised language, confirms that no claims changed, and records the task as Level 1.

Example 2 - Moderate risk: An analyst asks AI to suggest themes from anonymised open-ended survey responses. The team reviews source excerpts, checks for minority views, documents the prompt and output, and treats the work as Level 3.

Example 3 - High risk: A consultant asks AI to draft recommendations for a politically sensitive evaluation using interview notes. This is Level 4 or Level 5 depending on the data. It requires approval, evidence mapping, stakeholder validation, and strict data protection review.

Glossary

Term	Meaning
AI-assisted output	Any text, table, summary, classification, suggestion, or draft produced with support from an AI tool.
Evidence-to-claim map	A check showing which evidence supports each finding, conclusion, or recommendation.
Human accountability	The principle that qualified people remain responsible for final decisions, judgement, interpretation, and reporting.
Prompt log	A record of instructions given to AI tools, usually kept for moderate or high-risk uses.
Risk level	A practical classification showing how much review, documentation, and approval a task needs.
Verification	The process of checking AI-assisted outputs against evidence, context, ethics, and professional judgement.

Final

A practical commitment

Responsible AI is a practice, not a checkbox.

A responsible Monitoring and Evaluation team does not need to reject AI. It needs to keep the discipline of evaluation at the centre: evidence, ethics, context, participation, transparency, and judgement.

Use this manual to support better project conversations, stronger documentation, clearer disclosure, and more careful professional judgement across Monitoring and Evaluation work.

GREEN
Use with normal checks

AMBER
Use with review and records

RED
Avoid or escalate first

Final field rule
AI can accelerate tasks, but it cannot carry professional accountability. The final responsibility belongs to the people doing the evaluation.

By EvalCommunity

The Ecosystem for Evaluation, AI, and Development Professionals

https://www.evalcommunity.com/