Australian Evaluation Society Ltd (AES)

ACN 606 044 624

Consultation paper

Responsible Use of AI in Evaluation

November 2026

Prepared by the AES Project Team

1. About the project

The AES is developing a policy and practical guidance on the responsible use of artificial intelligence (AI) in evaluation practice. This work is being undertaken by a project team of AES volunteers which convened in April 2026. The project team is seeking to deliver a draft policy to the AES Board in early 2027.

To inform the development of the policy we are seeking your views on how AI is currently used, its opportunities and risks as well as what guidance and support is required by evaluators.

We are defining AI as "*a collection of technologies that can be employed to perform tasks that would typically be done by a human, including at a scale, speed and complexity that humans cannot do. AI can also enable and augment human teams to:*

· *support better decision-making*
· *discover new ways of operating*
· *increase efficiency and resilience.*

AI may also refer to process automation or mathematical algorithms. It may also include machine learning, which is the ability for a computer to ingest and learn from data in order to make predictions or identify patterns." (page 6, Australian Signals Directorate, Ethical Artificial Intelligence in the Australian Signals Directorate, 2023)

The project has four phases:

1. **Explore and scope the issue:** through a literature review, identify the key uses, risks, opportunities and gaps related to AI use in evaluation.
2. **Consult and test emerging positions:** test proposed principles with AES members, First Nations evaluators and relevant partners.
3. **Establish a formal AES policy position:** set expectations for ethical, transparent and culturally responsible use of AI, for Board approval.
4. **Support professional capability:** identify implications for evaluator capability, professional learning and standards, and refer them to the Pathways Committee.

This paper supports the second phase. The principles below are a working draft, not an AES position.

2. About the consultations

The consultation runs in three phases:

1. **Member survey (October to November 2026):** a broad survey of AES members on current AI use, where they see value, and what concerns them.
2. **Focus groups (November to December 2026):** in-depth online discussions with specific groups to test the draft principles. This paper is for these sessions.
3. **Peer consultations (November to December 2026):** conversations national or international research and evaluation organisations who have written AI policies for their own members.

3. Summary of the literature review

The literature shows AI offers real opportunities for some evaluation tasks, particularly where it supports evaluative judgement rather than replacing it, and that responsible use depends on safeguards that take time and expertise to apply.

The project team reviewed 162 sources: peer-reviewed studies, reviews, conference papers, professional guidance and policy documents. The review investigated three questions:

- What are the current and potential uses of AI in evaluation practice?
- What are the risks of AI to current and future evaluation practice?
- What best practice guidelines and guardrails support safe AI use in evaluation practice?

Current use and opportunities for AI in evaluation. AI is being used or tested at most stages of evaluation, although much of the evidence comes from pilots, tests and proposals rather than routine practice.

- **Designing and framing:** AI can draft evaluation questions, program theory and plans for discussion. This is an emerging use with little testing. The evidence supports AI as an aid to design conversations, with evaluators keeping decisions about purpose and method.
- **Gathering evidence:** Retrieving and screening large volumes of documents is one of the more clearly documented uses in evaluation practice. AI can also widen the scale and timeliness of data collection and monitoring. AI-assisted interviewing and consultation is an emerging possibility for extending the reach of stakeholder engagement.
- **Analysing and synthesising:** Qualitative coding, where AI can generate codes, organise excerpts and propose themes, is the most extensively tested use, although whether AI belongs in reflexive, interpretive analysis is contested. AI can also support forecasting and parts of counterfactual analysis, but prediction does not by itself establish a program's effect.
- **Judging and reporting:** AI can provide scores, rankings and other inputs to appraisal, and help draft, translate and summarise findings for different audiences. In both cases AI helps with inputs and expression. A score or a fluent report is not, by itself, a warranted evaluative judgement.

Some sources also see potential to automate routine tasks, freeing evaluators for more complex, judgement-based work. Overall, the evidence supports AI as a task-specific assistant whose outputs are checked, not as a substitute for human interpretation.

What are the risks of AI to evaluation practice? The literature identifies risks that fall into three groups:

- **Validity of evidence:** bias and under-representation, weak causal claims, and fabricated or omitted information.
- **Rights and equity:** privacy and data security, intellectual property, human agency and public trust.
- **Institutional assurance:** accountability, capability, transparency, auditability, sustainability and resourcing.

How serious each risk is depends on the evaluation task, the data involved and the context. Coding community interview data, for example, carries different risks from screening published literature.

What safeguards support the responsible use of AI in evaluation? The literature identifies a range of safeguards that can help evaluators manage risks and use AI responsibly. They fall into four groups, which broadly correspond to the risks described above.

- **Accountability:** human responsibility for conclusions, with the expertise and authority to reject AI outputs; governance that names who reviews risk; and training so evaluators can question AI outputs, not just operate the tools.
- **Validity:** checking outputs for accuracy, fabrication and unsupported conclusions, and treating how AI is set up (prompts, grounding in sources, task design) as part of the method.
- **Rights and equity:** data governance, informed consent and attention to provider terms; checking for bias and exclusion; and respecting the rights, knowledge and perspectives of those involved or affected.
- **Transparency, monitoring and audit:** being clear about where, how and why AI was used, and monitoring and correcting AI use over time, with ways for people to challenge conclusions.

Most safeguards rest on professional guidance rather than testing. They also take time, expertise and resources, which can offset efficiency gains, and some sources raise AI's environmental costs. Both reinforce the need to consider when AI genuinely adds value.

The literature also highlights that safeguards are not without costs. Effective oversight and verification require time, expertise and resources, and may offset some of the efficiencies gained through AI use. Some sources also raise the environmental and resource burden of AI, although the evidence does not yet point to a specific safeguard for evaluation. This reinforces the importance of considering when and how AI adds value to evaluation practice.

Where is the evidence weak or contested? The evidence base for the responsible use of AI in evaluation is still emerging. Much of it draws on adjacent fields or broader AI governance rather than evaluation practice, and little comes from Australian evaluation practice. There is limited research on how safeguards are applied in evaluation and whether they work, or on culturally appropriate uses of AI, particularly in First Nations contexts. Findings are also often tied to particular tools and versions, which change quickly, so results may not carry over to newer tools.

Several questions remain unsettled. Views differ on the role of generative AI in reflexive and interpretive qualitative work. Productivity claims are hard to compare, because some studies count only task time and others include the checking and correction needed afterwards. The evidence does not show that prediction alone establishes a program's effect. Transparency is widely supported, but there is little evidence on what level of disclosure lets reviewers detect errors and challenge conclusions. Further research is needed, particularly on how AI is being used in Australian evaluation and with what effects.

4. Draft principles to test

Seven draft principles have emerged from the literature review, and we want to know whether they are right, clear and complete.

The principles are meant as a basis for judgement, not a checklist. How much weight each one carries will depend on the purpose, context and risk of the evaluation and the AI task involved.

1. **Evaluators remain accountable for evaluative judgement.** AI can assist with evaluation tasks, but responsibility for findings, conclusions and recommendations stays with people. Evaluators need the time, expertise and authority to question, correct or reject AI outputs, and the skills to use AI critically.
2. **AI use is fit for purpose and proportionate.** Evaluators choose whether and how to use AI based on the evaluation question, the method, the data and the risks. Some tasks and methods may not suit AI at all, and choosing not to use it is a legitimate decision.
3. **AI outputs are checked before they are relied on.** AI can be wrong, biased or fabricated in ways that look plausible. Evaluators verify outputs in ways that suit the task, keep a clear link between evidence and conclusions, and do not assume that accuracy on one task carries over to another.
4. **AI use is transparent.** Evaluators are open with commissioners, participants and audiences about when and how AI was used and what its limits were. They keep enough records for the process to be understood and scrutinised, while respecting confidentiality.
5. **Data, privacy and rights are protected.** Evaluators protect participant and program data through the whole AI workflow, including what tools retain, where data goes and how providers use it. Participants are informed and, where appropriate, consent to AI use. Indigenous data sovereignty is respected.
6. **AI use promotes fairness, inclusion and respect for cultural context.** Evaluators take reasonable steps to identify and address potential bias and exclusion, and consider how AI use may affect meaningful participation. Where AI use involves or affects First Nations peoples, knowledge or data, it respects their rights, cultural protocols and authority over their own knowledge and data.
7. **AI use strengthens evaluation practice and the profession.** Evaluators use AI in ways that build their expertise rather than replace it, and they keep practising the core skills of design, analysis and judgement. The AES and organisations support professional learning, and share the benefits and pressures of AI fairly across the evaluation workforce, including people early in their careers.