

THE AI INSTITUTE / OPERATING INTELLIGENCE

EXECUTIVE CASE NOTE · 17 OCT / 2026

Ask Which AI Errors the Average Hides



A reassuring score is not a reason to ignore the cases your organisation cannot afford to miss.

Boards, C-suite and senior management

Research period: October 2026 – October 2026

First published: 17 October 2026

Current to 17 October 2026

PUBLICATION RECORD

An Institute thesis and decision framework

Approve an AI system against the errors that matter in its intended job, not an average score that can hide them.



WHAT LEADERS SHOULD TAKE FROM THIS PAPER

- | | |
|----|--|
| 01 | Compare the result with a simple alternative before calling it progress. |
| 02 | Separate recognising a situation from choosing the right response. |
| 03 | Inspect costly misses and local conditions before approving wider use. |
| 04 | Keep the test's limitations attached to the investment decision. |

Research method

Institute synthesis of evidence from two versioned research papers, with their scopes retained. The approval check and warehouse illustration are original proposals, not reported experimental findings. The road-image case is not an incident report or a driving recommendation.

First published 17 October 2026 · Updated 17 October 2026 · Research period October 2026 – October 2026 · Version 1.0 · Suggested citation: The AI Institute, *Ask Which AI Errors the Average Hides* (2026).

CONTENTS

Inside this paper

01

A good average can answer the wrong question
Risk Casebook

02

Put a simple alternative beside the model
First ask what progress means

03

Separate an observation from a decision
Then inspect what matters

04

Keep changes inside the evidence boundary
Test the system you will use

05

Local evidence belongs in the purchase
Make transfer a question

06

Bring an error review, not another leaderboard
The next approval meeting

M

Method and limitations

R

References

How to use this brief

Start with the decision and use the framework to challenge the current plan. Figures retain their original populations and definitions. Institute analysis is labelled, and limitations are recorded before the references.

A good average can answer the wrong question

Before approving an AI pilot, ask for the mistakes behind its average score. A system can perform well on the easiest or most common cases while missing the exception that justifies buying it. The approval meeting should examine that exception, not only the summary slide.

A study by researchers at Rajshahi University of Engineering & Technology tested three open vision-language models on 500 selected Bangladesh road images. Hazard-presence accuracy was 88.0–89.2%; always choosing the majority category scored 89.0%. Identifying the hazard and its timing was much weaker. This was a diagnostic image test, not a live driving trial.¹

The paper was submitted on 28 September and appeared in the 16 October arXiv batch; its authors report acceptance at an ECCV workshop. Selected images, annotation limitations and differing model precision constrain the result. It does not establish failure rates on public roads or for other models.¹

The useful executive question is broader than transport: does our evidence test the decision we intend to delegate? The following review is the Institute's proposed practice, not a procedure evaluated by the study.

Put a simple alternative beside the model

Ask the team to show what happens without the proposed AI system. Depending on the task, that comparison might be the existing process, a straightforward rule or a person following the current checklist. A sophisticated model should not receive credit for solving a problem that a simpler method handles just as well.

Keep the comparison fair. Use the same cases, the same definition of success and the same deadline. Include the work needed to check and correct the output. A result that looks better before review may be less useful once that work is included.

Do not make the baseline artificially weak. If the current team has access to useful information, compare the proposed workflow with that reality. Equally, do not assume the current process is safe merely because it is familiar. Document its mistakes too.

The funding question becomes concrete: what improves, for whom, and at what additional cost? If the answer is not yet clear, fund a bounded comparison rather than describe the pilot as proven.

Separate an observation from a decision

Recognising a situation, explaining it and choosing an action are different jobs. A test of one should not silently authorise the next. Write down the actual decision the system will influence and what evidence supports each step towards it.

For a fictional warehouse example, spotting a damaged carton is different from deciding whether its contents can be dispatched. The second decision may require information that the photograph cannot supply. A fluent explanation should not be allowed to fill that gap with an assumption.

Choose a small set of errors that would change the approval decision. Include missed problems, unnecessary alarms and cases where the system should ask for help. The relative cost depends on the workflow; there is no universal acceptable accuracy percentage.

Show the number of examples behind each result. If only a handful of consequential cases were tested, say so. An empty row is missing evidence, not a perfect result. Have someone with operating knowledge review disputed labels before treating them as unquestionable truth.

Keep changes inside the evidence boundary

A separate October preprint, AlphaDiana, compares agent configurations on specified reasoning benchmarks. Adding tools, skills or retained history did not consistently improve correctness in those tests. It is not a production reliability study, but it supports a useful question: which complete configuration produced the result being presented? ²

Record the model, instructions, available tools, input format and review process used in your own trial. If one changes, decide which tests need to run again. Do not carry an approval across a material change simply because the product name remains the same.

Also specify what happens outside the tested conditions. A system might return work to a person, decline a request or stay in an advisory role. That fallback needs an owner and enough capacity to handle the expected workload. A theoretical escalation path is not an operating safeguard.

Local evidence belongs in the purchase

The Bangladesh case should keep its geography. It is not a proxy for Asia-Pacific, and no result here establishes comparative performance in North America, Europe, Africa, the Middle East or Latin America. The question that travels is whether the test resembles the intended use.

For a multinational organisation, compare sites rather than assume a country label explains them. Language, equipment, document conventions, connectivity and the availability of experienced reviewers may change the task. Two sites in one country can differ more than two across a border.

Ask local staff and affected users which examples the central team would overlook. Obtain suitable rights to use test material and minimise personal information. Local privacy and sector requirements need their own review; a benchmark result cannot stand in for it.

Make the supplier's evidence gaps visible in procurement. Specify who will produce local validation, who pays for it and what happens if the proposed use is unsupported. Avoid turning the customer's uncertainty into an unqualified claim of global readiness.

Bring an error review, not another leaderboard

Ask for one page containing the intended decision, a fair baseline, the consequential errors, the conditions tested and the fallback. Attach a few reviewed examples so the discussion stays connected to actual work.

Approve only the responsibility that the evidence supports. A useful assistant may still be unsuitable for independent action. Where an important question remains unanswered, give the next test a clear owner and completion condition before expanding access.

EXHIBIT / A PROPOSED APPROVAL CHECK

COMPARE A fair baseline What improves over the alternative? Institute proposed check	INSPECT Important errors Which misses change the decision? Institute proposed check	LIMIT Tested responsibility Where must the system hand back? Institute proposed check
--	---	---

RESEARCH RECORD

Method and limitations

Method

Institute synthesis of evidence from two versioned research papers, with their scopes retained. The approval check and warehouse illustration are original proposals, not reported experimental findings. The road-image case is not an incident report or a driving recommendation.

Limitations

No live deployment or independently replicated operational benefit is established. The selected research concerns particular data, models and configurations. The proposed management practice requires testing in the organisation's own workflow.



Sources and limitations

Source facts link to the original publication. Institute analysis reconciles definitions and turns the material into a management proposition. Frameworks are decision aids, not claims of universal causality. Read every important statistic with its population, date, definition and caveat.

Research cutoff: 17 October 2026. Review cycle: six months, or earlier after a material change in the underlying facts, standards or policy.

REFERENCES

Evidence behind the thesis

01 AI Maruf and colleagues: BD-HazardVLM

Version 1, submitted 28 September 2026; listed 16 October. arXiv comments report ECCV 2026 SDAD workshop acceptance, not main-conference acceptance. Full text inspected.

02 AlphaDiana: a measurement system for agents

Version 1, submitted 14 October 2026, listed 16 October. Preprint; no peer-review acceptance listed. Controlled reasoning benchmarks, not a production outcome evaluation.

THE AI INSTITUTE

Make the decision.
Then measure what changes.

Start with seven questions and a free readiness report. Then choose an optional 10–15 minute verbal assessment with Maya, our AI interviewer.

Start your free assessment: readiness.theai.institute

Research · Readiness · Capability

theai.institute