UNIVERSITY OF OXFORD

AI Competency Centre

AIIRS

AI Inherent Risk Scale

Overview

- The **AI Inherent Risk Scale** (**AIIRS**) provides a structured approach for classifying tasks that use generative artificial intelligence (GenAI) by evaluating them against three evidence-informed criteria to establish an overall inherent risk classification.
- AIIRS classifies tasks into **LOW**, **MEDIUM**, or **HIGH** inherent-risk bands. Classification is determined via three criteria—**epistemic dependence, verifiability, and consequences of error**—that define the nature and significance of a task's reliance on GenAI. These criteria consider the extent to which GenAI is expected to supply information, the degree to which the output can be independently verified, and the seriousness of any potential errors.
- AIIRS provides a consistent and defensible basis for **assessing the inherent risk associated with GenAI-assisted tasks.**

Purpose

- The purpose of AIIRS is not to determine whether GenAI should or should not be used, but **to establish the level of inherent risk associated with a task** that may need to be actively managed.
- AIIRS classifies the **inherent risk of a task** prior to the application of safeguards, mitigations, or governance controls.
- Once a task's inherent risk is understood, any **additional safeguards, mitigations, or design choices may be applied where warranted**, in line with any applicable governance arrangements.

Scope

- AIIRS is a **classification instrument** only, which indicates the **level of risk that must be actively managed** for a task that uses GenAI. It does not determine whether GenAI use is permitted, prohibited, ethical, compliant, or appropriate in any given context.
- AIIRS is designed for task-bounded human use of GenAI and **does not cover autonomous or agentic AI systems**, which introduce additional risks beyond the scope of this classification instrument.
- In AIIRS, “task” is used in its ordinary sense. It refers to the bounded piece of work the user is relying on GenAI to perform, whether a single action or a broader activity.

Alignment

- AIIRS does not replace or override institutional policy, regulatory obligations, assessment design decisions, or the exercise of human judgement. Classification outcomes must be interpreted and acted upon within existing governance, policy, and decision-making frameworks.
- The Australian Higher Education Standards Framework (HESF) requires providers to identify risks to academic quality and integrity and to manage those risks through informed judgement and established governance processes.
- AIIRS supports this requirement by providing a **shared, task-focused method for classifying the inherent risk of GenAI use** that can be applied by staff and students, while ensuring that decisions about safeguards, assessment design, and integrity responses remain within existing institutional governance, policy, and quality-assurance frameworks.

AIIRS

AI Inherent Risk Scale

LOW RISK TASK	MEDIUM RISK TASK	HIGH RISK TASK
		
Low epistemic dependence *and* **Embedded** verifiability *and* **Minimal** consequence of error	**Moderate** epistemic dependence *or* **External** verifiability *or* **Moderate** consequence of error	**High** epistemic dependence *or* **Expert** verifiability *or* **Significant** consequence of error

The overall risk level is determined by the highest-risk criterion present.

AIIRS

AI Inherent Risk Scale

Examples

- Critiquing clarity, coherence or tone
- Identifying weak transitions
- Suggesting alternative phrasings
- Proposing structural suggestions
- Listing considerations or constraints
- Identifying gaps in logic

Examples

- Summarising user-authored documents
- Rewriting user-authored text in a different style
- Producing illustrative examples for user-defined concepts
- Recommending general strategies or actions in professional contexts

Examples

- Answering factual questions about external realities, data, or people
- Producing summaries of real-world topics that rely on unstated knowledge
- Creating descriptions of laws, regulations or professional standards

Classification Criteria

Epistemic Dependence

Verifiability

Consequences of Error

Epistemic Dependence

- Epistemic dependence captures whether a task requires the system's representations of the world to be correct in order for the task outcome to be usable.
- Tasks with lower epistemic dependence rely only on user-provided material, without requiring the system's representations of the world to be correct for the task outcome to be usable.
- Tasks with higher epistemic dependence require the system's representations of the world to be correct for the task outcome to be usable.

Epistemic Dependence

HIGH

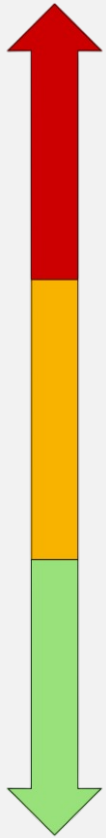
The task requires the system's representations of the world to be correct for the task outcome to be usable

MODERATE

The task relies on system-generated representations of the world, but correctness is not required for the task outcome to be usable

LOW

The task relies only on user-provided material, without requiring the system's representations of the world to be correct for the task outcome to be usable



Producing stand-alone factual claims, analyses, or advice about the world

Producing descriptions or profiles of real organisations, communities, or systems

Providing explanatory descriptions of real-world concepts or practices

Generating illustrative examples or analogies drawn from everyday contexts

Reordering or restructuring user-authored content

Correcting spelling, punctuation, and basic formatting

Verifiability

- Verifiability captures the basis on which the correctness of a GenAI system's output can be verified for the task.
- Verifiability is assessed independently of consequences. A task may be high risk due to unsourced verifiability, even where the immediate consequences of error are limited.
- Tasks with embedded verifiability enable quick, reliable verification by the user or the surrounding process, without requiring specific domain expertise.
- Tasks with unsourced verifiability depend on specialised expertise or external investigation that requires evaluative judgement.

Verifiability

EXPERT

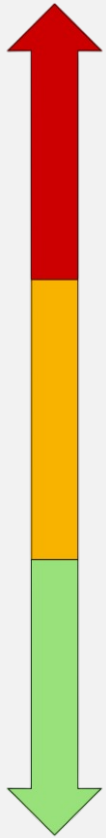
Correctness depends on specialised expertise or external investigation that requires evaluative judgement

EXTERNAL

Correctness requires verification using external sources, without specialised expertise

EMBEDDED

Correctness can be checked quickly and reliably using information already available for the task, without requiring specific domain expertise



Assessing stand-alone analyses or advice about real-world situations

Assessing the accuracy of descriptions or profiles of real organisations or systems

Evaluating whether a drafted public message is appropriate for a stated audience

Assessing whether a suggested timeline is plausible given a described sequence of tasks

Extracting explicitly stated information from a provided source

Identifying duplicated or repeated text segments

Consequences of Error

- The consequences of error reflect the extent to which incorrect, misleading, or incomplete GenAI outputs affect decisions, records, or outcomes related to the task.
- Tasks with minimal consequences of error are those in which errors have minimal impact on understanding or outputs and do not affect decisions, records, or outcomes relating to people beyond the task.
- Tasks with significant consequences of error are those in which errors affect decisions about people, alter records relating to them, or compromise outputs that have consequences for individuals or groups beyond the task.

Consequences of Error

SIGNIFICANT

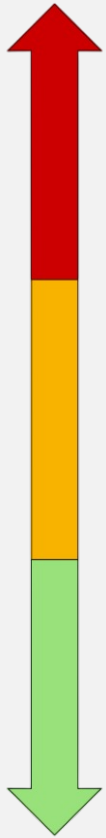
Errors affect decisions about people, alter records relating to them, with consequences for individuals or groups beyond the task

MODERATE

Errors influence reasoning, judgements, or actions taken within the task, but do not affect decisions, records, or outcomes beyond it

MINIMAL

Errors have minimal impact on understanding or outputs and do not influence decisions about people, records relating to them, or outcomes beyond the task



Producing outputs that determine or materially constrain outcomes for people

Generating information that directly affects individuals, groups, or communities

Producing preliminary analyses that feed into later decision-making processes

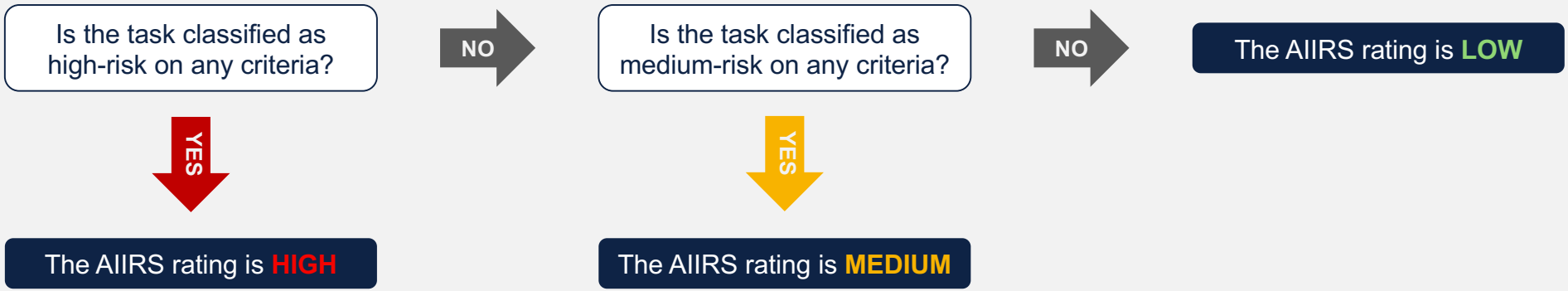
Drafting internal guidance or explanatory material for shared understanding

Generating placeholder or dummy content for layout, testing, or demonstration

Producing informal draft text for private reflection or experimentation

Calculating the AIIRS rating

AIIRS uses a classification approach in which the highest-risk element of a task determines its overall risk level, ensuring that any single high-risk feature of a task is not offset by lower-risk features elsewhere.



HIGH Risk Tasks

Tasks classified as **HIGH** must not proceed in their current form. One or more of the following interventions are required:

Accountable oversight

Proceed with the task only with explicit, accountable oversight, appropriate to its context.

Task redesign

Modify the task to reduce epistemic dependence on system output, improve verifiability, or limit downstream consequences.

Explicit safeguards

Apply explicit controls, such as expert review, independent verification, separation of GenAI output from final decisions, or documented human judgement.

Exclusion of GenAI

Where inherent risk cannot be reduced through redesign or appropriate safeguards, do not use GenAI for the task.

MEDIUM Risk Tasks

Tasks classified as **MEDIUM** require proportionate controls to manage identified risk. The following controls and conditions apply:

Risk-aware judgement

Apply appropriate oversight or reflective self-review to the task, proportionate to its context, with responsibility for identifying and managing risk resting with the person performing the task.

Targeted safeguards

Verify against external sources, clearly separate between GenAI output and final decisions, or apply documented checking processes, as appropriate to the task.

Task bounding

Bound the scope, audience, or reliance on GenAI output to support external or embedded verifiability and to reduce epistemic dependence or downstream impact.

Reclassification trigger

If the task's scope, use, or consequences change, reclassify and reassess the task.

LOW Risk Tasks

Tasks classified as **LOW** require routine care appropriate to the task and context. The following routine practices apply:

Routine practice

The task does not require specialised AI-specific safeguards beyond normal professional practice.

Assistive use only

GenAI contributes suggestions or transformations but is not treated as an authoritative source of information about the world.

Human checking

Review outputs for obvious errors, omissions, or misalignment with the user's intent.

Reclassification trigger

If the task's scope, use, or consequences change, reclassify and reassess the task.

AIIRS

AI Inherent Risk Scale