



unesco

Centre
Under the auspices
of UNESCO



ICAIRE

International
Center for
AI Research & Ethics

ICAIRE Framework Paper

Risk-Based AI Assurance in the Saudi Public Sector

Aligning International Assurance Practices with the National AI Risk Management Framework

Nawaf Alsahmah, Ph.D. (Independent Researcher)

<https://doi.org/10.68480/CEYS4608>

RECORD ICAIRE-EAI-2026-005



October 2026

Executive Summary

Saudi Arabia has established a national foundation for managing artificial intelligence (AI) risk through the National AI Risk Management Framework, issued by the Saudi Data and Artificial Intelligence Authority (SDAIA) in 2026. The framework structures risk management around context and scope, risk identification, risk assessment, risk treatment, and monitoring and review. The policy challenge addressed in this paper is therefore not the absence of AI risk governance, but how outputs from this established process can be connected to a distinct assurance process that supports justified confidence that applicable requirements have been satisfied.

The paper examines four international AI assurance practices — the United Kingdom, Australia, Canada, and the European Union — and analyses the Saudi framework using a common analytical lens of six conceptual dimensions. It also situates the proposed design against relevant international standards and frameworks, including ISO/IEC 42001, ISO/IEC 23894, ISO/IEC 42005, ISO/IEC 38507, and the NIST AI Risk Management Framework, and distinguishes its contribution from the nearest published framework of the same name (Muhammad & Yow, 2026).

The comparison shows substantial institutional variation but supports several common design lessons: defining assurance scope and applicable substantive requirements; applying assurance proportionately to risk and context; evaluating rather than merely generating evidence; addressing assurance across the AI lifecycle; and distinguishing assurance conclusions from subsequent organisational decisions. Notice, explanation, recourse, transparency, and information obligations are treated as substantive requirements where applicable rather than as a separate assurance dimension. The analysis does not support a universal mapping between risk levels and fixed assurance tiers.

The central need identified is an explicit and operationalised mechanism for Risk–Assurance Integration through which relevant outputs from the existing risk-management process can inform assurance without duplicating risk assessment. To address this need, the paper proposes a Risk-Based AI Assurance Framework comprising a Risk–Assurance Interface and four assurance-process components: Assurance Scope, Requirements, and Claims; Risk-to-Assurance Calibration; Assurance Execution and Evidence Evaluation; and Assurance Conclusion and Reassurance.

Within the proposed framework, assurance claims are linked to relevant existing risk records rather than mechanically mapped to risk categories. To make calibration operable without reinstating a mechanical mapping, the framework specifies four assurance rigour levels (R1–R4), five evaluator-independence levels (I1–I5), presumptive calibration floors, five evidence-adequacy criteria, and four assurance-conclusion categories. Existing risk-management records may contribute evidence, but that evidence must still be evaluated for adequacy. The resulting assurance conclusion remains distinct from residual-risk acceptability, informs rather than replaces organisational decision-making, and is reconsidered on material change or expiry.

For implementation, the paper recommends integration with existing risk-management and governance arrangements, mapping existing assurance-related activity onto the claim architecture, phased public-sector pilots with pre-stated measures and redesign triggers, and development of assurance capability. At the national level, it proposes SDAIA as the owner and issuer of national AI assurance guidance and the Risk–Assurance Interface specification, with organisational separation for higher-independence evaluations, together with common principles for proportionate assurance, standardised core documentation, and alignment of reassurance with existing lifecycle mechanisms.

The analysis is documentary and was conducted by a single analyst, using a corpus of three OECD member countries and the European Union. The framework has not been implemented or empirically validated; its scales, presumptive floors, evidence criteria and conclusion categories are design proposals to be tested through phased pilots before wider adoption.

Status. This framework paper proposes a design derived from comparative document analysis. The framework has not been implemented or empirically validated and should not be read as a settled national method; it is intended to be tested through the pilots specified in Section 8.3 before any wider adoption.

1 The Assurance Challenge in the Saudi Public Sector

1.1 From AI Risk Management to AI Assurance

Artificial intelligence is increasingly used across the public sector to support decision-making, service delivery, and the processing of large volumes of data. As adoption expands, public institutions must be able not only to identify and manage AI-related risks, but also to demonstrate that AI systems satisfy the legal, technical, ethical, and organisational requirements applicable to their use.

AI risk management and AI assurance address related but distinct questions. Risk management identifies, assesses, treats, and monitors potential risks. Assurance focuses on the evidential basis for justified confidence that applicable requirements have been met. An assessed risk level therefore does not, by itself, demonstrate that an AI system is safe, fair, reliable, transparent, or otherwise compliant. Risk and assurance should consequently be connected without being conflated: relevant risk-management outputs inform the assurance required, while evidence evaluation determines whether applicable requirements have been sufficiently demonstrated.

1.2 The Saudi Starting Point

Saudi Arabia already has a national foundation for managing AI risk. In 2026, the Saudi Data and Artificial Intelligence Authority (SDAIA) issued the National AI Risk Management Framework, providing a structured approach to identifying, assessing, treating, and monitoring AI-related risks across government and private-sector organisations (SDAIA, 2026).

The framework also addresses risk classification, residual-risk evaluation, escalation, monitoring, and lifecycle review, generating information that can support assurance. The policy challenge is therefore not to repeat risk identification and assessment, but to determine how outputs from the existing national process can inform assurance requirements and the evaluation of supporting evidence. The Saudi assurance challenge centres on integration, not replacement.

1.3 The Practical Assurance Gap

International approaches connect risk and assurance in different ways. For example, DSIT applies a general principle of proportionality (DSIT, 2024); Canada links Algorithmic Impact Assessment levels to corresponding requirements under the Directive on Automated Decision-Making (TBCS, 2025a); and the EU AI Act links regulatory risk classification to obligations and conformity-assessment requirements for high-risk AI systems (European Parliament & Council of the European Union, 2024). These approaches provide useful design prece-

dents, but no common operational method for the Saudi situation, where an independently established national AI risk-management framework already exists and its outputs must inform a complementary assurance process.

This paper refers to this broader relationship as **Risk-Assurance Integration**. A central operational component is **Risk-to-Assurance Calibration**: the use of relevant risk-management outputs to determine proportionate assurance requirements. Calibration considers relevant risk level(s), risk nature, and context of use without repeating or recalculating the underlying risk assessment.

The practical gap is not the absence of testing, review, monitoring, or documentation, which may already exist within risk-management and governance processes. It is the absence of an explicit and operationalised mechanism connecting risk-management outputs to a distinct assurance process that clarifies what must be demonstrated, what assurance is proportionate, and whether the available evidence is sufficient to support the relevant assurance claims and a distinct assurance conclusion, thereby supporting justified confidence that applicable requirements have been met. It must also accommodate material changes that may require renewed assurance and preserve the distinction between the assurance conclusion and the organisation's authority to decide on deployment, restriction, or continued operation.

On the evidential status of this gap. The gap is established analytically, by reading the published national framework against the six conceptual dimensions in Section 2.5, not by survey, interview, incident review or observation of practice inside Saudi public-sector entities. It should be read as a documented property of the published framework rather than a measured deficiency in institutional behaviour. Establishing its practical magnitude is the first objective of the pilots specified in Section 8.3.

1.4 Purpose and Scope

This paper proposes a risk-based AI assurance framework for the Saudi public sector that builds on, rather than duplicates or replaces, the National AI Risk Management Framework. It addresses the following guiding question:

How can international AI assurance practices be aligned with the Saudi National AI Risk Management Framework to enable proportionate assurance that supports justified confidence that applicable requirements have been met in public-sector AI use?

The paper examines selected international practices, identifies design lessons for risk-based AI assurance, assesses the Saudi framework from an assurance perspective, and proposes an operational approach for connecting existing risk-management outputs with assurance requirements, evidence

evaluation, and an assurance conclusion.

The proposed framework is policy-oriented and conceptual. It provides a basis for implementation and further testing rather than an empirically validated national assurance standard. Its scope conditions are summarised in Section 3.7 and its limitations are stated in full in Section 11.

1.5 Intended Audience and Use

The paper is addressed to three readers, and the distinction matters because the recommendations in Sections 8 and 9 fall to different parties.

1. **Public-sector entities deploying AI** — the users of Section 7 and Section 8. For them the paper is a design an entity could pilot within existing governance arrangements, using existing risk records, without waiting for national guidance.
2. **The prospective owner of national AI assurance guidance** — SDAIA, as the proposed owner and issuer of that guidance — the primary reader of Section 9.
3. **Assurance practitioners and reviewers** — internal audit functions, second-line risk functions, and external evaluators, who are the parties the calibration levels in Section 7.3 actually assign work to.

2 Foundations for Risk-Based AI Assurance

2.1 What AI Assurance Means

AI assurance builds on systems and safety-engineering practices in which evidence is used to justify confidence that defined requirements have been satisfied in context (Ashmore et al., 2021). The assurance-case tradition extends this logic through structured arguments linking claims and supporting evidence to specific system properties (Porter et al., 2024).

As its scope has expanded, AI assurance has also encompassed legal, ethical, and organisational concerns through techniques such as testing, audit, impact assessment, and conformity assessment (DSIT, 2024; Thomas et al., 2025). Guo et al. (2026) further distinguish *hierarchical* assurance, centred on documentary evidence against predefined criteria, from *relational* assurance, concerned with stakeholders' confidence in system trustworthiness. The framework proposed in this paper is primarily hierarchical in construction. Where assurance conclusions are communicated beyond the organisation, that communication may also contribute to relational assurance.

International normative instruments also extend the context

within which AI assurance may operate. UNESCO's Recommendation on the Ethics of Artificial Intelligence establishes a global, human-rights-centred normative framework and explicitly calls for impact-assessment mechanisms to identify and assess the benefits, concerns, and risks associated with AI systems, together with appropriate prevention, mitigation, and monitoring measures (UNESCO, 2022). UNESCO's Ethical Impact Assessment (EIA) provides a structured implementation tool for assessing the alignment of specific AI systems with the values and principles of the Recommendation (UNESCO, 2023).

For this paper, AI assurance is defined as:

A systematic, evidence-based process in which relevant evidence is collected and evaluated against defined requirements for an AI system within a specific context of use, in order to provide justified confidence that those requirements have been met.

This definition separates assurance from the individual techniques or assessment instruments used to generate or examine evidence.

2.2 Risk Management, Trustworthy AI, and Assurance

Trustworthy AI concerns the qualities expected of AI systems, such as safety, fairness, privacy, transparency, reliability, accountability, and human oversight. Risk management identifies and treats threats to those qualities and determines whether residual risk is acceptable. In contrast, assurance applies the evidence-based logic defined in Section 2.1 to determine whether applicable requirements have been sufficiently demonstrated.

This distinction is important in Saudi Arabia because an established national AI risk-management process already exists. A complementary assurance framework should therefore use relevant outputs from that process rather than repeat its functions.

2.3 Standards, Management Systems, and Assurance

International standards provide an important part of the infrastructure through which aspects of AI governance and assurance can be formalised and, in some cases, certified.

ISO/IEC 42001 specifies requirements for an artificial intelligence management system (AIMS) and provides a basis against which organisations may be certified by accredited certification bodies. Its relevance to this paper is twofold. First, certification may provide relevant externally generated evidence and may support evaluator independence where its scope is appropriate, but it does not by itself establish the

independence level required for claim-specific assurance. Second, it operates at a different unit of analysis from the framework proposed here. An AIMS certification provides an organisational-level conclusion about management-system conformity, whereas the assurance conclusion specified in Section 7.5 is a system-and-use-specific conclusion about whether identified substantive requirements have been demonstrated for a particular AI system in a particular context. The two are complementary, and an entity holding AIMS certification should be able to reuse a substantial part of its management-system evidence under Section 7.4 rather than generating it again.

ISO/IEC 23894 provides guidance on AI risk management and is the natural international reference point for the risk side of the interface. Where a Saudi entity operates its risk process under the national framework, alignment between that process and ISO/IEC 23894 concepts eases both evidence reuse and external evaluation.

ISO/IEC 42005 addresses AI system impact assessment, which is directly relevant to Section 7.4: impact-assessment output is one of the evidence types most likely to be available and most likely to be over-read as assurance in itself.

ISO/IEC 38507 addresses the governance implications of the use of AI by organisations, and speaks to the boundary in Section 7.5 between an assurance conclusion and a governing-body decision.

The NIST AI Risk Management Framework (Tabassi, 2023) organises AI risk management around the Govern, Map, Measure and Manage functions. Its Measure function provides a useful reference point for the evidence-evaluation logic used here, and its explicit separation of measurement from management supports the same distinction this paper draws between evidence generation and evidence-adequacy determination.

Two points follow for the design proposed in Section 7. First, adopted standards are a source of *substantive* requirements in their own right and are treated as such in Section 7.2, alongside law, regulation, policy, sectoral requirements and contract. Second, standards conformity is not a substitute for the claim-level adequacy determination specified in Section 7.4: a certificate is evidence about a management system, and it must still be evaluated for what it does and does not support about the specific claim in question.

These instruments are not analysed as comparative cases, because they are international standards and frameworks rather than jurisdiction-specific governmental assurance arrangements (see Section 3.1).

2.4 Related Frameworks and How This Paper Differs

The closest published work to this paper is Muhammad and Yow (2026), whose paper carries the same name as the framework proposed here (“Risk-Based AI Assurance Framework”) and addresses the same underlying construct: assurance graded to risk, with an explicit determination of whether assurance evidence is adequate.

Muhammad and Yow develop a generic, quantified instrument. Their approach combines a risk-severity assessment with modifying factors and produces composite scores — including an assurance-adequacy measure and a governance-readiness measure — that are portable across organisations and sectors and that support comparison between systems. Their contribution is measurement: rendering assurance adequacy as a value that can be computed, tracked and compared.

The framework proposed in this paper differs in three respects, each of which is a deliberate design choice rather than an omission.

1. **Unit of design.** This framework is not a free-standing instrument. It is an *interface* to a specific, already-established national risk-management process, and most of its content concerns what may be taken from that process, what must not be recalculated within assurance, and what must be returned to it. A generic instrument cannot make those determinations because it cannot assume a particular risk process exists upstream.
2. **Refusal of a composite score.** Section 7.3 deliberately does not aggregate multiple relevant risks into a single assurance figure, and does not derive a required assurance level from a computed value. The comparative analysis in Section 4.3 found no universal mapping between risk levels and fixed assurance tiers, and a composite adequacy score reintroduces exactly such a mapping through arithmetic. This paper instead uses presumptive floors with documented deviation (Table 9), which preserves proportionality while keeping the reasoning visible and contestable.
3. **Object of the conclusion.** Muhammad and Yow’s scores describe an organisation’s or system’s assurance posture. The conclusion specified in Section 7.5 is claim-level and categorical (Table 12), reports which claims are unsupported or only partially supported, and is explicitly not a licence to deploy.

The two approaches are compatible. An entity that wanted comparability across a portfolio of systems could compute measures of the kind Muhammad and Yow propose *on top of* the claim-level records produced under Section 7.4, and Muhammad and Yow’s demonstration that comparable risk

levels can lead to different deployment outcomes when evidence adequacy differs provides a useful illustration supporting Section 7.3's refusal of a mechanical mapping.

Two further strands of recent work bear on the design and are noted here rather than analysed. The assurance-case and claims-based tradition (Porter et al., 2024) supplies the claim–evidence structure used in Section 7.2, and the broader-harms critique (Thomas et al., 2025) is the reason Section 7.2 treats affected parties as material to how a claim is framed and evidenced.

2.5 Core Logic of Risk-Based AI Assurance

Across the reviewed sources, *context*, *requirements*, *evidence*, *evaluation*, and *justified confidence* recur as core elements of AI assurance (CDEI, 2021; NTIA, 2024; Porter et al., 2024). Context defines the system and its use; requirements establish what must be satisfied; evidence provides relevant information; evaluation determines what that evidence can support; and justified confidence reflects the defensible basis provided by evaluated evidence for believing that applicable requirements have been met.

This paper distinguishes evidence generation, evidence evaluation, and evidence-adequacy determination. The existence of a test report, audit record, monitoring result, or impact assessment does not itself establish assurance; evidence must be evaluated against defined requirements and judged adequate for the conclusion drawn (CDEI, 2021; NTIA, 2024). It also distinguishes *substantive* requirements — what an AI system or its use must satisfy — from *assurance* requirements — what is needed to establish sufficient confidence that those substantive requirements have been met, including assurance rigour, form, evaluator independence, and supporting evidence (CDEI, 2021; DSIT, 2024).

For risk-based AI assurance, this evidential logic must be connected to risk management. Approaches differ in where they draw the boundary: DSIT positions risk as an input to proportionate assurance, while MITRE adopts a broader conception that incorporates risk discovery, assessment, and management across the AI-system lifecycle (DSIT, 2024; Robbins et al., 2024). In the Saudi context, this paper adopts a functional rather than strict procedural distinction: the National AI Risk Management Framework retains its established risk-management functions and may also generate assurance-relevant information and evidence, which assurance can use without repeating the underlying risk assessment (SDAIA, 2026).

For this paper, **Risk-Based AI Assurance** refers to an assurance approach in which relevant risk-management outputs inform proportionate assurance requirements, while evaluated evidence provides the basis for justified confidence that applicable substantive requirements have been met. This broader relationship is referred to as **Risk–Assurance**

Integration, with **Risk-to-Assurance Calibration** as its central operational construct. International practices apply proportionality in different ways, including risk-proportionate assurance, impact-based requirements and peer review, and risk-linked conformity-assessment requirements (DSIT, 2024; TBCS, 2025a, 2025b; European Parliament & Council of the European Union, 2024).

Risk-to-Assurance Calibration is used in this paper as an analytical construct rather than as a common model identified across the reviewed practices. It determines how relevant risk information influences the level and form of assurance required without assuming a fixed mapping between risk levels and universal assurance tiers or recalculating the underlying risk assessment. Relevant considerations include risk level(s), risk nature, and context of use. Muhammad and Yow (2026) further illustrate the limitation of relying on risk level alone, showing that comparable risk levels can lead to different deployment outcomes when the adequacy of assurance evidence differs.

Risk-based assurance must also remain responsive across the AI lifecycle. Material changes may affect previous risk assessments, evidence, or assurance conclusions and may therefore require renewed review or assurance-related activity (NTIA, 2024; Robbins et al., 2024; European Parliament & Council of the European Union, 2024). For conceptual purposes, this paper distinguishes monitoring, risk reassessment, and reassurance; this distinction is analytical rather than a common taxonomy across the reviewed practices. Similarly, the assurance conclusion concerns what evaluated evidence supports and remains distinct from the subsequent organisational decision concerning deployment, restriction, remediation, continuation, or other action (CDEI, 2021; NTIA, 2024).

Drawing on the reviewed literature and the conceptual distinctions developed above, the author synthesised and organised these concepts into six interrelated conceptual dimensions of risk-based AI assurance, as presented in Table 1. These dimensions are not presented as a universal taxonomy or as sequential stages of an assurance lifecycle. Because the six dimensions serve as the analytical instrument for the comparative analysis, Table 2 provides an independent conceptual grounding check for each dimension using sources outside the comparative case corpus, while separately identifying the analytical contribution developed in this paper.

The original development of the analytical lens drew on a broader literature that included some documents later selected as comparative cases. This creates source overlap but does not by itself establish analytical circularity. Table 2 records the independent conceptual grounding of each dimension and separately identifies the analytical contributions developed in this paper.

Table 1. Conceptual Dimensions of Risk-Based AI Assurance.

Conceptual Dimension	Conceptual Focus
1. Assurance Basis, Scope, and Requirements	What is being assured, within what context and boundaries, and against which applicable requirements.
2. Risk-Assurance Relationship and Integration	How risk management and assurance are functionally related and how relevant risk-management outputs inform assurance.
3. Risk-to-Assurance Calibration	How relevant risk information influences the proportionate level and form of assurance required.
4. Assurance Execution and Evidence Evaluation	How assurance is performed and how evidence is generated or assembled, evaluated, and judged for adequacy.
5. Lifecycle and Reassurance	How monitoring, reassessment, and material change may affect the continuing validity of assurance.
6. Assurance Outcome	What outcome assurance produces, and how it relates to subsequent organisational decision-making or formal conformity.

2.6 A Design Assumption Stated Explicitly

This paper assumes that assurance is best positioned as a process *complementary* to the National AI Risk Management Framework rather than as an extension of it. The genuine alternative, adding the assurance functions identified in Section 5 inside the national risk framework, has real advantages: one process, one owner, one set of documentation, and no boundary to police.

The complementary design is preferred here for three reasons. It preserves the functional distinction between determining risk acceptability and determining evidential sufficiency, which Section 5.3 argues are different judgements requiring different criteria. It allows assurance to draw on requirement sources — law, sector regulation, contract, adopted standards — that lie outside the risk framework's remit. And it permits the higher independence levels in Section 7.3 to be met by parties outside the body that owns the framework, which a single integrated process makes structurally difficult.

The cost of the choice is the risk that entities operate two processes where one would do. Section 8.1 and Section 8.2 are written specifically to mitigate that risk, and the pilots in Section 8.3 include a measure designed to detect it.

3 Analytical Approach

This paper uses directed qualitative content analysis of institutional documents to examine international AI assurance practices and the Saudi National AI Risk Management Framework (Hsieh & Shannon, 2005). The analysis proceeded in

five stages: selection of the international corpus; analysis of the selected practices using a common conceptual lens; application of the same lens to the Saudi framework; synthesis of the international and Saudi findings; and development of the proposed framework. The coding procedure and audit trail are available from the author on request.

3.1 International Practice Selection

Four institutional practices, represented by five official documents, were selected to provide a focused and diverse corpus relevant to public-sector AI assurance. The selection criteria are available from the author on request.

The selected practices are analysed as institutional approaches rather than as comprehensive representations of their respective AI governance systems.

The corpus is deliberately focused rather than exhaustive, comprising three OECD member countries and the European Union, with no comparator from a governance context closer to Saudi Arabia's (Section 11). The effect is mitigated because the design lessons carried forward in Section 6 concern the internal logic of assurance rather than institutional form, and because the framework in Section 7 is built against the Saudi risk framework's own outputs rather than transplanting any of the four arrangements. Adding a comparator from a comparable administrative setting is a priority for future comparative work.

3.2 Analytical Lens and International Practice Analysis

The six conceptual dimensions developed in Section 2.5 were applied as a common analytical lens to the selected international practices. They were used as non-sequential areas of analysis rather than as stages of an assurance process. The analysis examined how each practice defines and scopes assurance, relates it to risk management, calibrates assurance requirements, evaluates evidence, addresses lifecycle change, and produces assurance outcomes.

3.3 Saudi Framework Analysis

The same analytical lens was then applied to the Saudi National AI Risk Management Framework without presuming the existence of assurance gaps. The analysis examined how the framework addresses each of the six conceptual dimensions and the extent to which assurance-specific functions are explicitly operationalised.

Because several findings in Section 5 are statements about what the framework does *not* specify, each such finding is accompanied by a locator identifying the part of the document that was examined in reaching it. Negative findings are only

Table 2. Independent Conceptual Grounding of the Six Analytical Dimensions.

Dimension	Independent conceptual grounding	Author's analytical contribution
1. Assurance Basis, Scope, and Requirements	Ashmore et al. (2021); Porter et al. (2024); CDEI (2021); NTIA (2024)	Distinguishing substantive requirements from assurance requirements and treating adopted standards as a potential source of substantive requirements.
2. Risk-Assurance Relationship and Integration	Robbins et al. (2024); Tabassi (2023); NTIA (2024)	Framing risk management and assurance as functionally distinct but connected through a two-way Risk-Assurance Interface with an explicit return path for material new risk information.
3. Risk-to-Assurance Calibration	NTIA (2024); Tabassi (2023); Muhammad and Yow (2026)	Developing Risk-to-Assurance Calibration as an explicit construct, including separate rigour and evaluator-independence dimensions and rebuttable presumptive floors.
4. Assurance Execution and Evidence Evaluation	CDEI (2021); NTIA (2024); Ashmore et al. (2021); Porter et al. (2024)	Distinguishing evidence generation, evidence evaluation, and evidence-adequacy determination, and specifying the five evidence-adequacy criteria used in Section 7.4.
5. Lifecycle and Reassurance	Ashmore et al. (2021); Robbins et al. (2024); NTIA (2024)	Distinguishing monitoring, risk reassessment, and reassurance, and treating material change and expiry as factors affecting the continuing validity of assurance.
6. Assurance Outcome	Porter et al. (2024); CDEI (2021); NTIA (2024)	Separating the assurance conclusion from residual-risk acceptability and organisational decision-making, and specifying the categorical conclusion set in Table 12.

Table 3. Selected International AI Assurance Practices and Official Sources.

Jurisdiction	Practice examined	Official source(s) analysed
United Kingdom	Risk-proportionate AI assurance	DSIT, <i>Introduction to AI Assurance</i> (2024)
Australia	Public-sector AI assurance	Australian Government et al., <i>National Framework for the Assurance of Artificial Intelligence in Government</i> (2024)
Canada	Impact-based requirements and peer review for automated decision-making	TBCS, <i>Directive on Automated Decision-Making</i> (2025a); <i>Guide to Peer Review of Automated Decision Systems</i> (2025b)
European Union	Conformity assessment for high-risk AI systems	Regulation (EU) 2024/1689, relevant provisions concerning high-risk AI systems and conformity assessment

as good as the reader's ability to check that the analyst looked in the right place, and the locators exist for that purpose.

3.4 Cross-Practice Synthesis

The findings from the international practices and the Saudi analysis were synthesised across the six conceptual dimensions. The synthesis examined areas of alignment and difference, identified assurance-related functional requirements relevant to the Saudi context, and distinguished between observed institutional practices, cross-practice findings, and design implications. These synthesised requirements informed the subsequent development of the proposed framework.

3.5 Proposed Framework Development

The synthesis was used to develop the proposed Risk-Based AI Assurance Framework. The framework was designed to build on relevant risk-management outputs and mechanisms already established in the Saudi National AI Risk Management Framework while introducing assurance-specific functions where required, without repeating the underlying risk-assessment process.

3.6 Coding Procedure

Directed qualitative content analysis begins with an a priori coding structure informed by prior literature, while retaining material that does not fit the initial categories for further examination (Hsieh & Shannon, 2005). The six dimensions in Table 1 constituted that structure. The unit of analysis, coding rules and disposition of the analytical record are available from the author on request, and the single-analyst constraint is recorded in Section 11.

3.7 Scope Conditions

The analysis is based on published institutional documents and does not observe implementation in any jurisdiction, including Saudi Arabia. A finding that a framework does not explicitly operationalise a function concerns the published document only, not unpublished guidance, annexes or practice. The proposed framework is a design proposal that has not been implemented, piloted or validated. Section 11 states the resulting limitations in full.

4 What International Practices Show

The selected international practices show that AI assurance does not follow a single institutional model. The United Kingdom, Australia, Canada, and the European Union differ in the basis and object of assurance, its relationship with risk management, the operationalisation of proportionality, the treatment of evidence, lifecycle arrangements, and the form of assurance outcome.

The purpose of the comparison is not to identify a single model for replication, but to derive design lessons for aligning international assurance practices with an established national AI risk-management framework.

4.1 Comparative Overview

The comparison shows that the object and basis of assurance differ materially across the selected practices. The immediate design implication is that assurance should begin by defining what is being assured and the applicable substantive requirements that form its basis, rather than with a generic catalogue of assurance techniques.

4.2 Assurance Basis, Scope, and Requirements

The selected practices differ materially in what is being assured and in the requirements or reference points against which assurance is performed. In the United Kingdom, AI assurance centres on measuring, evaluating, and communicating system trustworthiness, using context-relevant standards, benchmarks, regulatory requirements, and verifiable claims (DSIT, 2024).

Australia focuses specifically on government use of AI, grounded in Australia's AI Ethics Principles and the relevant legislative, policy, and operational context. Assurance is directed toward demonstrating, through evidence, that such use is safe and responsible (Australian Government et al., 2024). The Australian framework is also the clearest of the four in directing entities toward international standards as reference points for assurance activity, naming AI management-system, AI risk-management, and AI-governance standards among the instruments with which government assurance arrangements should align. This provides the most direct precedent in the corpus for treating adopted standards as a source of substantive requirements in Section 7.2.

Canada and the European Union adopt more formally prescribed assurance bases. In Canada, automated decision systems used to make or support administrative decisions are subject to defined requirements, including, where applicable, obligations relating to notice, explanation, and recourse. In the European Union, high-risk AI systems are assessed against

legally prescribed requirements, with relevant transparency and information duties applying in specified circumstances (TBCS, 2025a; European Parliament & Council of the European Union, 2024). Where applicable, these obligations form part of the substantive requirements against which assurance claims may be framed; they do not constitute a separate assurance dimension.

Overall, the comparison shows that there is no single common assurance basis or object. Assurance may centre on trustworthiness, responsible government use, policy compliance, or formal legal conformity. The transferable lesson is to define what is being assured, within what scope and context, and against which substantive requirements before selecting assurance techniques.

4.3 Risk-Assurance Integration and Calibration

Risk-Assurance Relationship. The selected practices differ in how risk relates to assurance. The United Kingdom positions AI assurance within wider organisational risk-management frameworks, while Australia places assurance within broader AI governance and identifies a risk-based approach as one of its cornerstones (DSIT, 2024; Australian Government et al., 2024). Canada provides a more explicit operational link, with the Algorithmic Impact Assessment determining the impact level that activates applicable requirements within the same policy architecture (TBCS, 2025a). In the European Union, risk management is itself a legal requirement within the conformity-assessment architecture for high-risk AI systems (European Parliament & Council of the European Union, 2024).

These arrangements differ institutionally, and none provides a directly transferable mechanism for linking outputs from an independently established national risk-management framework to a complementary assurance process.

Calibration. Risk classification identifies the relevant risk level or category, while proportionality reflects the principle that greater risk may require more demanding assurance. In this paper, Risk-to-Assurance Calibration refers to translating relevant risk information into proportionate assurance requirements.

Canada provides the clearest explicit link between impact level and subsequent requirements, including peer review at specified levels (TBCS, 2025a, 2025b). The United Kingdom and Australia apply proportionality more generally (DSIT, 2024; Australian Government et al., 2024), while EU conformity-assessment routes depend on regulatory classification and additional legal and technical factors (European Parliament & Council of the European Union, 2024). Overall, the practices show no universal mapping between risk levels and fixed assurance tiers.

It is worth being precise about what that finding does and

Table 4. Comparative Overview of Selected International AI Assurance Practices.

Conceptual Dimension	United Kingdom	Australia	Canada	European Union
1. Assurance Basis, Scope, and Requirements	AI system trustworthiness; context-relevant standards, benchmarks, regulation, and verifiable claims	Government use of AI; AI Ethics Principles and relevant legislative, policy, and operational context	Automated decision systems; defined requirements under the Directive on Automated Decision-Making	High-risk AI systems; legally prescribed requirements
2. Risk-Assurance Relationship and Integration	Assurance positioned as a component of wider organisational risk-management frameworks	Assurance positioned within broader AI governance, with a risk-based approach identified as a cornerstone of assurance practice	The Algorithmic Impact Assessment determines the applicable Directive requirements	Risk management is a legal requirement within the conformity-assessment architecture
3. Risk-to-Assurance Calibration	Higher risk may require a more robust combination of assurance techniques	Different risk levels may require different standards of assurance	Impact Levels I–IV activate scaled requirements; peer review applies from Level II upward	High-risk classification activates regulatory requirements; conformity route depends on additional legal and technical factors
4. Assurance Execution and Evidence Evaluation	Measure–evaluate–communicate; techniques include testing, audit, impact assessment, and conformity assessment	Testing, verification, monitoring, evaluation, and conformity assessment	Peer review evaluates the system, the AIA, and supporting evidence	Internal control or notified-body assessment; additional evidence or testing may be required where applicable
5. Lifecycle and Reassurance	Assurance techniques used across the AI lifecycle; no distinct renewed-assurance process specified	Lifecycle review primarily risk-management based; no distinct reassurance mechanism specified	AIA regularly reviewed and updated throughout the system lifecycle	Substantial modification requires a new conformity assessment
6. Assurance Outcome	Communication of system trustworthiness; verifiable claims	Evidence-based demonstration of safe and responsible government use of AI	Peer-review findings, recommendations and conclusion, considered separately before launch	EU declaration of conformity; CE marking; certificates where applicable

does not support. It does not support the conclusion that risk information should have no structured influence on assurance — Canada demonstrates the opposite. What it does not support is the *portability* of any particular mapping: the four practices scale assurance to risk in four different ways, keyed to their own legal instruments and institutional capacities. A framework designed for a fifth jurisdiction therefore cannot adopt one of these mappings, but it can adopt the underlying discipline of scaling deliberately and recording why. Section 7.3 implements this as presumptive floors rather than as fixed tiers.

4.4 Assurance Execution, Lifecycle, and Outcome

Assurance Execution and Evidence Evaluation. The practices differ in how assurance activities and evidence evaluation are operationalised. The United Kingdom uses a measure–evaluate–communicate approach supported by techniques such as testing, audit, and impact assessment (DSIT, 2024). Australia emphasises testing, verification, monitoring, and evaluation, while also noting that conformity assessments may assist testing and verification (Australian Government et al., 2024). Canada provides a more structured evaluative mechanism through peer review, which examines the system, the Algorithmic Impact Assessment, and supporting evidence (TBCS, 2025b). In the European Union, conformity

assessment evaluates compliance with applicable legal requirements; where notified-body assessment applies, further evidence or testing may be required (European Parliament & Council of the European Union, 2024).

Across these approaches, the existence of assurance activities or evidence does not by itself establish that applicable requirements have been satisfied; the evidence must be evaluated against the relevant basis of assurance.

Lifecycle and Reassurance. Lifecycle treatment also varies. The United Kingdom encourages the use of assurance techniques across the AI lifecycle (DSIT, 2024). Australia addresses lifecycle change primarily through risk monitoring and re-evaluation (Australian Government et al., 2024), while Canada requires the Algorithmic Impact Assessment to be reviewed and updated throughout the system lifecycle (TBCS, 2025a). The European Union provides the clearest example of renewed formal assessment, requiring a new conformity assessment where a previously assessed high-risk AI system undergoes a substantial modification (European Parliament & Council of the European Union, 2024). These differences reinforce the distinction between ongoing monitoring, risk reassessment, and renewed assurance.

Assurance Outcome. Outcomes also differ across the practices. The United Kingdom emphasises communicating system trustworthiness and supporting verifiable claims (DSIT,

2024). Australia focuses on evidence-based demonstration of safe and responsible government use of AI (Australian Government et al., 2024). In Canada, where peer review is required, it produces findings, recommendations, and a conclusion for consideration before system launch (TBCS, 2025b). The European Union produces formal regulatory conformity outcomes, including the EU declaration of conformity and CE marking (European Parliament & Council of the European Union, 2024).

These differences show that assurance-related outputs should not be conflated with subsequent organizational decisions or, where applicable, formal regulatory conformity status. Where assurance outcomes are communicated beyond the organization, communication forms part of the assurance outcome rather than a separate assurance dimension.

5 The Saudi National AI Risk Management Framework Through an Assurance Lens

The Saudi National AI Risk Management Framework provides an established national foundation for managing AI-related risks. It structures risk management through context and scope definition, followed by risk identification, assessment, treatment, and monitoring and review across government and private-sector organisations (SDAIA, 2026). These processes generate risk-management outputs and mechanisms that may also be relevant to AI assurance.

Accordingly, this section uses the six conceptual dimensions of assurance as an analytical lens to examine the risk-management outputs and mechanisms provided by the Saudi framework, the extent to which they can support relevant assurance functions, and whether any assurance-specific functions are not explicitly operationalised.

5.1 Existing Risk-Management Foundations

Unlike the selected international practices, which were examined for their assurance-oriented arrangements, the Saudi National AI Risk Management Framework is explicitly a risk-management framework. The analysis therefore does not expect it to reproduce the assurance structures identified in Section 4; rather, it examines which existing risk-management outputs and mechanisms may support a complementary assurance process.

The framework establishes context and scope, risk identification, assessment, treatment, and monitoring and review, supported by regulatory and ethical reference points. The framework also provides structured information on risk levels, controls, residual risk, responsibilities, escalation, and

lifecycle reassessment. Its illustrative scenario additionally demonstrates practices such as testing, audit logging, human review, and, in specific circumstances, independent review. These provide an existing foundation that assurance can draw upon rather than recreate.

5.2 Saudi Framework through the Assurance Analytical Lens

The fourth column of Table 5 records the part of the published framework examined in reaching each observation, so that any reader may verify the basis of a negative finding.

5.3 Interpreting the Analysis

Five distinctions are important when interpreting Table 5.

1. **Residual-risk evaluation should not be equated with assurance evaluation.** The framework reassesses likelihood and impact after controls are implemented and compares the resulting residual risk with approved risk tolerance. This is an evaluative process, but its purpose is to determine whether the remaining risk is acceptable or requires further treatment or escalation. It does not determine whether applicable requirements have been demonstrated through sufficient assurance evidence. The two judgements use different criteria and can diverge: a system may carry acceptable residual risk while the evidence that a specific legal requirement has been satisfied remains thin, and conversely a well-evidenced system may still present unacceptable residual risk.
2. **The framework already operationalises proportionality within risk management.** Risk level and context influence treatment, controls, acceptance, and escalation. This differs from Risk-to-Assurance Calibration, which concerns how risk information influences the level and form of assurance. The existence of risk-based proportionality in the Saudi framework therefore does not, by itself, constitute assurance calibration.
3. **The framework already generates substantial assurance-relevant information** through controls, indicators, documentation, monitoring, reassessment, and review; its illustrative scenario additionally demonstrates testing, audit logging, human review, and independent review in specific cases. Evidence generation, however, should be distinguished from evidence evaluation and evidence-adequacy determination. The framework does not establish a separate process for determining whether the available evidence is sufficient for assurance purposes.
4. **Lifecycle monitoring and risk reassessment should be distinguished from reassurance.** The framework provides for continuous monitoring and scheduled or event-triggered reassessment, including following changes in data, model

Table 5. Saudi National AI Risk Management Framework through the Assurance Analytical Lens.

Conceptual Dimension	Existing Saudi Framework Provision	Assurance Observation	Source examined
1. Assurance Basis, Scope, and Requirements	Context and scope definition; intended use; system boundaries; affected parties; and regulatory and ethical reference points	Provides contextual and normative foundations relevant to assurance, but does not define a distinct assurance object or assurance-specific requirements	SDAIA (2026), "Reference Foundations," pp. 14–16; "Context and Scope," p. 17
2. Risk–Assurance Relationship and Integration	Risk records, risk levels, treatment decisions, controls, residual risk, escalation, monitoring, and reassessment outputs	Provides structured risk-management outputs, but does not specify an explicit mechanism for translating them into a complementary assurance process	SDAIA (2026), "Risk Assessment," pp. 22–25; "Risk Treatment," pp. 26–27; "Monitoring and Review," pp. 28–29
3. Risk-to-Assurance Calibration	Risk level and context influence treatment, controls, acceptance, and escalation	Proportionality is operationalised within risk management, but no assurance-specific calibration mechanism is defined	SDAIA (2026), "Risk Assessment," pp. 22–25; "Risk Treatment," pp. 26–27
4. Assurance Execution and Evidence Evaluation	Controls, measurable indicators, documentation, monitoring, reassessment, and review; the illustrative scenario additionally demonstrates testing, audit logging, human review, and independent review in specific cases.	Existing mechanisms can generate assurance-relevant evidence, but no distinct process specifies how such evidence is evaluated for assurance or when it is sufficient	SDAIA (2026), "Risk Treatment," pp. 26–27; "Monitoring and Review," pp. 28–29; illustrative scenario, pp. 42–45, where relevant
5. Lifecycle and Reassurance	Continuous monitoring and scheduled or event-triggered risk reassessment	Lifecycle monitoring and risk reassessment are explicit, but their relationship to renewed assurance is not specified	SDAIA (2026), "Monitoring and Review," pp. 28–29
6. Assurance Outcome	Residual risk is assessed against approved risk tolerance and informs acceptance or escalation, including decisions on conditional launch, restricted use, or temporary or permanent cessation	The framework produces risk-management evaluations and organisational decisions, but does not specify a distinct evidence-based assurance conclusion	SDAIA (2026), "Risk Treatment," pp. 26–27

updates, expanded use, or new regulatory requirements. These mechanisms revisit the system's risk position, but do not specify whether or when such changes require renewed assurance.

- 5. The framework distinguishes residual-risk assessment from subsequent organisational decision-making**, including risk acceptance, conditional or restricted use, and temporary or permanent cessation, with escalation where required. These are risk-management and governance decisions. The framework does not separately define an evidence-based assurance conclusion as an additional output.

Two qualifications apply to all five. First, each concerns the published framework as examined at the locators recorded in Table 5; implementation guidance, annexes or sector instructions not publicly available may address these functions, and if they do, the corresponding observation in Table 5 no longer holds. Second, none of these observations is a criticism of the framework. A risk-management framework is not deficient for not being an assurance framework, and the analysis was conducted without presuming that assurance gaps would be found.

Taken together, the Saudi National AI Risk Management Framework provides substantial risk-management foundations that can support a complementary assurance process, while several assurance-specific functions are not

explicitly operationalised in the published document. These observations are carried forward into the cross-practice synthesis in Section 6.

6 Cross-Practice Synthesis

The findings from the selected international practices and the Saudi National AI Risk Management Framework were synthesized across the same six conceptual dimensions. The synthesis examines what these findings mean when considered together and identifies the functional requirements for a complementary risk-based assurance approach.

Overall, the synthesis indicates that the Saudi framework already provides substantial foundations on which risk-based AI assurance can build. A complementary assurance approach therefore need not recreate existing risk-management functions. Rather, it should use relevant outputs from them to determine whether sufficient evidence supports confidence that applicable requirements have been met.

The comparison also indicates that evaluation arrangements may vary according to the required assurance rigour and institutional context. However, the reviewed practices do not establish evaluator independence as a universal or uniformly risk-graded requirement. FR3 and FR4 therefore present a design tension that Section 7.3 must resolve rather than restate: the analysis supports scaling assurance to risk, provides no

Table 6. Cross-Practice Synthesis and Functional Requirements.

Conceptual Dimension	Synthesised Finding	Functional Requirement
1. Assurance Basis, Scope, and Requirements	International practices use different assurance bases, while the Saudi framework already provides context, scope, and relevant regulatory and ethical reference points. Adopted standards are treated as a requirement source in at least one practice, and some practices additionally impose notice, explanation, recourse, transparency, or information obligations that form part of the applicable substantive requirements.	FR1. Define what is being assured, which requirements apply, including notice, explanation, recourse, or information obligations where applicable, and what must be demonstrated.
2. Risk-Assurance Relationship and Integration	International practices connect risk and assurance in different ways. Canada provides the clearest link between impact assessment and subsequent requirements, while the Saudi framework already produces structured risk-management outputs that can inform assurance.	FR2. Establish an explicit connection through which relevant risk-management outputs inform assurance without repeating the underlying risk assessment.
3. Risk-to-Assurance Calibration	Proportionality is applied differently across the international practices, while the Saudi framework already applies risk-based proportionality within risk management. No universal mapping between risk levels and fixed assurance tiers emerges from the comparison.	FR3. Translate relevant risk information into proportionate assurance requirements without relying on a fixed one-to-one mapping — while still producing an output specific enough to act on.
4. Assurance Execution and Evidence Evaluation	Practices differ in testing, review, verification, peer review, conformity assessment, and evaluation arrangements. No common model of evaluator independence emerges. The Saudi framework already generates substantial assurance-relevant information.	FR4. Define appropriate evidence-evaluation arrangements and determine whether the available evidence is sufficient to support the relevant assurance claim.
5. Lifecycle and Reassurance	International practices address lifecycle change through different combinations of monitoring, reassessment, and renewed assessment. The Saudi framework already provides continuous monitoring and scheduled or event-triggered reassessment.	FR5. Determine when material change should trigger renewed assurance and its appropriate extent.
6. Assurance Outcome	International practices produce different assurance-related outcomes, including communicated trustworthiness, peer-review conclusions, and formal conformity outcomes. The Saudi framework already evaluates residual-risk acceptability and uses that evaluation to inform subsequent organizational decisions, but this differs from producing an evidence-based assurance conclusion and determining how that conclusion should be communicated where applicable.	FR6. Establish an evidence-based assurance conclusion distinct from residual-risk acceptability and organizational decision-making, and specify how that conclusion is recorded and communicated where applicable.

portable scale for doing so, and yet an entity cannot act on an instruction to scale without one. The resolution adopted is to specify scales explicitly as a proposal of this paper — distinguishable from the comparative findings, and testable in the pilots — rather than to present them as observed practice.

The resulting direction is therefore one of complementarity rather than replacement: the National AI Risk Management Framework remains the foundation for managing AI-related risks, while the functional requirements identified through the synthesis define what a complementary assurance approach needs to address.

7 Proposed Saudi Risk-Based AI Assurance Framework

The proposed framework was developed by translating the functional requirements identified through the cross-practice synthesis into an operational approach to risk-based AI assurance. The Saudi National AI Risk Management Framework remains the foundation on which the proposed assurance

process builds.

The proposed framework operationalises these functional requirements through a Risk-Assurance Interface and four inter-related assurance-process components: Assurance Scope, Requirements, and Claims (FR1); Risk-to-Assurance Calibration (FR3); Assurance Execution and Evidence Evaluation (FR4); and Assurance Conclusion and Reassurance (FR5, FR6). The interface addresses FR2. The Risk-Assurance Interface connects the existing risk-management process with assurance, while the four assurance-process components define what must be demonstrated, calibrate the required assurance, evaluate supporting evidence, and produce an assurance conclusion that may be reconsidered across the AI lifecycle.

The framework is designed to complement, rather than replace or duplicate, the existing national risk-management process. The resulting assurance conclusion remains distinct from residual-risk determination and informs, rather than replaces, organisational decision-making.

Everything specified in Sections 7.3 and 7.4 as a scale, floor or criterion is a **proposal of this paper**, not a finding of the comparative analysis. The distinction is maintained explicitly

throughout, because a reader deciding whether to adopt the framework needs to know which parts rest on observed practice and which rest on the author's design judgement. The latter are the parts the pilots in Section 8.3 are designed to test.

7.1 Framework Rationale and Risk-Assurance Interface

The proposed framework includes a Risk-Assurance Interface, a design mechanism that connects the existing national risk-management process with the assurance process while preserving their distinct functions. The interface identifies and routes relevant risk-management outputs to the appropriate assurance component without repeating the underlying risk assessment. It also returns new or materially changed risk information from assurance for risk reassessment and routes any revised risk-management outputs back to the relevant assurance component.

The Saudi National AI Risk Management Framework already produces relevant information, including system context and intended use, risk descriptions and levels, treatment decisions, implemented controls, residual-risk assessments, monitoring results, and reassessment findings (SDAIA, 2026). These outputs can support assurance scope and requirements, calibration, evidence evaluation, and lifecycle reconsideration. Applicable substantive requirements may also arise from relevant laws, regulations, policies, sector-specific requirements, contractual obligations, and adopted standards, rather than from the risk-management framework alone.

The interface also preserves the functional boundary between the two processes. Risk management remains responsible for identifying, assessing, treating, and reassessing risk and determining residual-risk acceptability. Assurance uses relevant outputs from that process but does not recalculate or independently revise the risk position. Where assurance identifies new information that may materially affect risk, that information is routed back through the interface for risk reassessment. Revised risk-management outputs may then re-enter assurance.

The interface therefore performs three functions:

1. Selecting relevant risk-management inputs for assurance.
2. Preserving the boundary between risk management and assurance.
3. Reconnecting the two processes when material change or new risk information requires reassessment.

This two-way relationship provides the operational basis for Risk-Assurance Integration. Responsibility for operating the interface should be assigned explicitly to a named role, be-

cause an interface that is everyone's responsibility is no one's; Section 8.1 addresses the assignment.

7.2 Assurance Scope, Requirements, and Claims

Building on the Risk-Assurance Interface, this stage defines the assurance scope, identifies the applicable substantive requirements, and articulates what must be demonstrated through assurance claims.

The **assurance scope** defines the boundaries within which the assurance conclusion will apply. It draws on contextual information already established through the national risk-management process, including the AI system or use, intended purpose, system boundaries, lifecycle stage, and affected parties. This step does not repeat context definition or risk identification; it uses existing information to establish the specific context within which assurance will be conducted.

Within that scope, **applicable substantive requirements** are identified from the relevant legal, regulatory, policy, sectoral, ethical, contractual, and adopted-standards context. These requirements establish what the AI system or its use is expected to satisfy. The Saudi risk-management framework may help identify relevant principles and regulatory references, but it is not treated as the sole source of applicable requirements.

Where the AI system supports decisions about individuals, the requirement-identification step should explicitly check for notice, explanation, and recourse obligations arising from applicable law or policy (FR1). These are substantive requirements in the same sense as accuracy or non-discrimination requirements: they can be expressed as assurance claims and evaluated through evidence. The Canadian practice examined in Section 4 illustrates that such obligations may scale with impact.

Affected parties are considered where their characteristics or the nature of the impact materially affect how an applicable requirement should be demonstrated. They do not create or alter the requirement itself; rather, they may influence how the corresponding assurance claim is framed and what evidence is relevant in the specific context.

Applicable substantive requirements are then articulated as **assurance claims**. A claim states what must be demonstrated about an applicable requirement before confidence in its satisfaction can be justified. Claims do not create new legal, ethical, technical, or organisational obligations; rather, they express existing requirements in a form that can be supported and evaluated through evidence.

As a proposed design approach, each assurance claim should be:

1. **Clear** — stated in terms that admit of evidence for or

against, without evaluative language that cannot be evidenced.

2. **Context-specific** — bounded to the system, use, population and lifecycle stage in the assurance scope.
3. **Traceable** — linked to a named applicable substantive requirement, so that the claim's authority is visible.
4. **Evaluable** — capable of being supported or not supported by evidence obtainable at proportionate cost.
5. **Bounded in time** — stated for a defined period or system configuration, so that its expiry is visible rather than implicit.

This sequence prevents a direct transition from risk classification to assurance technique. Risk level alone does not determine whether testing, audit, independent review, or another assurance activity is appropriate. What must be demonstrated is first established through the applicable substantive requirement and corresponding assurance claim; the proportionate strength and form of assurance are determined subsequently through Risk-to-Assurance Calibration.

Where material changes affect the system, context, affected parties, or applicable requirements, the assurance scope, requirements, and claims should be reconsidered and revised where necessary.

The output of this stage is a defined assurance scope, an identified set of applicable substantive requirements, and explicit assurance claims establishing what must be demonstrated. Appendix B provides a template for recording them.

7.3 Risk-to-Assurance Calibration

Once the assurance scope, applicable substantive requirements, and assurance claims have been defined, Risk-to-Assurance Calibration determines the proportionate level and form of assurance required for each claim.

For each assurance claim, calibration first identifies the existing risk record or records relevant to that claim, where such records exist. This linkage is based on the relationship between the claim and the identified risk, rather than on a fixed one-to-one correspondence between requirements and risk categories. Calibration uses the existing risk assessment; it does not repeat that assessment or create a separate assurance risk score.

Where no relevant risk record exists, the claim is not automatically excluded. If failure of the claim could materially affect the system's risk position, the issue is referred through the Risk-Assurance Interface to the risk-management process to determine whether a new or revised risk record and risk assessment are required. Where the claim reflects a mandatory substantive requirement, assurance of that claim continues while the risk-management review is undertaken. Any new or

revised risk-management outputs are then returned through the interface and incorporated into calibration.

Calibration considers three main inputs for each claim: the relevant **risk level**, the **risk nature**, and the **context of use**. Risk level is therefore important, but it does not by itself determine the assurance required. Risk nature matters because risks of different kinds are evidenced differently — a risk to individual rights is not demonstrated by the same evidence as a risk to service continuity — and context of use matters because the same model in a different decision setting bears differently on the same requirement.

Where several risks are relevant to one claim, they are considered separately rather than combined into a new risk score. As a proposed design rule, the most demanding applicable condition establishes the presumptive minimums for assurance rigour and evaluator independence, while the resulting assurance approach should address all relevant risks.

Calibration determines two aspects for each claim: how strong the assurance must be, and how independent the evaluation must be.

Assurance Rigour

Rigour is a property of the depth and intensity of the evaluation performed, not of the evaluator's independence or the technique chosen. Testing, audit, impact assessment, conformity assessment, and other assurance techniques may each be used at more than one rigour level; what changes with the level is the extent of evaluation, verification, challenge, and supporting documentation.

Evaluator Independence

Independence is specified separately from rigour because the two do not move together in the reviewed practices. The Canadian arrangement introduces peer review — a form of independence — at defined impact levels while leaving evaluation technique to the entity; the EU regime varies the independence of the conformity route on legal and technical grounds that are not purely a function of risk level (Section 4.3). Separating the two dials preserves that observed variation instead of collapsing it.

Presumptive Floors

The comparative analysis found no portable mapping from risk level to assurance tier (Section 4.3), and this framework does not create one. It does, however, need to produce an output an entity can act on. The mechanism proposed is a set of **presumptive floors**: a minimum rigour and independence level that applies unless the entity records a documented justification for departing from it. A floor differs from a tier in

Table 7. Assurance Rigour Levels. *(proposed in this paper)*

Level	Designation	What is done	Documentation expected
R1	Basic documented evaluation	Existing and available evidence is reviewed against the assurance claim, and a documented judgement is made on whether that evidence is adequate to support the claim. Expanded verification or additional testing is not required by default.	Claim record, evidence register, and documented adequacy judgement with reasons.
R2	Structured evaluation	Evidence is assessed systematically against predefined evaluation criteria. Targeted additional evidence is obtained where the existing evidence is insufficient to evaluate the claim.	R1 documentation plus predefined evaluation criteria, evaluation findings, and any additional evidence obtained.
R3	Enhanced verification	The evaluation goes beyond summaries to examine underlying evidence through sampling, rechecking, targeted testing, or other proportionate verification activities. An explicit adequacy determination is made for each claim.	R2 documentation plus verification records, methods used, findings, and an explicit per-claim adequacy determination.
R4	Comprehensive assurance evaluation and challenge	The claim and supporting evidence are subjected to the highest level of rigour, including broader coverage, deeper verification, challenge of material assumptions and results, and testing or reperformance where appropriate. Gaps, limitations, and unresolved issues are explicitly documented.	R3 documentation plus the full evaluation and challenge record, testing or reperformance results where applicable, material limitations, unresolved issues, and a reviewable basis for the conclusion.

Table 8. Evaluator Independence Levels. *(proposed in this paper)*

Level	Evaluator	Independence characteristic
I1	The development or operating team	No separation; self-assessment team
I2	A colleague or function within the same business area, not involved in developing or operating the system	Separation of individuals, not of reporting line
I3	Internal audit or an equivalent assurance function outside the delivery chain	Separation of reporting line within the entity
I4	An external party commissioned by the entity	Organisational independence; commercial relationship remains
I5	An accredited or mandated third party where law, regulation or policy requires it	Organisational independence with external accountability for the evaluation itself

three respects — it sets a minimum rather than a prescription, it is rebuttable, and rebutting it leaves a reviewable record.

Three constraints apply to the table.

- Legal minimums are not reducible.** Any assurance or conformity requirement already prescribed by applicable law, regulation, policy, or another binding requirement remains a minimum that calibration cannot lower, whatever Table 9 would otherwise indicate.
- Departures are recorded, not merely permitted.** A presumptive floor may be departed from downward with documented justification, or exceeded upward where the circumstances warrant. Any downward departure must be recorded in the claim register (Appendix B) and forms part of the assurance documentation available to any subsequent evaluator.

Table 9. Presumptive Calibration Floors. *(proposed in this paper; rebuttable with documented justification)*

Condition (any one applies)	Presumptive floor
The claim concerns a system whose failure could affect the rights, safety, eligibility, or entitlements of individuals, and the relevant risk is assessed at the highest level in use by the entity.	R3 / I3
As above, where the relevant risk is assessed at a middle risk level.	R2 / I2
The claim concerns a binding requirement imposed by applicable law, regulation, policy, or another binding instrument.	R2 / I2 minimum, or any higher minimum prescribed by the binding instrument. Calibration cannot reduce that prescribed minimum.
The claim concerns low-risk internal, operational, service-continuity, or comparable matters with no direct material effect on individuals.	R1 / I1
The claim cannot yet be linked to an existing risk record and the referral under Section 7.3 remains open.	R2 / I2 as a precautionary interim floor pending the risk-management determination; this does not classify the claim as medium risk.

- The floors are a proposal, not an observed standard.** They are the author's design judgement, calibrated to be operable rather than derived from any reviewed practice, and they are among the specific elements the pilots in Section 8.3 are designed to test. Entities adopting the framework before national guidance exists should treat the levels as local and expect them to change.

The output of calibration is therefore a **claim-specific assur-**

Table 10. Claim-Level Evidence and Evaluation Elements.

Element	Function
Evidence and source/activity	Define the evidence needed to support the claim and identify the existing source or additional assurance activity required to obtain it
Evaluator	Assign evaluation responsibility consistent with the degree of independence determined during calibration (I1–I5)
Evidence-adequacy criteria	Define how the combined evidence will be judged sufficient to support the claim at the calibrated rigour level (R1–R4)

ance configuration defining the required assurance rigour (R1–R4), the appropriate assurance technique or techniques, and the required degree of evaluator independence (I1–I5), together with any recorded departure from the presumptive floor.

7.4 Assurance Execution and Evidence Evaluation

The claim-specific assurance configuration established through calibration is operationalised through assurance activities and evidence evaluation.

For each claim, relevant evidence may be drawn from existing risk-management records, including treatment documentation, control indicators, testing results, monitoring records, audit logs, and review findings. Additional assurance activity should be undertaken where the available evidence is insufficient to support the claim at the calibrated level of assurance. Where several risks relate to one claim, the evidence should address the relevant implications of all of them.

Structured impact-assessment instruments may also contribute assurance-relevant evidence where appropriate. UNESCO’s Ethical Impact Assessment provides a structured approach to examining the ethical implications of a specific AI system in context and may contribute evidence relevant to corresponding assurance claims (UNESCO, 2023); ISO/IEC 42005 addresses the same territory in standards form. Where an entity holds AI management-system certification, evidence arising from that certification may contribute to claims concerning governance and process requirements. Its relevance remains claim-specific: certification does not by itself establish that the substantive requirement underlying a particular assurance claim has been satisfied.

As a proposed operational structure, three elements should be specified for each assurance claim.

The framework distinguishes between evidence generation, evidence evaluation, and evidence-adequacy determination. The existence of evidence does not itself demonstrate that a claim has been sufficiently supported. Five evidence-adequacy criteria are proposed for that judgement, each expressed as a question the evaluator must answer in

writing.

Sufficiency is assessed last and at the level of the whole body of evidence, because the adequacy question is collective: individually relevant, valid, current and appropriately scoped evidence items may still leave a claim only partly supported, and the criteria are designed to make that outcome visible rather than to force a binary verdict.

Evidence evaluation must also remain distinct from residual-risk evaluation. The Saudi framework assesses the effectiveness of risk treatment and controls and determines whether residual risk is acceptable against approved risk tolerance. These findings may contribute relevant evidence to assurance, but their primary purpose remains to determine risk acceptability rather than whether the evidence sufficiently supports an assurance claim (SDAIA, 2026).

Where evidence is insufficient, additional evidence or assurance activity may be required, remediation may occur before re-evaluation, or the assurance conclusion may be limited to what the available evidence supports. The EU conformity-assessment regime provides a specific precedent for requesting additional evidence or testing where submitted material is insufficient (European Parliament & Council of the European Union, 2024).

If assurance reveals information that could materially affect an existing risk assessment, that information is returned through the Risk-Assurance Interface for reassessment by the risk-management process rather than being reclassified within assurance.

The output of this stage is a documented evidence-adequacy determination for each assurance claim, providing the basis for the assurance conclusion.

7.5 Assurance Conclusion and Reassurance

The evidence-adequacy determinations for individual assurance claims are synthesised into an assurance conclusion. The conclusion should identify which claims are sufficiently supported, which are unsupported or partially supported, and any material evidence limitations. It should not extend beyond what the evaluated evidence can justify.

To keep conclusions comparable across systems and entities, four categories are proposed at claim level, with the system-level conclusion reporting the distribution rather than averaging it.

The distinction between *not supported* and a finding of non-compliance is important and easily lost. An assurance conclusion of “not supported” states that the entity cannot currently demonstrate satisfaction of a requirement. Whether the requirement is in fact breached is a different question, answered by a different process, and the assurance record should not be written in a way that conflates them.

Table 11. Evidence-Adequacy Criteria. *(proposed in this paper)*

Criterion	The question the evaluator answers	Typical failure
Relevance	Does this evidence speak to the claim <i>as stated</i> — the same system, the same use, the same population, the same lifecycle stage?	Evidence from a development environment offered for a claim about production use
Validity	Does the method that produced the evidence support the inference drawn from it?	A functional test offered as evidence about outcomes across population groups
Currency	Does the evidence reflect the system and context as currently configured, within the claim's stated time bound?	Pre-retraining performance evidence offered after a model update
Coverage	Does the evidence address the claim's full scope, including any subgroup, edge case or use variant within it — and where it does not, is the gap stated?	Aggregate performance offered for a claim that is stated across groups
Sufficiency	Taken together, and at the calibrated rigour level, does the body of evidence support the claim — and if only in part, what part?	Individually adequate evidence items treated as collectively sufficient despite unresolved gaps in the overall body of evidence.

Table 12. Assurance Conclusion Categories. *(proposed in this paper)*

Category	Meaning
Supported	The evaluated evidence is adequate to support the claim as stated, at the calibrated rigour level, within its stated time bound
Partially supported	The evidence supports the claim for a defined subset of its scope; the excluded part is stated explicitly
Not supported	The available evidence does not support the claim; this is a finding about evidence, not a finding that the requirement is breached
Not assessed	The claim was in scope but not evaluated; the reason is recorded

The assurance conclusion remains distinct from both residual-risk determination and organisational decision-making. The Saudi National AI Risk Management Framework determines residual-risk acceptability and supports organisational decisions concerning risk acceptance, conditional or restricted use, temporary suspension, or permanent cessation, with escalation where required (SDAIA, 2026). The assurance conclusion provides an additional evidence-based input to this decision process; it does not itself authorise or prohibit deployment.

Communication of the conclusion. Beyond its internal governance function, communication of an assurance conclusion may contribute to wider stakeholder confidence. Its form and audience should depend on applicable requirements and context.

- 1. Internal record (all cases).** The full claim register, evidence, evaluations, and conclusion are retained and made available to the entity's relevant governance bodies and any subsequent evaluator.
- 2. Governance report (all cases).** The assurance conclusion by category, unsupported and partially supported claims, material evidence limitations, and any documented de-

partures from presumptive calibration floors are reported to the body responsible for the organisational decision before that decision is taken.

External communication, where applicable. Any communication beyond the organisation should be determined by applicable law, policy and context, taking account of security, confidentiality, data-protection and operational considerations. Its appropriate form and extent should be tested through the proposed pilots.

Reassurance. Assurance should be reconsidered when material changes affect the system, its context, applicable requirements, relevant risks, assurance claims, or supporting evidence. Existing monitoring and risk-reassessment mechanisms under the Saudi framework can provide important triggers for this reconsideration. Where risk-management outputs change, the revised information re-enters assurance through the Risk-Assurance Interface.

The extent of renewed assurance may range from targeted reconsideration of affected claims or evidence to broader renewal of the assurance process, depending on the nature and materiality of the change. International practices provide precedent for renewed assessment following material change: the EU AI Act provides for renewed conformity assessment in relevant cases involving substantial modification of a high-risk AI system (European Parliament & Council of the European Union, 2024).

Because each claim carries a stated time bound (Section 7.2), expiry is itself a reassurance trigger, independent of change. A claim whose stated time bound has expired should no longer be treated as currently supported and should be marked as requiring reassessment until re-evaluated.

The proposed framework therefore produces a traceable assurance conclusion that informs, but does not replace, organisational decision-making; may be communicated beyond the organisation where required or appropriate; and is reconsid-

ered where material change or the passage of time affects its basis.

7.6 Worked Example

Appendix A carries a single assurance claim through every stage of the framework — from substantive requirement, through claim formulation, risk-record linkage, calibration against Table 9, evidence assembly, adequacy evaluation against Table 11, conclusion under Table 12, and the return path through the interface — for an AI-supported eligibility screening system. It is included for two reasons: to give an adopting entity a concrete pattern to follow, and to test the framework's own executability. The example is illustrative and does not describe any actual system.

8 Implementation Considerations

Implementation of the proposed Risk-Based AI Assurance Framework should build on existing AI risk-management and governance arrangements within Saudi public-sector entities rather than create a parallel process.

8.1 Integrate with Existing Arrangements

Existing risk-management responsibilities and decision pathways should be preserved. Three assignments are needed in addition.

1. **Ownership of the Risk-Assurance Interface.** A named role should be accountable for selecting risk-management inputs, maintaining the documented links between assurance claims and risk records, and routing material new risk information back for reassessment. Locating this role inside the risk function keeps the boundary clear but risks assurance being treated as a risk output; locating it in a governance or assurance function keeps the distinction but requires access to risk records. Either is workable; leaving it unassigned is not.
2. **Evaluator assignment consistent with calibration.** Where calibration produces an I3 requirement, the entity must have a function reporting outside the delivery chain able to take the work. Entities without one should either establish the capability, use I4 external evaluation, or record the departure from the presumptive floor under Section 7.3.3 — the third option being legitimate but visible.
3. **Coordination of conclusions with decision points.** The governance report specified in Section 7.5 must reach the deciding body *before* the decision, which requires the assurance cycle to be timed against the entity's own approval gates rather than run in parallel with them.

8.2 Map Existing Activity onto the Claim Architecture

Entities already performing testing, internal audit, model review, privacy impact assessment or control monitoring will ask whether that activity counts as assurance under this framework. It frequently does, and the framework is designed so that adoption begins by relabelling and linking rather than by generating new evidence. A practical sequence:

1. **Inventory what is already produced** for the system in scope: test reports, monitoring dashboards, control indicators, audit findings, impact assessments, treatment records.
2. **Draft the claims first, independently of that inventory**, from the applicable substantive requirements. Doing it in the other order produces claims shaped to the available evidence, which is the principal way assurance becomes self-confirming.
3. **Map each existing item to the claims it bears on**, noting the criteria in Table 11 it satisfies and those it does not.
4. **Read off the gaps.** The claims with no mapped evidence, and those for which the mapped evidence does not adequately satisfy the relevant criteria in Table 11, are the entity's actual assurance backlog. This list — not a new testing programme — is the starting point.
5. **Calibrate, and compare with what the existing activity delivers.** Where existing activity already meets or exceeds the calibrated rigour and independence levels, no new work is required beyond documenting the evaluation. Where it falls short, the shortfall is specific and costable.

This sequence also produces the measure that Section 8.3 uses to test the framework's central premise: the proportion of claims that can be evidenced, wholly or in part, from material the entity already holds.

8.3 Pilot and Refine Before Wider Adoption

Adoption should begin with limited, phased public-sector pilots across different risk profiles and operational contexts, with the findings used to refine guidance, templates, documentation requirements, and responsibilities before broader implementation. A pilot that reports only whether participants found the framework useful will not answer the questions that matter, so the following design is proposed.

Scope. A small number of entities — enough to span at least three distinct settings: a system making or supporting decisions about individuals; a system used for internal or operational purposes only; and a system procured from a third party rather than built in-house. The third case is included because evidence access is materially different when the entity

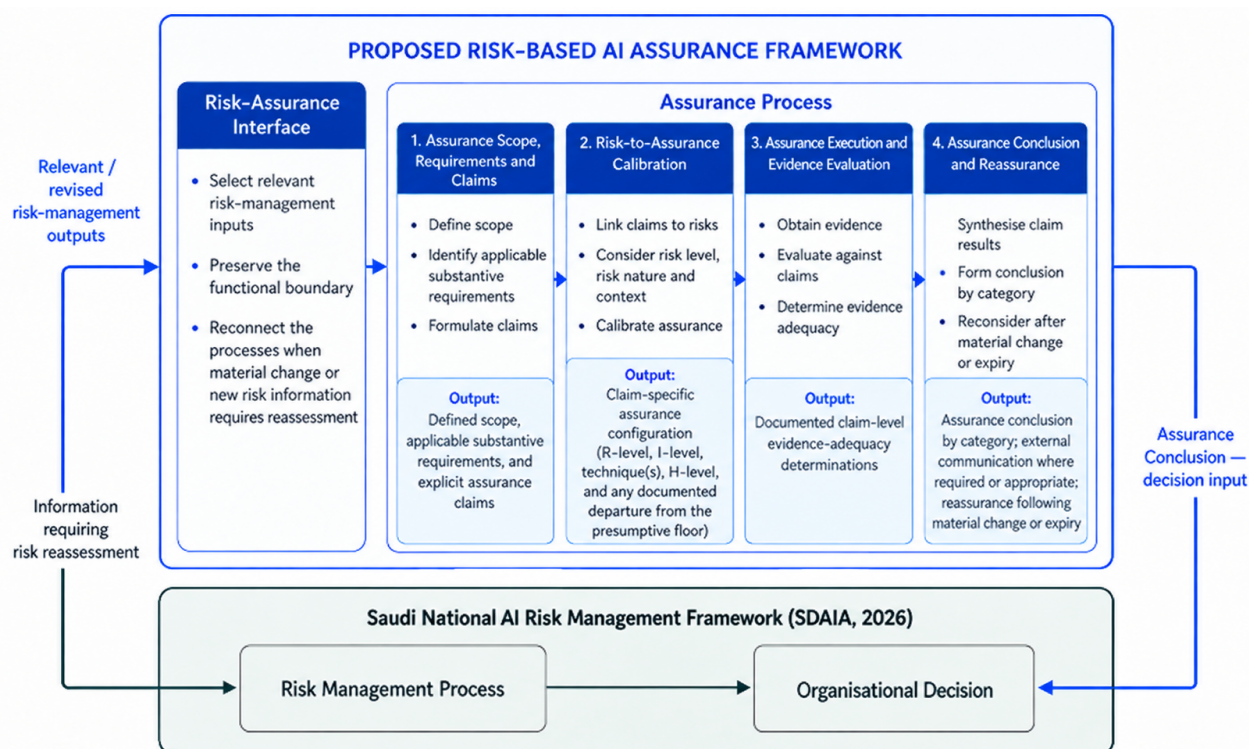


Figure 1. Proposed Risk-Based AI Assurance Framework for the Saudi Public Sector.

Figure note. The lower layer represents the existing Saudi National AI Risk Management Framework (SDAIA, 2026), whose risk-management process generates outputs and informs organisational decision-making. The upper layer represents the proposed framework: the Risk-Assurance Interface and four sequential assurance-process components, producing defined assurance claims, a claim-specific assurance configuration, claim-level evidence-adequacy determinations, and an assurance conclusion by category. Relevant risk-management outputs pass into the interface, information requiring risk reassessment is returned through it, and the assurance conclusion informs, but does not replace, the organisational decision.

Note. International normative instruments may inform the identification of applicable substantive requirements in Stage 1. For example, UNESCO's *Recommendation on the Ethics of Artificial Intelligence* may provide relevant normative reference points, while its *Ethical Impact Assessment* may contribute assurance-relevant evidence in Stage 3. Adopted standards, including ISO/IEC 42001 and ISO/IEC 23894, may serve both functions.

does not own the development record. The number of participating entities should be determined by the guidance owner, provided that the pilot portfolio includes sufficient variation in risk profile, system type, and evidence-access conditions to test the framework's principal design assumptions.

Duration. Long enough for each pilot to include at least one lifecycle event — a model update, a scope extension, a monitoring exception, or a scheduled risk reassessment — because the reassurance provisions in Section 7.5 are otherwise untested.

What the pilot measures. The measures below are stated in advance, and each is paired with the interpretation that would indicate the framework needs redesign rather than refinement.

On effort and cost. This paper offers no estimate of the effort or cost of operating the framework, because none is available: the analysis is documentary and no implementation has been observed. The number of claims a typical system generates, the hours a claim consumes at each rigour level, and the proportion of evidence reusable from existing records are the three parameters that determine whether the framework is affordable at scale, and all three are pilot outputs rather than inputs. An entity considering early adoption should treat the

first pilot as the means of establishing them.

8.4 Use Existing Resources and Build the Capability the Framework Requires

Relevant evidence already generated through risk management should be reused where appropriate. Beyond evidence, the framework depends on access to technical, domain, risk, legal, and governance expertise, including evaluator competence at the calibrated level of independence.

This is the framework's largest unexamined dependency. Levels I3 to I5 assume internal audit functions competent to evaluate AI systems and an external market able to supply independent AI evaluation, neither of which can be assumed at the scale national adoption would require. Three mitigations are available and are not mutually exclusive: staged introduction of the higher independence levels, beginning with systems presenting the highest risk or most significant potential impact; shared or pooled evaluation capacity across entities; and recognition of accredited AI management-system certification as a source of reusable evidence and as one factor that may support evaluator independence where its scope is relevant, without by itself establishing the required indepen-

Table 13. Pilot Measures and Redesign Triggers. *(proposed in this paper)*

Measure	Why it matters	Result indicating redesign, not refinement
Proportion of assurance claims linkable to an existing risk record without creating a new one	Tests the interface premise — that the risk process already holds what assurance needs	A persistently low proportion across diverse pilots would indicate that the interface premise requires reconsideration, including whether existing risk records are the appropriate primary connection point.
Proportion of claims evidenced wholly or partly from existing records	Tests the “build on, A don’t duplicate” premise	persistently low proportion across pilots would weaken the “build on, don’t duplicate” premise and require reconsideration of the expected level of evidence reuse.
Distribution of calibrated rigour levels across claims	Tests whether calibration discriminates	Clustering at a single level regardless of risk and context: the presumptive floors are doing no work and calibration is decorative
Rate of documented departures from presumptive floors, and their direction	Tests whether the floors are set plausibly	Departures on a large majority of claims: the floors are miscalibrated and should be reset from pilot data
Effort per claim, by rigour level	Tests feasibility at scale	Effort per claim that, multiplied by a realistic claim count, exceeds what entities can absorb: the claim granularity is wrong and requirements should be assured in groups
Number of claims per system, by system type	Establishes the scaling parameter absent from this paper	— (this measure exists to be established, not to be passed)
Frequency and cause of interface returns to risk management	Tests whether the two-way path is used or ornamental	No returns across the pilot portfolio should prompt examination of whether the return path is functioning as intended; it should not by itself be treated as evidence of design failure.
Whether external communication, where required or appropriate, can provide meaningful information without exposing security-sensitive, confidential, or personal information.	Tests the feasibility of context-dependent external communication	Repeated inability to communicate meaningful assurance information, where external communication is required or appropriate, without exposing protected or sensitive information would indicate that the communication approach requires redesign.

dence level. Which combination is appropriate is a national policy question, addressed in Section 9.

9 National Policy Recommendations

The proposed framework is intended to complement the Saudi National AI Risk Management Framework rather than establish a parallel risk-management regime. Wider application would benefit from national-level measures that promote consistency in AI assurance while allowing proportionate adaptation across different public-sector contexts.

9.1 Who Should Own National Assurance Guidance

The recommendations in this section require a national owner for AI assurance guidance. The Organizational Arrangements of the Saudi Data and Artificial Intelligence Authority (SDAIA) were issued by Council of Ministers Decision No. 292 dated 27/4/1441H (Council of Ministers, 1441 AH). Against this institutional background, three arrangements are relevant for the ownership of national AI assurance guidance.

1. **The owner of the national risk framework issues companion assurance guidance.** This arrangement places the risk-management framework and the assurance guidance

under the same national owner and facilitates maintenance of the Risk–Assurance Interface. It also requires a clear separation between guidance-setting and claim-specific evaluation where higher levels of evaluator independence apply.

2. **A separate national assurance function is established.** This arrangement separates assurance-framework ownership from ownership of the national risk framework, but would require a defined institutional boundary and coordination mechanism between the two.
3. **Sector regulators issue assurance requirements within common national principles.** This arrangement allows assurance requirements to reflect sector-specific legal, operational, and risk contexts while requiring mechanisms to maintain consistency across government.

This paper proposes SDAIA as the owner and issuer of national AI assurance guidance and of the Risk–Assurance Interface specification, given its existing national role in data and artificial intelligence and its ownership of the National AI Risk Management Framework. This role would concern framework-setting, guidance, and integration; it would not require SDAIA itself to perform claim-specific assurance evaluations. Higher-independence evaluations should remain organisationally separate from framework ownership, through appropriately independent internal or external functions.

9.2 Recommendations

The recommendations below are presented in an order that reflects their dependencies.

1. **Develop national AI assurance guidance.** National guidance should clarify how AI assurance complements the existing risk-management process and how the Risk-Assurance Interface operates, including the transfer of relevant risk-management outputs into assurance and the return of information requiring risk reassessment. It should also state the boundary between an assurance conclusion and an organisational decision, since conflating the two is the failure mode with the most serious consequences.
2. **Establish common principles for proportionate assurance.** National guidance should establish common principles for determining assurance rigour, appropriate assurance techniques, and evaluator independence based on relevant risk information and context. The rigour and independence scales and presumptive floors proposed in Section 7.3 are offered as a starting point for that work, to be reset against pilot data rather than adopted as they stand. These principles should support consistency without creating a fixed relationship between risk categories and predefined assurance levels, and should require that departures from any national floor be documented rather than prohibited.
3. **Standardise core assurance documentation.** Common templates or minimum documentation requirements should support consistent recording of assurance scope, applicable substantive requirements, assurance claims, relevant risk records, calibration decisions and departures, evidence evaluation, and assurance conclusions, while allowing sector-specific adaptation where needed. Appendix B provides a candidate structure. Standardized documentation also supports greater consistency and comparability of assurance conclusions across entities.
4. **Align assurance with existing lifecycle mechanisms.** National guidance should link reconsideration of assurance to existing monitoring and risk-reassessment mechanisms. Material changes in the system, context, applicable requirements, relevant risks, or supporting evidence should prompt proportionate reconsideration of assurance where required, and claim expiry should be treated as a trigger in its own right.

A fifth measure follows from Section 8.4 and is properly a national rather than an entity-level matter: address evaluator capacity explicitly, whether through staged introduction of the higher independence levels, pooled evaluation capability across entities, recognition of accredited certification as a potential source of reusable evidence and as one factor that

may support evaluator independence where its scope is relevant, or a combination. Assurance requirements that outrun the capacity to meet them produce documentary compliance rather than justified confidence, which is the outcome this framework exists to avoid.

10 Conclusion

Saudi Arabia already has a substantial national foundation for managing AI risk through the National AI Risk Management Framework. The framework establishes a structured process for defining context, identifying and assessing risks, applying treatment, and supporting monitoring, reassessment, and organisational decision-making (SDAIA, 2026). The principal need identified in this paper is therefore not for an additional risk-governance framework, but for an explicit and operationalised mechanism for Risk-Assurance Integration that connects relevant risk-management outputs to a distinct assurance process without duplicating risk assessment.

To address this need, the paper proposes a Risk-Based AI Assurance Framework comprising a Risk-Assurance Interface and four assurance-process components: Assurance Scope, Requirements, and Claims; Risk-to-Assurance Calibration; Assurance Execution and Evidence Evaluation; and Assurance Conclusion and Reassurance. Together, these components use relevant risk-management outputs to define what must be assured, calibrate claim-specific assurance configurations, evaluate the adequacy of supporting evidence, and produce an evidence-based assurance conclusion that informs, but does not replace, organisational decision-making.

The proposed approach complements the Saudi National AI Risk Management Framework rather than creating a parallel risk-management regime, keeping risk management and assurance functionally distinct while operating in an integrated and iterative relationship across the AI lifecycle. It has not been implemented or empirically validated, and its practical operation should be tested through the phased public-sector pilots specified in Section 8.3, including the measures and redesign triggers stated there, before wider adoption.

11 Limitations

The following limitations apply to the analysis and to the proposed framework. They are stated together rather than dispersed, and they should be read as qualifying every finding in Sections 4 to 7.

Limitations of the method.

1. **Single-analyst coding.** The content analysis was performed by one analyst. No second coder was used, no

inter-coder agreement was computed, and the coding was not externally audited. The mitigation is transparency and traceability rather than reliability statistics: the coding rules and the retained coding and traceability record are available from the author on request, and the source locators in Table 5 allow relevant findings to be re-examined.

2. **Documents only.** The analysis reads published institutional documents. It does not observe, audit or assess implementation in any jurisdiction, and it therefore cannot say how any of the four international practices, or the Saudi framework, operates in fact.
3. **Negative findings are bounded by what is published.** Where the analysis reports that the Saudi framework does not explicitly operationalise an assurance function, that finding concerns the published document at the locators recorded in Table 5. Implementation guidance, annexes, sector instructions or internal material not publicly available may address these functions; if they do, the corresponding observations do not hold. The author has no access to non-public material and makes no claim about it.
4. **A focused comparative corpus.** Five documents from four comparative settings: three OECD member countries and the European Union. No comparator was included from a governance context closer to Saudi Arabia's.
5. **Transferability is assumed, not demonstrated.** The design lessons carried into Section 6 are treated as potentially transferable to the Saudi context, but that transferability has not been empirically tested.
6. **Source overlap in conceptual development.** Some documents that informed the original analytical lens were later included in the comparative corpus. Table 2 shows that each of the six dimensions can be independently grounded in sources outside that corpus. The overlap is therefore acknowledged as a limitation of source independence, but not treated as analytical circularity.
7. **Standards not analyzed as cases.** The international standards discussed in Section 2.3 inform the design but were not analysed through the six dimensions, so the paper cannot say how their assurance logic compares with the four governmental practices.

Limitations of the proposed framework.

1. **Not implemented, not validated.** The framework has never been operated. No claim is made that it improves assurance outcomes, reduces effort, or produces conclusions that would withstand review.
2. **The calibration scales are design judgements.** The rigour levels, independence levels and presumptive floors in Sec-

tion 7.3 are proposed by the author to make calibration operable. They are not derived from any reviewed practice and are not validated. Their plausibility is the specific subject of two of the pilot measures in Table 13.

3. **No effort, cost or scaling estimates.** The paper cannot say how many claims a typical system generates, what a claim costs to assure at each rigour level, or what proportion of evidence is reusable. Section 8.3 treats all three as pilot outputs.
4. **Capacity is assumed.** Independence levels I3–I5 depend on evaluator capability, including internal functions competent to assess AI systems and an external market capable of providing sufficiently independent evaluation. The availability of that capability at national scale has not been established.
5. **Stakeholder participation was absent from the design.** No public-sector entity, affected community, or practitioner group was consulted in developing the framework. Its usability is therefore asserted from documentary analysis rather than established with the people expected to operate it.
6. **External communication remains context-dependent and untested.** Where assurance outcomes are communicated outside the organization, the appropriate content and level of disclosure may be constrained by security, confidentiality, data-protection, legal, or operational considerations. The framework does not assume mandatory public disclosure in all cases.
7. **The complementary-process design carries a duplication risk.** Positioning assurance alongside rather than inside the risk framework (Section 2.6) risks entities operating two processes where one would serve. Sections 8.1 and 8.2 mitigate this; Table 13 measures it.
8. **Limited treatment of AI systems procured rather than built.** Where an entity does not hold the development record, evidence access differs materially. The framework recognizes this issue and includes procured systems within the pilot scope, but does not yet specify a complete approach for addressing evidence-access constraints in such cases.

Declarations

Funding

This work received no external funding. It was conducted by the author independently and without financial support from any organisation, and no organisation had any role in the design, analysis, drafting or decision to publish.

Conflict of Interest and Independence

The author is an independent researcher working in AI governance. The author declares the following in relation to this paper: no current or recent employment, consultancy, advisory, contractual or financial relationship with the Saudi Data and Artificial Intelligence Authority or with any Saudi public-sector entity in respect of AI assurance or AI risk management; and no financial interest in any product, service, certification scheme or advisory business that would be commercially advantaged by adoption of the recommendations in Sections 8 and 9.

Institutional Endorsement

This paper is an independent, unsolicited analysis. It was not commissioned, funded, reviewed or endorsed by the Saudi Data and Artificial Intelligence Authority or by any other public body (other than review for publication by the ICAIRE Scientific Committee), and it should not be read as representing the position of any institution. The characterisation of the National AI Risk Management Framework in Section 5 is the author's own reading of the published document, bounded as stated in Section 11.

No institutional logo, official mark, or other identifier belonging to any organisation appears in this paper. Figure 1 refers to the National AI Risk Management Framework by name and citation only.

Ethics and Data Protection

This study involved no human participants and no collection or processing of personal data. The corpus consisted exclusively of published institutional documents listed in Table 3 and the References. No research-ethics approval was required or sought. Should the framework be piloted as recommended in Section 8.3, any collection or processing of personal data in that context would require separate ethics and data-protection review under the applicable national requirements.

Declaration of Generative AI Use

During the preparation of this work, the author used ChatGPT (OpenAI) to assist with manuscript structuring, language refinement, review for clarity and internal consistency, drafting support for parts of Sections 2.3, 2.4, 2.6, 7.3, 7.5, 8.2, 8.3, 9.1 and 11 and of the appendices, and the initial visual rendering of Figure 1. The figure structure and content were defined by the author and subsequently reviewed and manually edited. All AI-assisted outputs were reviewed and edited by the author, who takes full responsibility for the final content, including the accuracy of every citation, the characterisation

of every cited source, and every design proposal advanced in Section 7.

Supplementary documents

The manuscript's two appendices are published as separate supplementary documents accompanying this paper:

- **Appendix A.** Worked Example — One Assurance Claim End to End.
- **Appendix B.** Assurance Claim Register — Template.

References

Group authors are listed with the abbreviation used in the text, so that every in-text citation resolves directly. The list is alphabetised by the form used in the text.

Ashmore, R., Calinescu, R., & Paterson, C. (2021). Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys*, 54(5), Article 111, 1–39.
<https://doi.org/10.1145/3453444>

Australian Government et al. (2024). *National framework for the assurance of artificial intelligence in government* (Version 1.0). Australian Government.
<https://www.finance.gov.au/sites/default/files/2024-06/National-framework-for-the-assurance-of-AI-in-government.pdf>

Centre for Data Ethics and Innovation [CDEI]. (2021). *The roadmap to an effective AI assurance ecosystem*. UK Government.
<https://www.gov.uk/government/publications/the-roadmap-to-a-n-effective-ai-assurance-ecosystem>

Council of Ministers. (1441 AH). *Organizational arrangements of the Saudi Data and Artificial Intelligence Authority* (Decision No. 292, 27/4/1441H). National Center for Archives and Records.
<https://ncar.gov.sa/document-details/eyJpdil6lnpJZnhXODlFVUFVNE1qZlB1WlVsVHc9PSIsInZhbHVlIjoieY2tSNG81UkErYoV5S1dwSohkcXJYUTo9IiwibWFJfjIjoINTZiN2RjOTA5ODRkNzUyYjUxMTQxMGVjYjTlZTBmMDRmNzUwMjVmOWI1NWVlYjViMWI3NjRhZDczNWlZNlFhMSIsInRhZyl6IiJ9>

Department for Science, Innovation and Technology [DSIT]. (2024). *Introduction to AI assurance*. UK Government. <https://www.gov.uk/government/publications/introduction-to-ai-assurance>

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*, L, 2024/1689.
<https://eur-lex.europa.eu/eli/reg/2024/1689/oj>

Guo, Y., Foster, J., Buckley, A. R., & Zoukas, G. (2026). Assuring trustworthy AI: An integrative review of hierarchical and relational assurance approaches. *International Journal of Information Management Data Insights*, 6(1), 100415. <https://doi.org/10.1016/j.jiimei.2026.100415>

- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288. <https://doi.org/10.1177/1049732305276687>
- International Organization for Standardization & International Electrotechnical Commission [ISO/IEC]. (2022). *Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations* (ISO/IEC 38507:2022).
- International Organization for Standardization & International Electrotechnical Commission [ISO/IEC]. (2023a). *Information technology — Artificial intelligence — Management system* (ISO/IEC 42001:2023).
- International Organization for Standardization & International Electrotechnical Commission [ISO/IEC]. (2023b). *Information technology — Artificial intelligence — Guidance on risk management* (ISO/IEC 23894:2023).
- International Organization for Standardization & International Electrotechnical Commission [ISO/IEC]. (2025). *Information technology — Artificial intelligence (AI) — AI system impact assessment* (ISO/IEC 42005:2025).
- Muhammad, A. E., & Yow, K.-C. (2026). Risk-based AI assurance framework. *Information*, 17(3), 263. <https://doi.org/10.3390/info17030263>
- National Telecommunications and Information Administration [NTIA]. (2024). *AI accountability policy report: Earned trust through AI system assurance*. U.S. Department of Commerce. <https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report>
- Porter, Z., Habli, I., McDermid, J., & Kaas, M. H. L. (2024). A principles-based ethics assurance argument pattern for AI and autonomous systems. *AI and Ethics*, 4, 593–616. <https://doi.org/10.1007/s43681-023-00297-2>
- Robbins, D., Eris, O., Kapusta, A., Booker, L., & Ward, P. (2024). *AI assurance: A repeatable process for assuring AI-enabled systems*. MITRE. <https://www.mitre.org/news-insights/publication/ai-assurance-repeatable-process-assuring-ai-enabled-systems>
- Saudi Data and Artificial Intelligence Authority [SDAIA]. (2026). *National AI risk management framework* (Version 1.0, Document No. SDAIA-P145EN). Retrieved August 13, 2026, from <https://sdaia.gov.sa/en/MediaCenter/KnowledgeCenter/Pages/SDAIPublications.aspx>
- Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>
- Thomas, C., Roberts, H., Mökander, J., Tsamados, A., Taddeo, M., & Floridi, L. (2025). The case for a broader approach to AI assurance: Addressing “hidden” harms in the development of artificial intelligence. *AI & Society*, 40, 1469–1484. <https://doi.org/10.1007/s00146-024-01950-y>
- Treasury Board of Canada Secretariat [TBCS]. (2025a). *Directive on automated decision-making*. Government of Canada. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>
- Treasury Board of Canada Secretariat [TBCS]. (2025b). *Guide to peer review of automated decision systems*. Government of Canada. <https://publications.gc.ca/site/eng/9.946875/publication.html>
- United Nations Educational, Scientific and Cultural Organization [UNESCO]. (2022). *Recommendation on the ethics of artificial intelligence*. <https://unesdoc.unesco.org/ark:/48223/pfo000381137>
- United Nations Educational, Scientific and Cultural Organization [UNESCO]. (2023). *Ethical impact assessment: A tool of the Recommendation on the Ethics of Artificial Intelligence*. <https://unesdoc.unesco.org/ark:/48223/pfo000386276>

Copyright and publication

© 2026 Nawaf Alsahmah. Published by the International Center for Artificial Intelligence Research and Ethics (ICAIRE), Riyadh, Saudi Arabia. All rights reserved.

Review

This paper was reviewed by the ICAIRE Scientific Committee and approved for publication by the Committee Chair, subject to revisions, which were incorporated by ICAIRE in accordance with the ICAIRE Research Policy.

Disclaimer

The views expressed are those of the author and do not necessarily reflect the positions of ICAIRE or UNESCO.

Suggested citation

Alsahmah, N. (2026). Risk-based AI assurance in the Saudi public sector: Aligning international assurance practices with the National AI Risk Management Framework (ICAIRE Framework Paper). International Center for Artificial Intelligence Research and Ethics. <https://doi.org/10.68480/CEYS4608>