Building Trust in Government AI: Redesigning the NSW AI Assurance Framework

The NSW Government has released a redesigned Artificial Intelligence Assessment Framework (AIAF) to strengthen the safe and responsible use of AI across public services. The NSW team led the redesign and CSIRO's Data61 contributed scientific and technical methods to enhance specific components of the framework, improving risk assessment logic, validation, and usability in line with global standards.

A next-generation framework for responsible AI

Artificial Intelligence (AI) is rapidly transforming how governments deliver services and make decisions. As agencies explore new use cases of AI, from automating some administrative tasks to supporting critical decisions in health, policing, and social services, the need for a consistent, use case-based approach to managing AI risk has never been greater. Agencies need practical tools that identify and mitigate risks early without overwhelming non-technical users or duplicating existing compliance processes.

Many existing risk assessment frameworks face practical challenges. Assessments are often manual and time-consuming, requiring significant expert interpretation and occasionally producing inconsistent outcomes across similar use cases. Although the original AIAF tried to address these challenges, some limitations remained. Subject Matter Experts (SMEs) were often involved too early or too frequently, stretching limited capacity and diverting effort from high-value assessments. In addition, fragmented governance pathways meant that privacy, cybersecurity, and procurement reviews were conducted separately, creating duplication and slowing delivery. These factors made it difficult for agencies to apply the framework efficiently or at scale.

Working from a model first developed by the NSW Government, which introduced the shift to intrinsic risk assessment and moved away from traditional likelihood-versus-impact methods, Data61 collaborated to validate and extend the design. The partnership strengthened the scientific foundations of the framework, refining its logic, calibrating its scoring mechanisms, and testing its performance across diverse AI contexts. The resulting NSW AIAF is a dynamic, multi-layered framework that enables agencies to identify, classify and mitigate AI-related risks in line with international standards and NSW policy obligations, supporting early detection of high-risk or critical cases, proportionate mitigations, and accountable oversight across the AI lifecycle.

This work supports government agencies to manage AI risks confidently and consistently, ensuring responsible innovation that earns and sustains public trust.

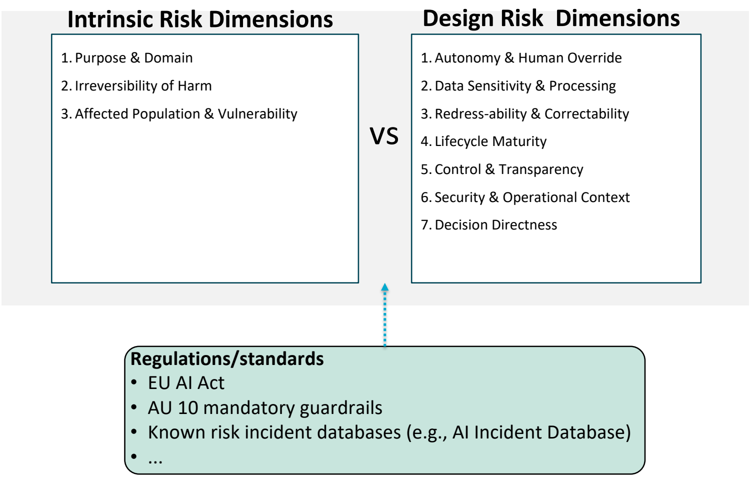
Separating intrinsic risk from design choices

One of our key contributions for the new framework is its clear separation between intrinsic risks and design risks.

Intrinsic risks, those that exist regardless of whether a task is performed by AI, traditional software, or human decision-making, represent baseline risks in any system. By recognising these shared risk foundations, the AIAF can be applied bidirectionally: as an AI-focused framework or as an extension of broader digital assurance processes.

Design risks, meanwhile, arise from specific design choices or implementation decisions.

Intrinsic risks include areas that inherently involve health, safety, or fundamental rights, or where harms could be irreversible. These are treated as high-risk by default. Design risks, such as the level of autonomy, hosting arrangements, or the presence of redress mechanisms, can be reduced through stronger design and governance controls.


Intrinsic Risk Dimensions
1. Purpose & Domain
2. Irreversibility of Harm
3. Affected Population & Vulnerability
VS
Design Risk Dimensions
1. Autonomy & Human Override
2. Data Sensitivity & Processing
3. Redress-ability & Correctability
4. Lifecycle Maturity
5. Control & Transparency
6. Security & Operational Context
7. Decision Directness
Regulations/standards
• EU AI Act
• AU 10 mandatory guardrails
• Known risk incident databases (e.g., AI Incident Database)
• ...

This distinction helps agencies understand that while intrinsic risks cannot be eliminated, the overall system risk (intrinsic + design) can be reduced through thoughtful design. It allows domain experts to focus on whether an AI use case should be pursued at all, while technical experts work on making its design safer. By separating these dimensions, the framework avoids conflating distinct issues and supports clearer, more accountable decisions.

A layered approach to assurance

Managing AI risk in government is complex. No single assessment method can identify all potential risks correctly. Rule-based approaches offer consistency but can miss context; score-based systems can become inconsistent if thresholds are not well calibrated or adjusted for context.

To address this, the redesigned AIAF uses a multi-layered structure that blends rule-based and score-based logic: Intrinsic-risk layer, pattern-matching layer, safety-net layer, and score-based layer.

Intrinsic-risk and pattern-matching layers identify high-risk combinations (for example, use of sensitive data without redress). Safety-net layer catches potentially missed high risks and identifies new design-risk patterns. Scoring layer then uses cumulative score for risk classification and compares the results with prior layers to detect edge cases based on mismatch and anomaly pattern analysis.

This layered design ensures that if one logic misses a risk, another will likely detect it. It allows thresholds to be adjusted, combining precision with flexibility. The result is a framework that provides a deeper, more explainable assessment of AI risk and increases confidence that major risks will not go unnoticed.

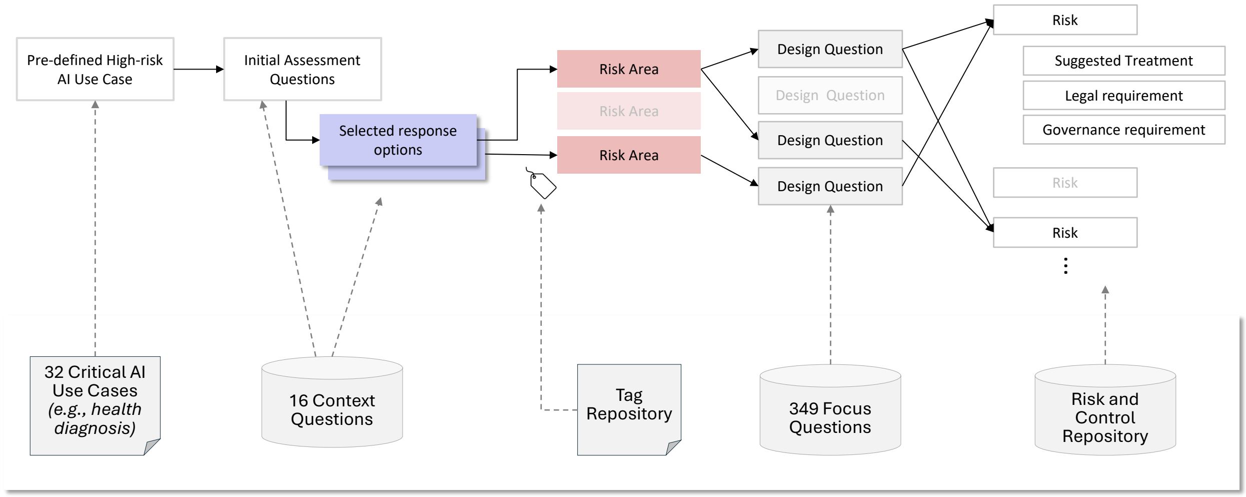
Dynamic navigation of complex risks

Government AI systems operate in diverse and evolving contexts. A static checklist approach often fails to account for how multiple risk factors interact. The redesigned AIAF addresses this by dynamically interpreting thousands of possible risk combinations drawn from regulatory standards, ethical frameworks, and NSW policy controls.

Instead of a long checklist, a concise series of questions recognises how risk factors interact, for example, how autonomy, external hosting, and lack of human oversight together may elevate overall risk. This approach reflects how risks amplify each other in practice, producing a more realistic assessment.

More importantly, the dynamic structure of the framework helps practitioners focus on what truly matters. In traditional frameworks, long lists of risks and controls often overwhelm assessors, many of which may not apply to a specific context. The AIAF's dynamic logic narrows this scope, filtering out irrelevant risks and surfacing only those applicable to a system's intrinsic characteristics, design choices, and operating environment. This ensures that assessments remain both comprehensive and efficient.

Data61 contributed to developing and validating this question logic, ensuring that the right mix of intrinsic and design-choice questions captures underlying risk conditions rather than narrowly checking for specific safeguards. To manage complexity and keep the framework adaptable, we avoided control-style questions (e.g. "Have you used X safeguard?").



Instead, we focused on assessing underlying risk conditions and risk amplifiers and providing relevant treatments and policy requirements.

This approach keeps the assessment tool flexible and evidence-based, while avoiding duplication with established governance frameworks.

Usability by design

A persistent challenge in AI governance is usability. AI risk frameworks can be comprehensive yet difficult for non-experts to apply consistently. To address this, the framework was designed to be practical for non-technical users across government.

It begins with a short set of yes/no context questions that quickly establish a baseline risk profile. Based on the user answers, the tool automatically generates focused follow-up questions linked to relevant NSW laws and policies.

This structure allows assessors to draw on existing subject-matter expertise and governance processes, such as privacy impact assessments, cybersecurity attestations, procurement reviews, and digital assurance checkpoints without duplicating effort.

Through usability testing with non-technical agency staff and subject-matter experts, we confirmed that this structure reduces assessment fatigue while improving clarity and traceability. The outcome is a framework that lowers barriers to responsible AI adoption while maintaining scientific and policy rigour.

Evidence-based validation

Assessing whether the redesigned framework works as intended required large-scale validation. The AIAF logic was tested on more than 2,400 AI use cases, including 20 reference cases drawn from real NSW government AI use cases, examples listed in the EU AI Act, and documented incidents from AI risk databases. These reference cases served as a "control group" to verify that the framework correctly identified known high-, medium-, and low-risk AI cases.

Testing confirmed that the multi-layered logic consistently surfaces high-risk cases and aligns with expected outcomes. The framework demonstrated strong sensitivity to critical risk factors while maintaining interpretability for auditors and policy teams.

This validation demonstrated that the layered approach not only improves accuracy but also reveals where contextual judgment or expert review is still needed, supporting transparent, defensible decision-making across agencies. It also strengthens confidence that the framework can be applied consistently across diverse contexts while remaining adaptable to emerging AI systems. It provides a practical balance between automation and human oversight, essential for building sustained trust in government AI.

A collaborative achievement

The redesign of the NSW AIAF was a collaborative effort that built on a model first developed by the NSW Government, which created the core concept, intrinsic-risk methodology, and overall assurance architecture. CSIRO's Data61 partnered to validate and extend this foundation, applying formal methods, refining the logic, strengthening the scoring and calibration model, and testing the framework across a wide range of AI contexts.

Together, this work produced a rigorously tested, operational assurance system that offers agencies a clear and transparent way to evaluate AI systems, plan proportionate mitigations, and demonstrate accountability throughout the AI lifecycle.

By combining government-led design with scientific validation, the collaboration illustrates how policy innovation and technical evidence can be integrated to deliver trustworthy, practical governance tools. It shows that responsible AI practice is strongest when governments and researchers work together to translate principles into robust, real-world mechanisms that support safe and confident adoption.