kjr.com.au

AI Safety:
Emerging Policy, Governance and Assurance

Dr Kelvin Ross

Australian System Safety Conference 2024, Brisbane, 24 Oct 2024


AP

AI Safety: Emerging Policy, Governance, and Assurance

Since the mainstream adoption and rapid growth of GPT and Large Language Models (LLMs) in late 2022, unprecedented concerns about the responsible use of AI have emerged, reverberating from industry to the broader community and government. This shift has accelerated the creation and adoption of policies, standards, and regulations aimed at ensuring AI safety, driving a global dialogue on the ethical and secure deployment of AI systems.

Leading nations are now addressing these concerns by establishing dedicated institutes and initiatives focused on AI safety, designed to deepen the understanding of AI's capabilities, limitations, and inherent risks. The rapid evolution of standards for governance and assurance, combined with the rise of mandated policies, underscores the urgent need to address AI's role in both safety-critical systems and broader societal impacts.

In this keynote, Dr. Kelvin Ross will provide an in-depth overview of emerging AI safety policies and governance frameworks, with a particular focus on trends in Australia. This discussion will explore the influences of international efforts and cross-industry collaboration on shaping a responsible, secure future for AI technology. Additionally, Dr. Ross will examine the role of disruptive technologies in redefining assurance and governance standards, offering insights into how industries can navigate this fast-evolving landscape.

KJR est 1997

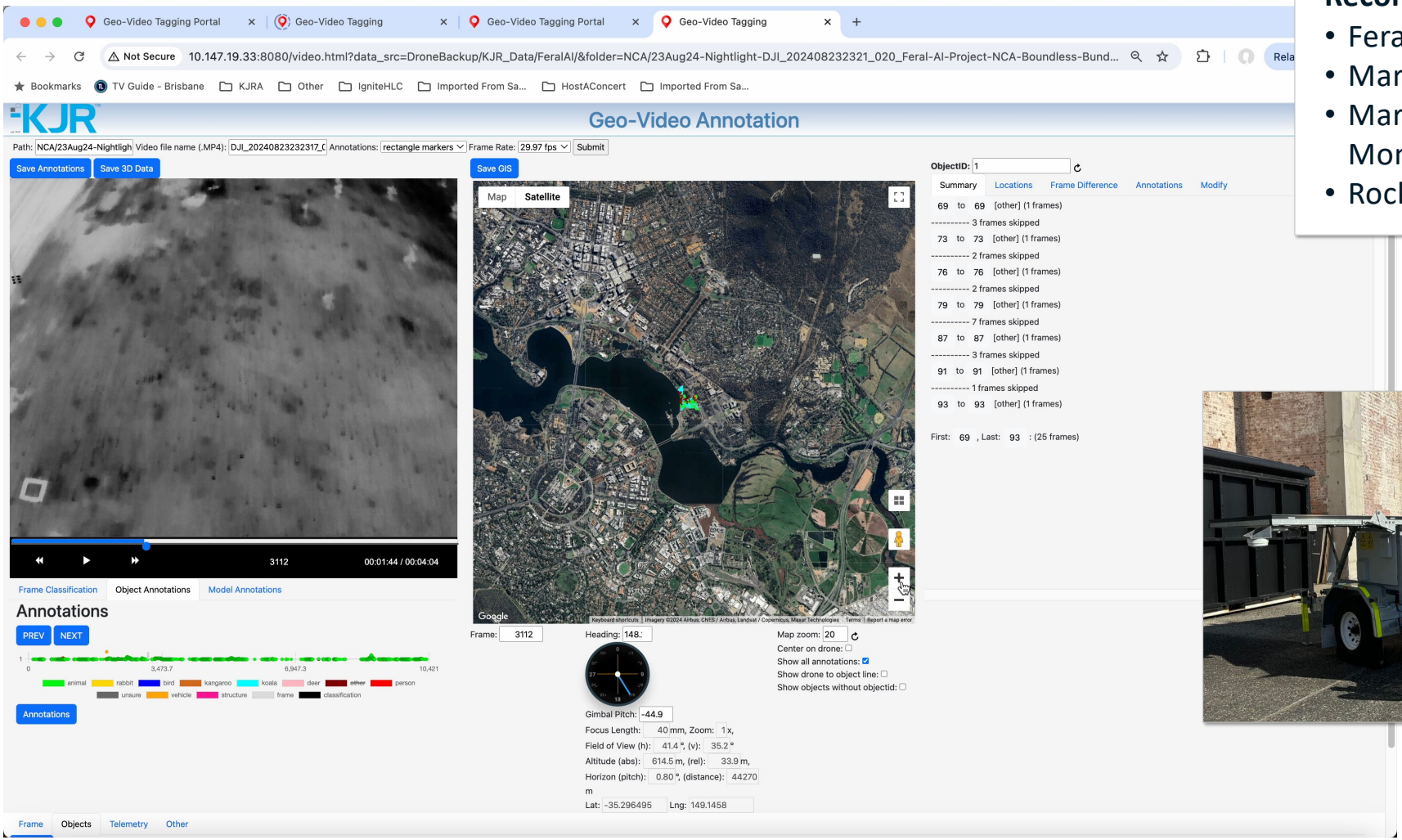
Dr Kelvin Ross is an entrepreneur, technologist and researcher. He currently holds a number of roles, including Founder & MD of KJR, a mid-tier IT consultancy, Adjunct Associate Professor in Intelligent and Integrated Systems at Griffith University, and Director at IntelliHQ, a non-profit innovation Centre focused on Artificial Intelligence in healthcare at Gold Coast University Hospital.

He has over 25 years of experience in advanced technology commencing with safety-critical systems in the military, then moving on to transportation, banking, financial markets, government and healthcare systems. He has participated in several successful and unsuccessful technology startups, as well numerous successful and unsuccessful technology implementation programmes in medium and large enterprises.

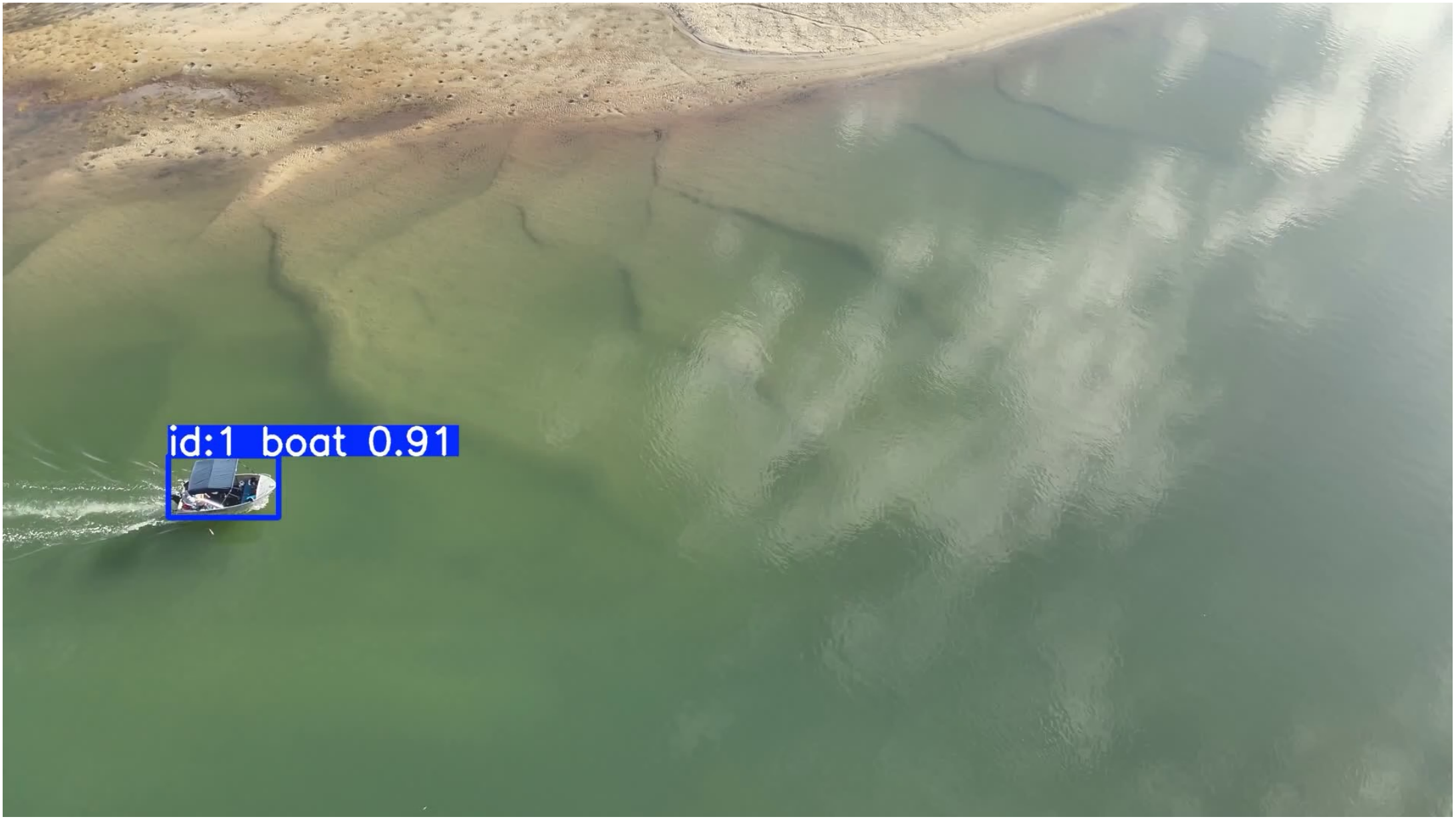
Computer Vision: Video Object Tracking

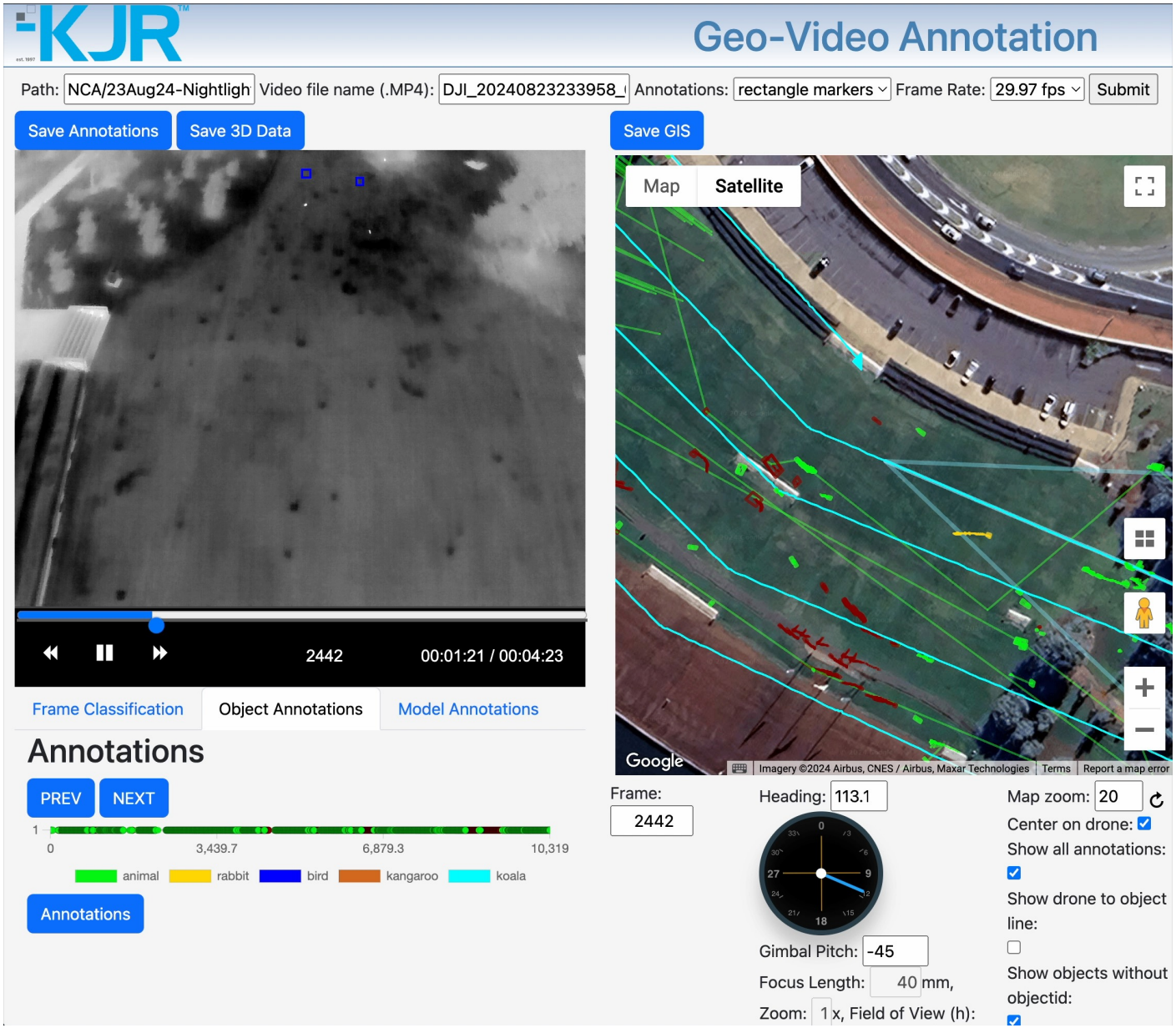
Drone Intelligence Surveillance and Reconnaissance (ISR)

- FeralAI
- Maritime Surveillance
- Marine Debris Monitoring
- Rock Art Conservation



KJR est 1997


id:1 boat 0.91


KJR
Geo-Video Annotation
Path: NCA/23Aug24-Nightligh
Video file name (.MP4): DJI_20240823233958_
Annotations: rectangle markers
Frame Rate: 29.97 fps
Submit
Save Annotations
Save 3D Data
Save GIS
2442
00:01:21 / 00:04:23
Map
Satellite
Google
Imagery ©2024 Airbus, CNES / Airbus, Maxar Technologies
Terms
Report a map error
Frame Classification
Object Annotations
Model Annotations
Annotations
PREV
NEXT
1
0
3,439.7
6,879.3
10,319
animal
rabbit
bird
kangaroo
koala
Annotations
Frame:
2442
Heading: 113.1
Gimbal Pitch: -45
Focus Length: 40 mm,
Zoom: 1 x, Field of View (h):
Map zoom: 20
Center on drone:
Show all annotations:
Show drone to object line:
Show objects without objectid:

Confidence range

0.40 1.00

0.00 1.00

Object size

0.00 10000.00

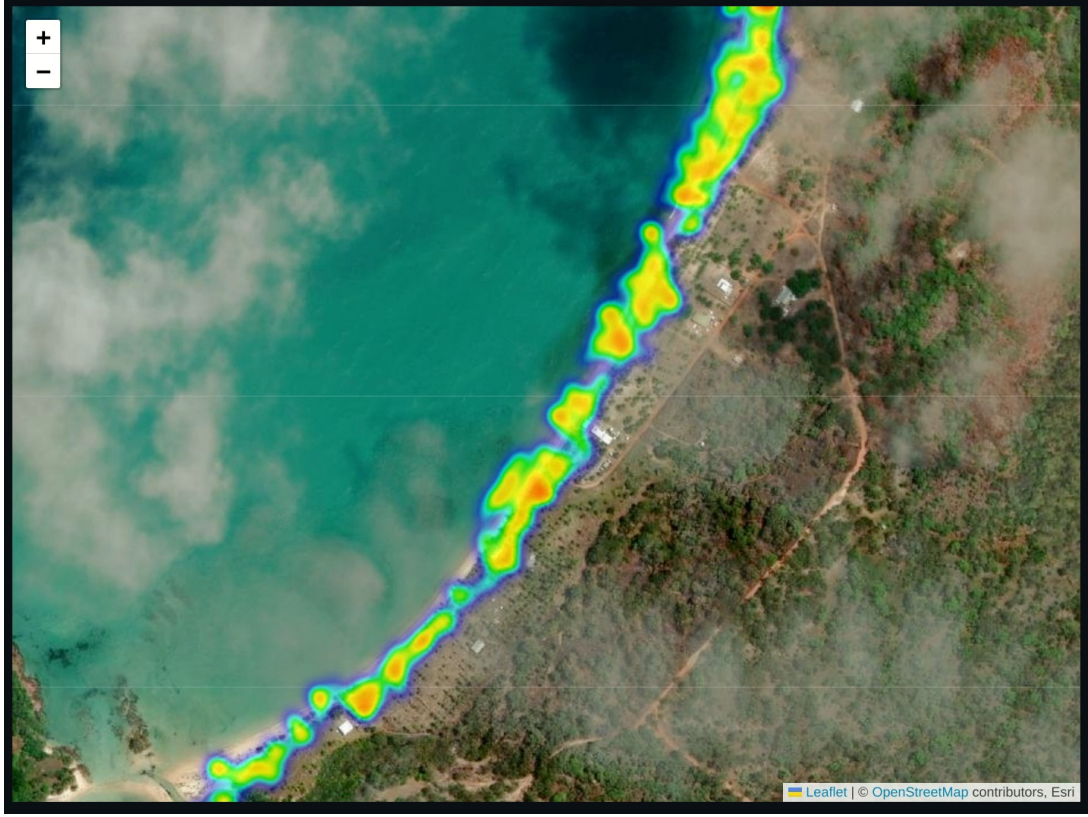
0.00 10000.00

Drone path

- (•) Hide
- ( ) Show

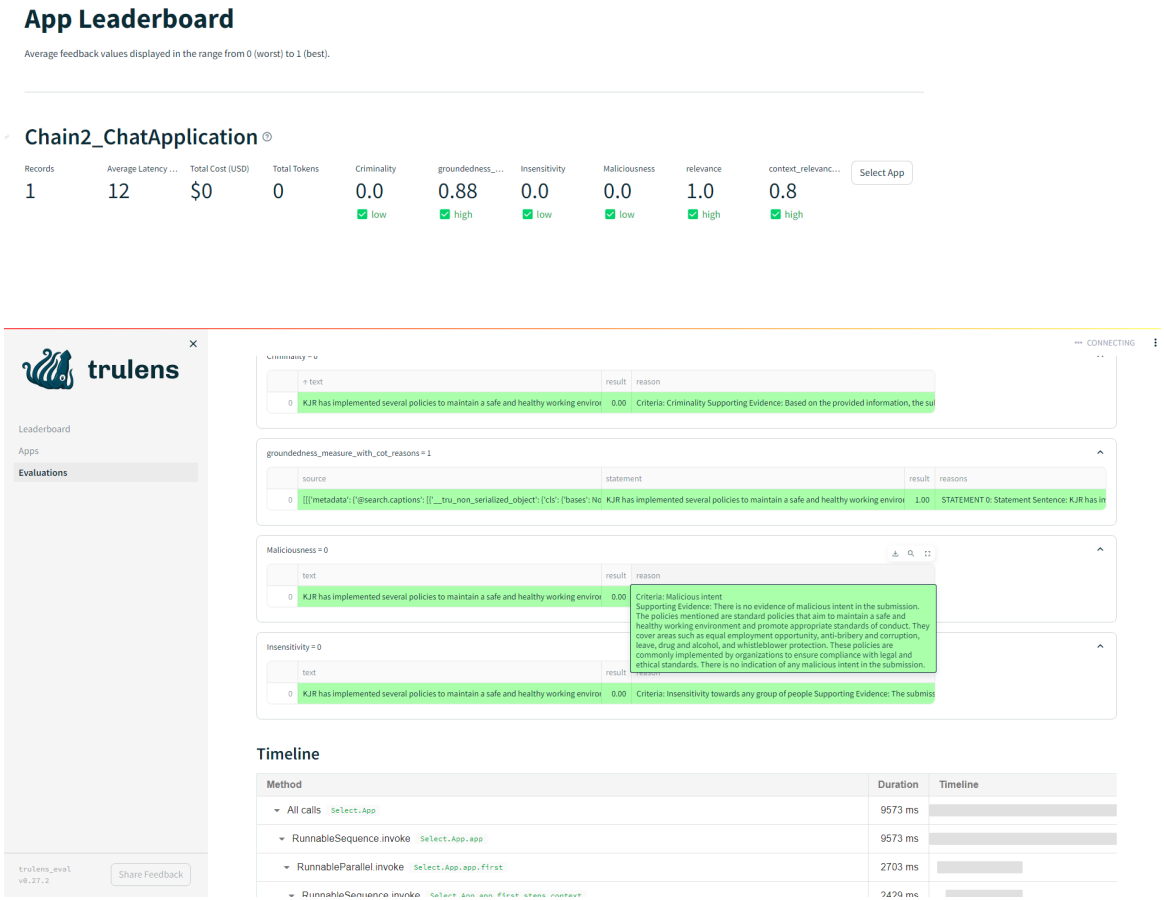
Mission

- (•) after
- ( ) before



KJR est 1997

LLM Automated Test Frameworks



- Automated evaluation
  - Targeted to your specific task
  - Can scale to 1000s of tests
  - Can be run repeatedly over time
  - Can cover functional correctness & safety
  - Provides an audit trail

https://www.trulens.org

KJR est. 1997

AI Project & Analysis Software

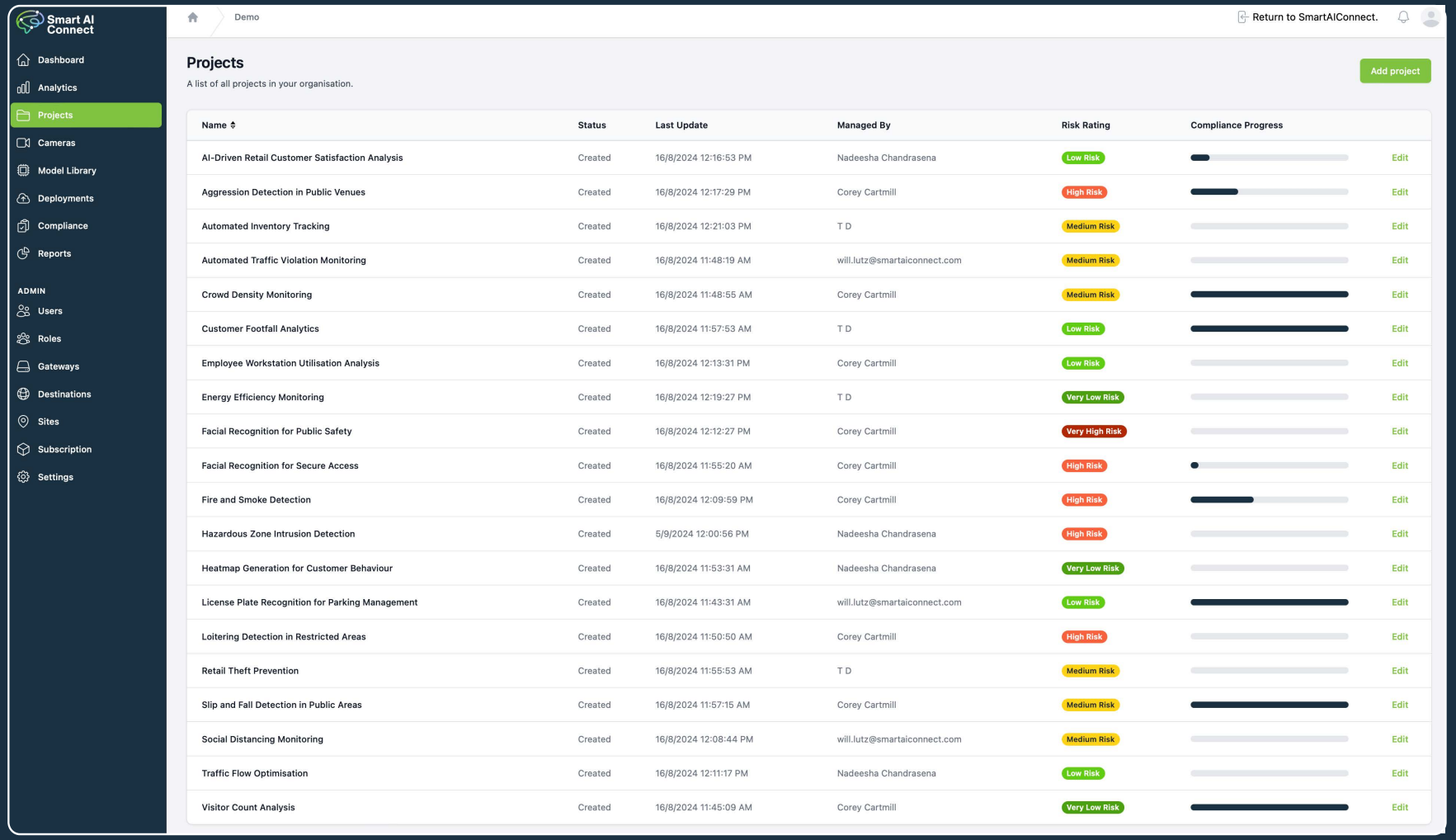
Monitor
identifies AI risks on the network

Manage (AI Project)
eliminates AI threats in the network, controls version deployments and collects data

Protect (AI Project)
ensures governance, risk and compliance processes to meet ISO and regulatory standards.

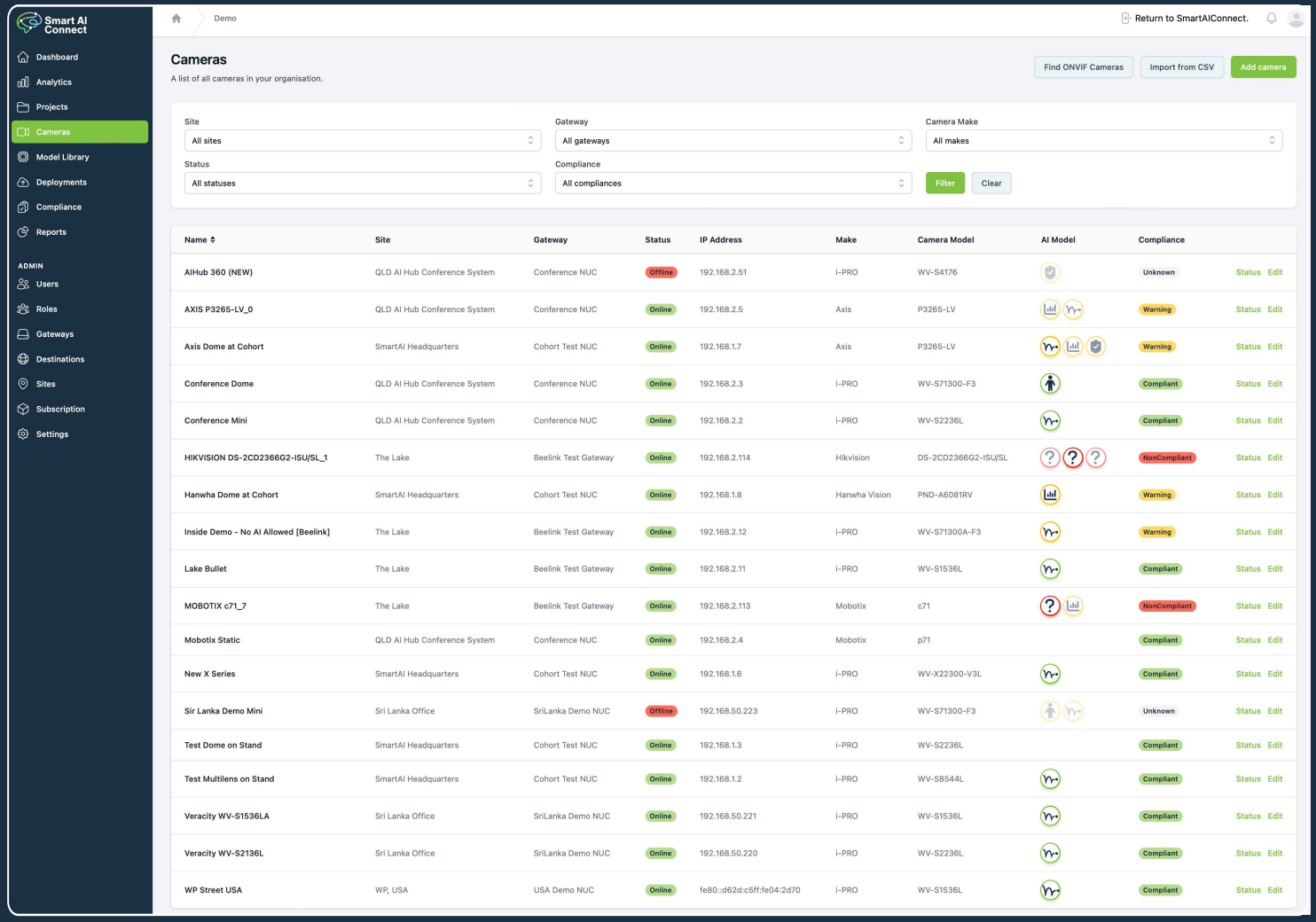
SmartAIConnect Responsible AI Framework supports mulitple manufactures including:
***i-PRO, Axis, Hanwha and Mobotix** cameras.*

AI Projects Overview

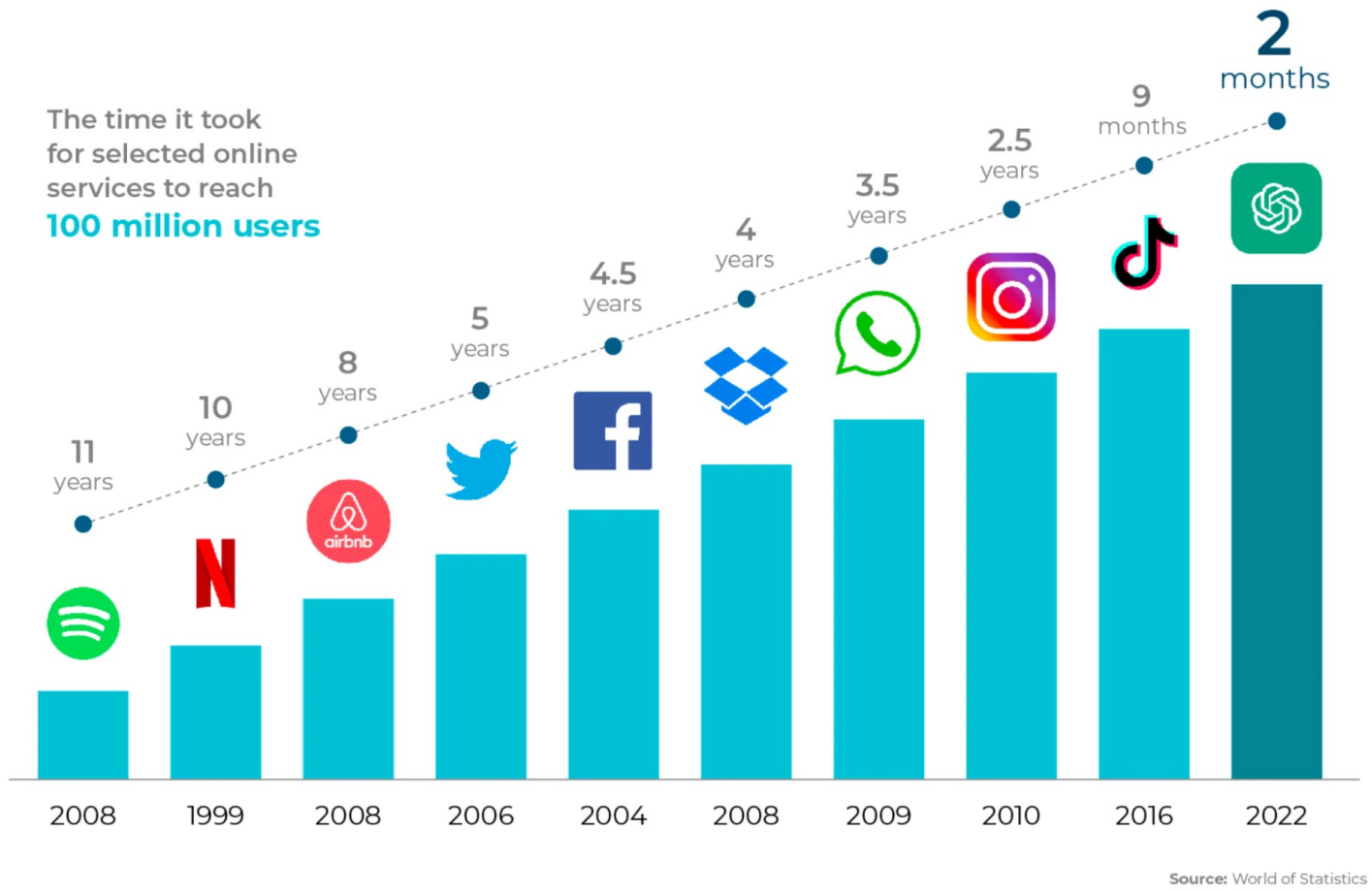


Name	Status	Last Update	Managed By	Risk Rating	Compliance Progress	
AI-Driven Retail Customer Satisfaction Analysis	Created	16/8/2024 12:16:53 PM	Nadeesha Chandrasena	Low Risk		Edit
Aggression Detection in Public Venues	Created	16/8/2024 12:17:29 PM	Corey Cartmill	High Risk		Edit
Automated Inventory Tracking	Created	16/8/2024 12:21:03 PM	T D	Medium Risk		Edit
Automated Traffic Violation Monitoring	Created	16/8/2024 11:48:19 AM	will.lutz@smartaiconnect.com	Medium Risk		Edit
Crowd Density Monitoring	Created	16/8/2024 11:48:55 AM	Corey Cartmill	Medium Risk		Edit
Customer Footfall Analytics	Created	16/8/2024 11:57:53 AM	T D	Low Risk		Edit
Employee Workstation Utilisation Analysis	Created	16/8/2024 12:13:31 PM	Corey Cartmill	Low Risk		Edit
Energy Efficiency Monitoring	Created	16/8/2024 12:19:27 PM	T D	Very Low Risk		Edit
Facial Recognition for Public Safety	Created	16/8/2024 12:12:27 PM	Corey Cartmill	Very High Risk		Edit
Facial Recognition for Secure Access	Created	16/8/2024 11:55:20 AM	Corey Cartmill	High Risk		Edit
Fire and Smoke Detection	Created	16/8/2024 12:09:59 PM	Corey Cartmill	High Risk		Edit
Hazardous Zone Intrusion Detection	Created	5/9/2024 12:00:56 PM	Nadeesha Chandrasena	High Risk		Edit
Heatmap Generation for Customer Behaviour	Created	16/8/2024 11:53:31 AM	Nadeesha Chandrasena	Very Low Risk		Edit
License Plate Recognition for Parking Management	Created	16/8/2024 11:43:31 AM	will.lutz@smartaiconnect.com	Low Risk		Edit
Loitering Detection in Restricted Areas	Created	16/8/2024 11:50:50 AM	Corey Cartmill	High Risk		Edit
Retail Theft Prevention	Created	16/8/2024 11:55:53 AM	T D	Medium Risk		Edit
Slip and Fall Detection in Public Areas	Created	16/8/2024 11:57:15 AM	Corey Cartmill	Medium Risk		Edit
Social Distancing Monitoring	Created	16/8/2024 12:08:44 PM	will.lutz@smartaiconnect.com	Medium Risk		Edit
Traffic Flow Optimisation	Created	16/8/2024 12:11:17 PM	Nadeesha Chandrasena	Low Risk		Edit
Visitor Count Analysis	Created	16/8/2024 11:45:09 AM	Corey Cartmill	Very Low Risk		Edit

Support for Mulitple Camera Manufactures and Software Providers



Chat-GPT sprints to 100 million users



Source: World of Statistics

National Safety Institutes

NIST AI SAFETY INSTITUTE

https://www.nist.gov/aisi

AISI AI SAFETY INSTITUTE

https://www.aisi.gov.uk

Key challenges:

- A lack of commonly accepted definitions for AI safety, as well as safety capabilities and measurements of those capabilities, especially for frontier models and advanced AI agents and systems
- Underdeveloped testing, evaluation, validation and verification (TEVV) methods and best practices to provide holistic assessments of risk – from model capabilities to human-AI interaction to system system-level and societal-level impacts
- An absence of scientifically-established risk mitigations across the lifecycle of AI design and deployment
- An insufficient understanding of the relationship between model architecture and design and model behaviour and performance, especially after deployment
- Limited and ad hoc coordination around safety practices among industry, civil society, and national and international actors

Mission:

- Test advanced AI systems and inform policymakers about their risks;
- Foster collaboration across companies, governments, and the wider research community to mitigate risks and advance publicly beneficial research; and
- Strengthen AI development practices and policy globally.

KJR est 1997

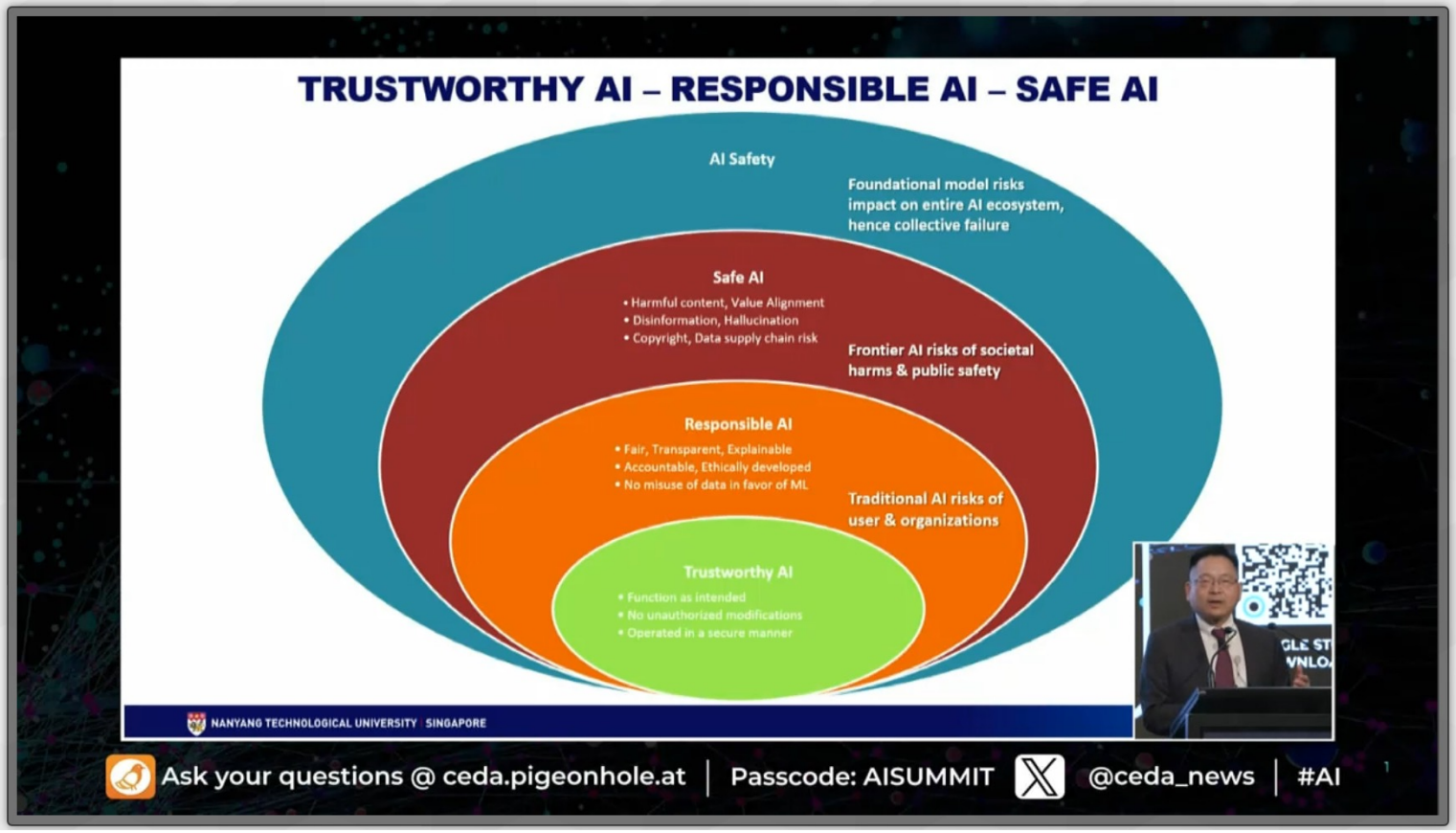
NIST - Potential AI Harms

https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

KJR est 1997

AI Leadership Summit

Pigeonhole – Q&A portal



Australian AI Ethics Principles

Beneficial

- AI systems should benefit individuals, society and the environment.

Human-centred

- AI systems should respect human rights, diversity, and the autonomy of individuals.

Fairness

- AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

Privacy and Security

- AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

Reliability and Safety

- AI systems should reliably operate in accordance with their intended purpose.

Explainability

- There should be transparency and responsible disclosure so people can understand when they are being significantly impacted by AI, and can find out when an AI system is engaging with them.

Contestibility

- When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or outcomes of the AI system.

Accountability

- Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles

KJR est 1997

https://architecture.digital.gov.au/responsible-use-of-AI-in-government

Effective 1 Sep 2024

Policy Requirements:

1. Accountable officials

30 Nov 2024

Agencies **must** designate accountability for implementing this policy to accountable official(s).

The responsibilities of the accountable officials are to:

- be accountable for implementation of this policy within their agencies
- notify the Digital Transformation Agency (DTA) where the agency has identified a new high-risk use case.
  This information will be used by the DTA to build visibility and inform the development of further risk mitigation approaches.
- be a contact point for whole-of-government AI coordination
- engage in whole-of-government AI forums and processes
- keep up to date with changing requirements as they evolve over time.

2. AI transparency statement

28 Feb 2025

Agencies **must** make publicly available a statement outlining their approach to AI adoption and use

The statement **must** be reviewed and updated annually or sooner, should the agency make significant changes to their approach to AI.

This statement must provide the public with relevant information about the agency's use of AI including information on:

- compliance with this policy
- measures to monitor effectiveness of deployed AI systems
- efforts to protect the public against negative impacts.

KJR est 1997

https://architecture.digital.gov.au/responsible-use-of-AI-in-government

AI Transparency Statement

28 Feb 2025

MUST provide the following information regarding their use of AI:

- the intentions behind why the agency uses AI or is considering its adoption
- classification of AI use according to usage patterns and domains
- classification of use where the public may directly interact with, or be significantly impacted by, AI without a human intermediary or intervention
- measures to monitor the effectiveness of deployed AI systems, such as governance or processes
- compliance with applicable legislation and regulation
- efforts to identify and protect the public against negative impacts
- compliance with each requirement under the Policy for responsible use of AI in government
- when the statement was most recently updated.

AI Definition

"An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment." [OECD definition]

Used in [Policy for the responsible use of AI in government]

"a technical and scientific field devoted to the engineered system that generates outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives" [ISO/IEC 22989:2022]

AI Principles

Image Source: https://www.ceda.com.au/researchandpolicies/research/data-digital-economy/artificial-intelligence-principles-to-practice

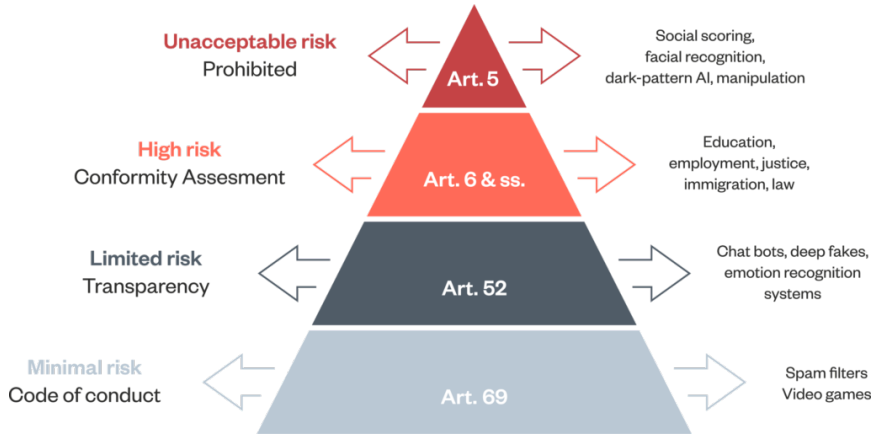
KJR est 1997

Voluntary AI Safety Standard Guardrails

1. Establish, implement and publish an accountability process including governance, internal capability and a strategy for regulatory compliance.
2. Establish and implement a risk management process to identify and mitigate risks.
3. Protect AI systems, and implement data governance measures to manage data quality and provenance.
4. Test AI models and systems to evaluate model performance and monitor the system once deployed.
5. Enable human control or intervention in an AI system to achieve meaningful human oversight across the life cycle.
6. Inform end-users regarding AI-enabled decisions, interactions with AI and AI-generated content.
7. Establish processes for people impacted by AI systems to challenge use or outcomes.
8. Be transparent with other organisations across the AI supply chain about data, models and systems to help them effectively address risks.
9. Keep and maintain records to allow third parties to assess compliance with guardrails.
10. Engage your stakeholders and evaluate their needs and circumstances, with a focus on safety, diversity, inclusion and fairness.

https://www.industry.gov.au/publications/voluntary-ai-safety-standard/10-guardrails

KJR est. 1997



Effective 1 Aug 2024

Enforced 2 Aug 2026

Prohibited AI (Late 2024)

- Systems for social credit scoring
- Employment and educational emotion recognition systems
- AI leveraging people's vulnerabilities (e.g., age, disability)
- Manipulation of behavior and undermining free will
- Indiscriminate collection of facial images for facial recognition
- Biometric categorization systems involving sensitive traits
- Particular predictive policing uses
- Law enforcement employing real-time biometric
- identification in public (except in restricted, pre-approved scenarios)

High-Risk AI (Mid 2026)

- Medical devices & Automobiles
- Hiring, human resources, and labor supervision
- Instruction and professional education
- Shaping political elections and the voting populace
- Entry to amenities (like insurance, banking, credit, benefits, etc.)
- Supervising crucial infrastructure (like water, gas, electricity, etc.)
- Systems for recognizing emotions and identifying individuals via biometrics
- Policing, regulating borders, migration, and asylum procedures and conducting legal affairs
- Particular merchandise or safety elements within particular products

No longer high-risk

If not intended to be used as a safety component of a product covered by the Union harmonisation legislation.*

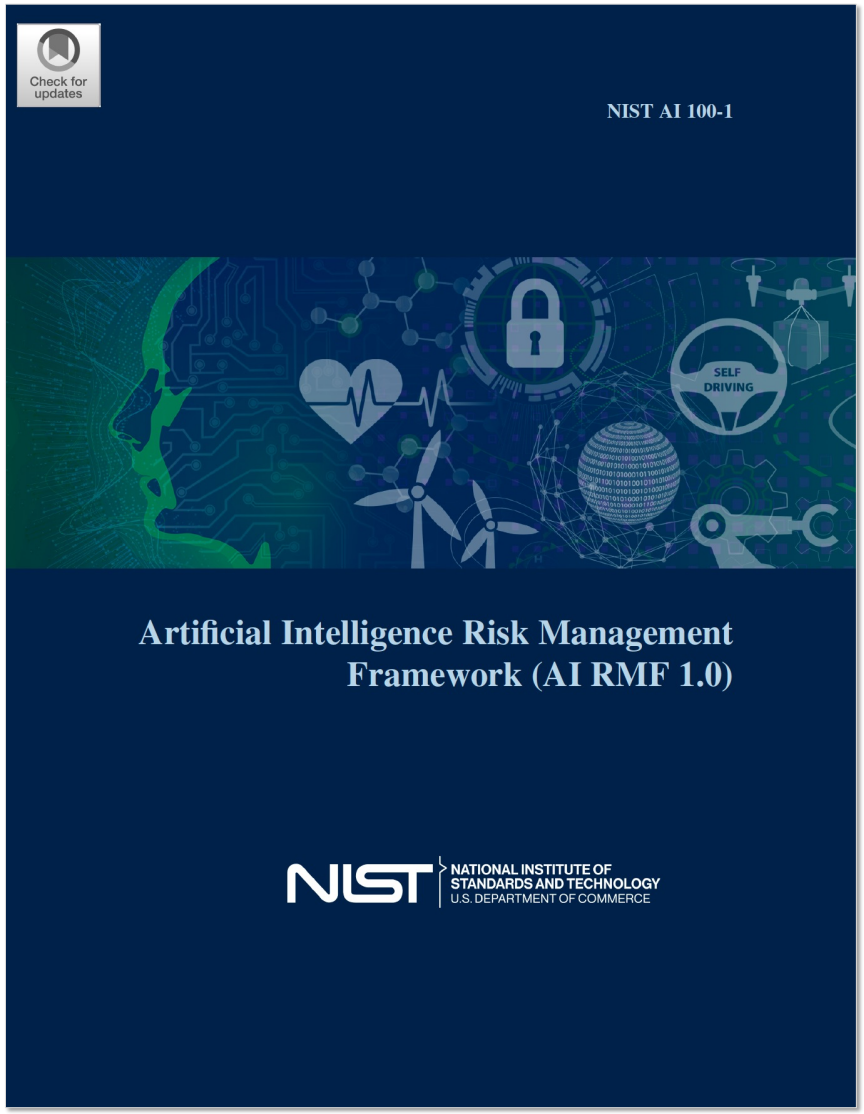
- Performs a narrow procedural task
- Intended to improve the result of a previously completed human activity
- Not meant to replace or influence the previously completed human assessment, without proper human review
- perform a preparatory task to an assessment relevant to high-risk use-cases.

KJR est 1997

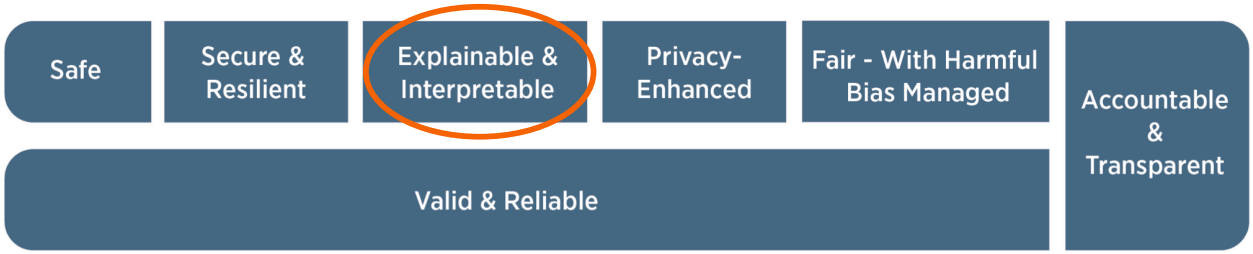
High risk AI providers must:

- Establish a **risk management system** throughout the high risk AI system's lifecycle;
- Conduct **data governance**, ensuring that training, validation and testing datasets are relevant, sufficiently representative and, to the best extent possible, free of errors and complete according to the intended purpose.
- Draw up **technical documentation** to demonstrate compliance and provide authorities with the information to assess that compliance.
- Design their high risk AI system for **record-keeping** to enable it to automatically record events relevant for identifying national level risks and substantial modifications throughout the system's lifecycle.
- Provide **instructions for use** to downstream deployers to enable the latter's compliance.
- Design their high risk AI system to allow deployers to implement **human oversight**.
- Design their high risk AI system to achieve appropriate levels of **accuracy, robustness, and cybersecurity**.
- Establish a **quality management system** to ensure compliance.

KJR est 1997



NIST AI Trustworthiness Characteristics

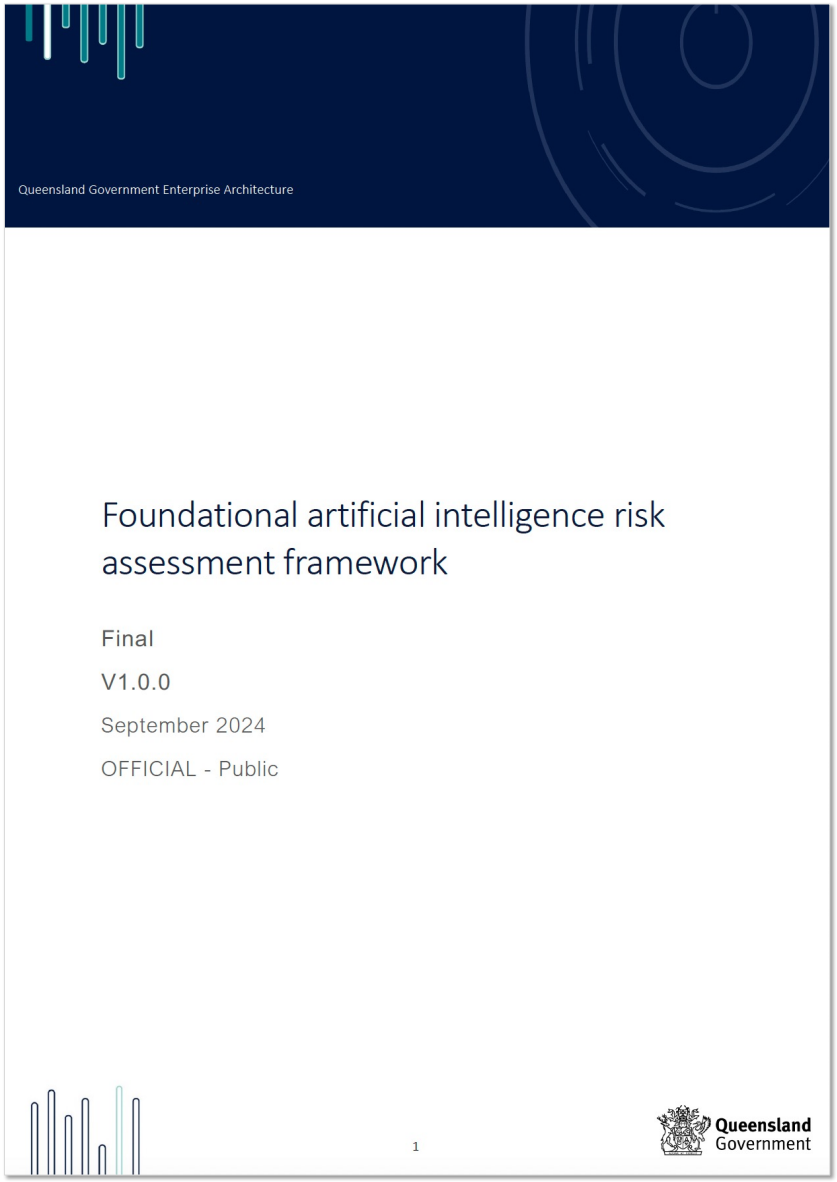


https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

KJR est. 1997

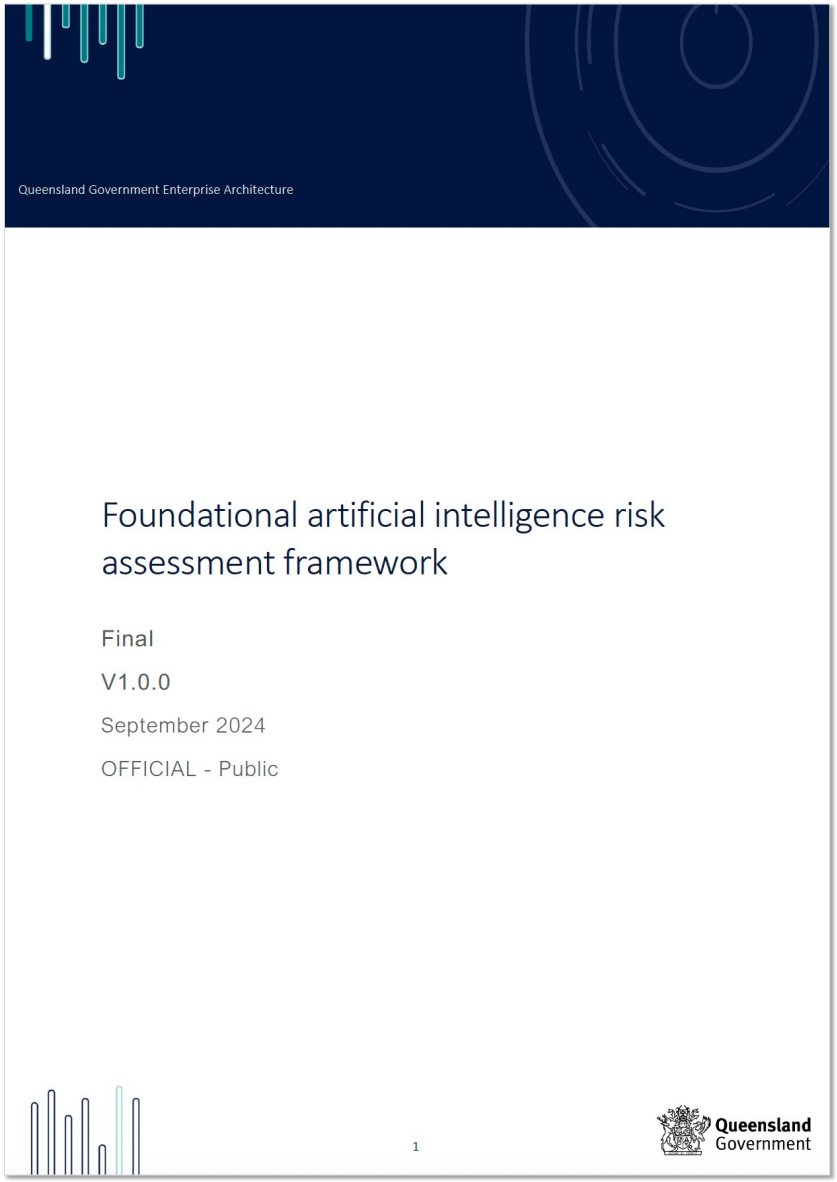
QLD Gov AI governance policy

- Agencies must use a consistent and evidence-based process which incorporates an ethical framework to evaluate their transparency, accountability, and risk associated with the AI lifecycle
  - FAIRA: Foundational artificial intelligence risk assessment guideline provides a comprehensive process for assessing AI risk within a specific context including alignment of tasks and responsibilities to relevant organisational roles
- Agencies must establish AI governance arrangements based on ISO 38507 (Governance implications of the use of artificial intelligence by organisations)
  - Aligns with existing IT governance processes
  - Provides specific advice regarding AI, including
    - usage governance
    - data governance

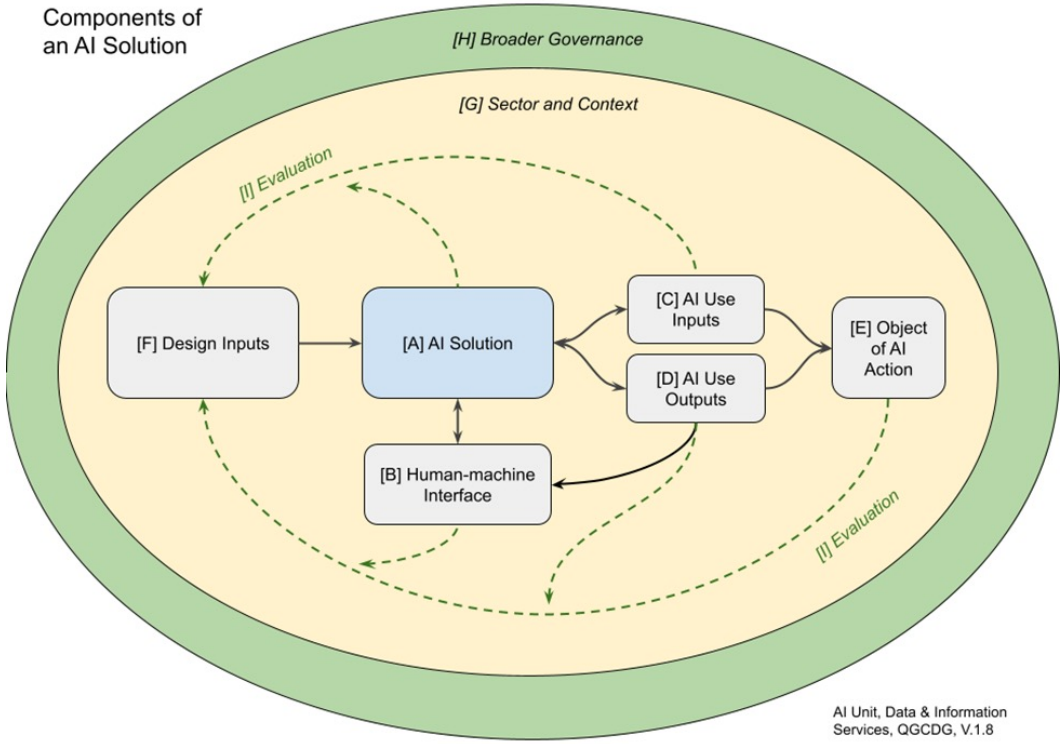


FAIRA Framework

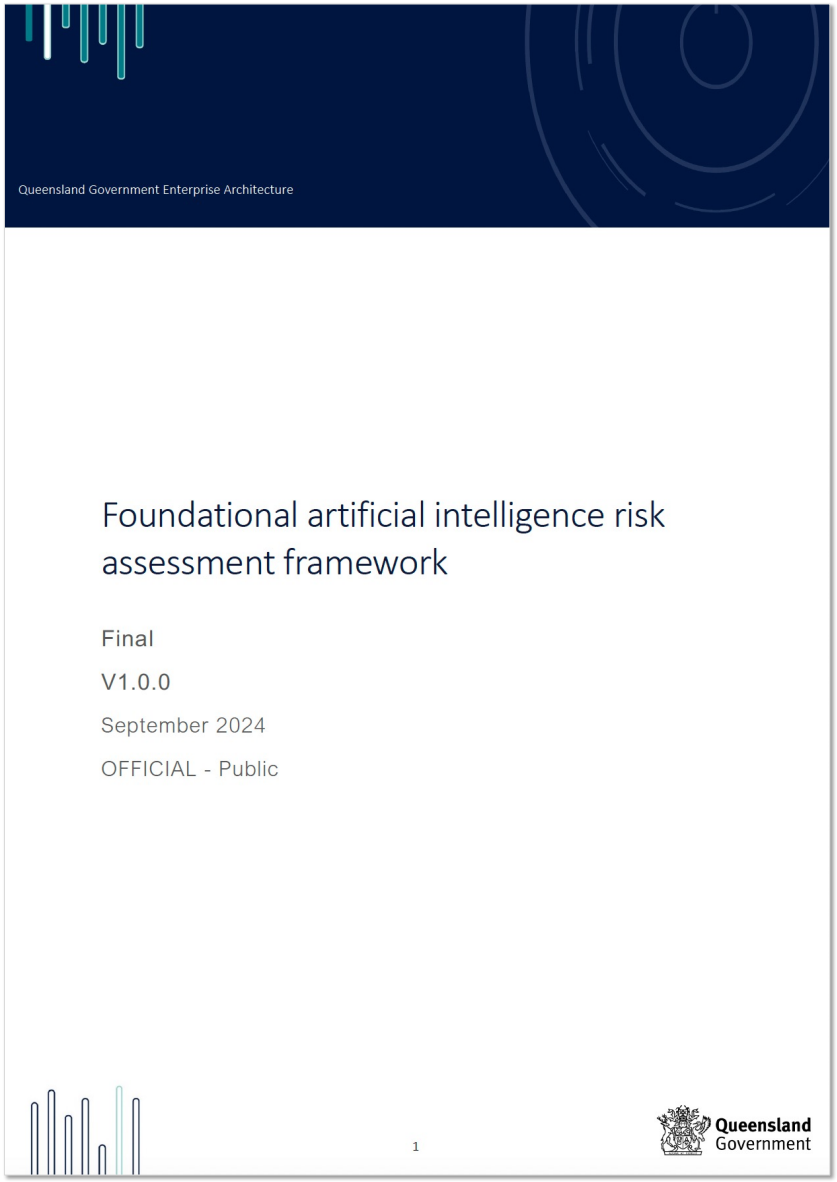
KJR est 1997



Part A – Components Analysis



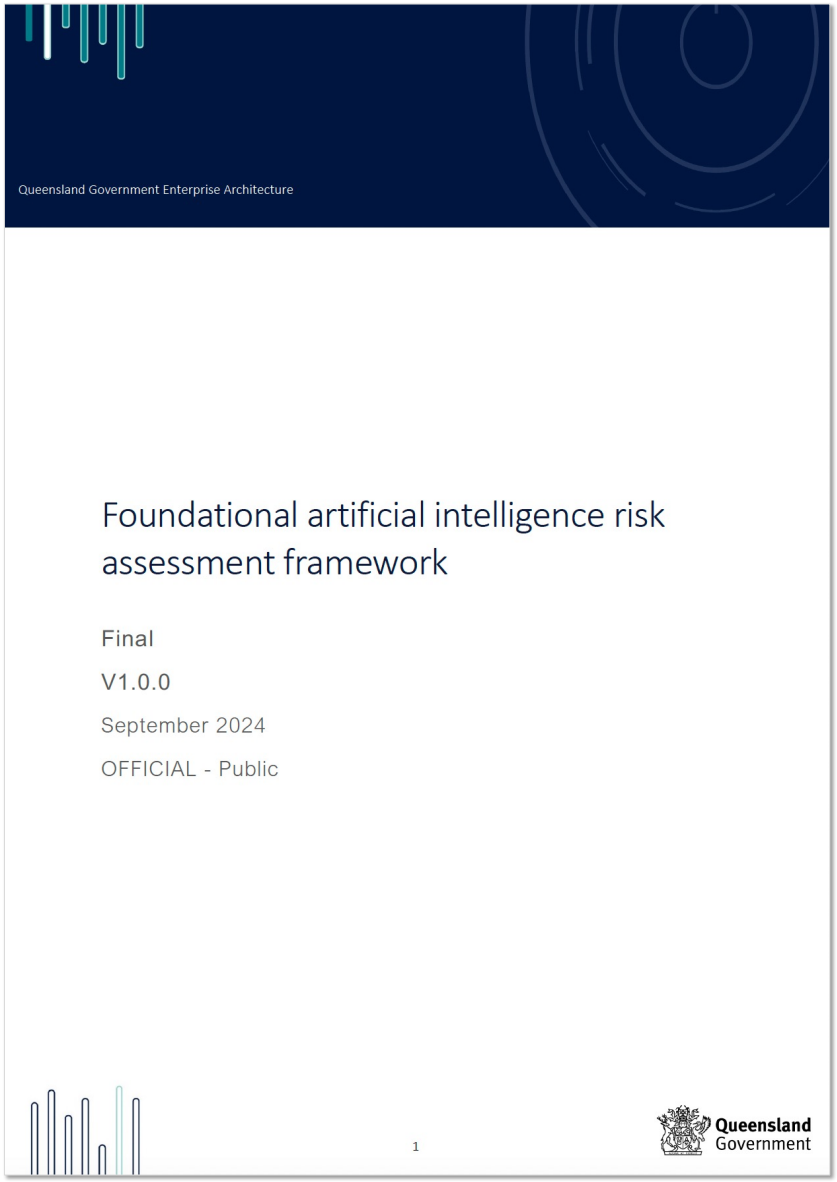
KJR est 1997



Part B – Components Analysis

- Human, societal, and environmental wellbeing
- Human-Centred Values
- Fairness
- Privacy Protection and Security
- Reliability and Safety
- Transparency and Explainability
- Contestability
- Accountability

KJR est 1997

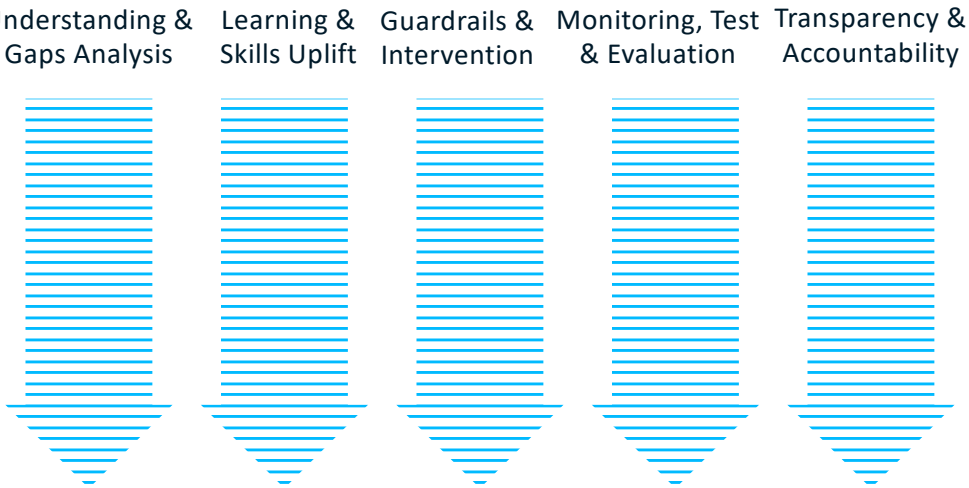


Part C – Controls for AI Risks

Control Areas

Business Function

Executive
Management
Technical
Policy
Engagement

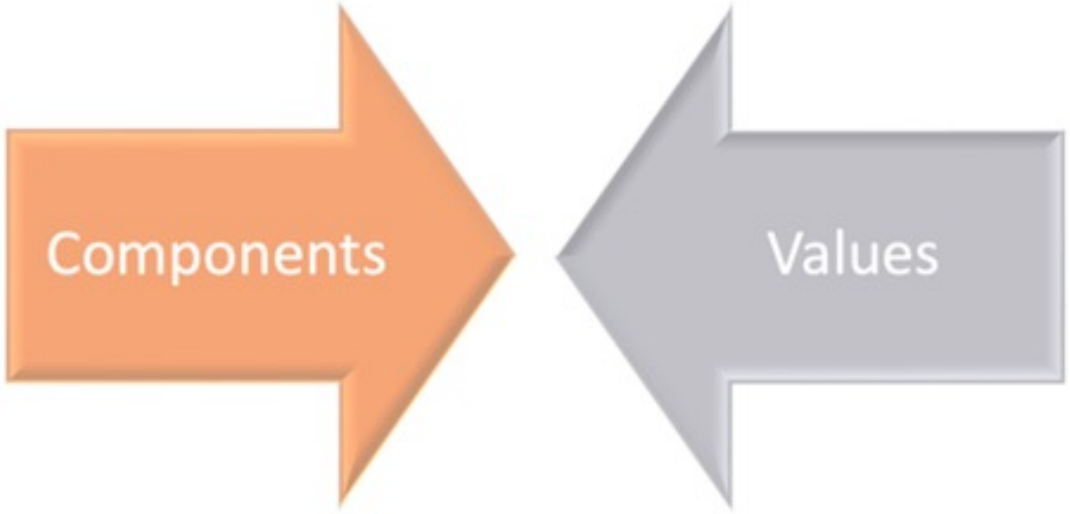


KJR est 1997

How FAIRA works

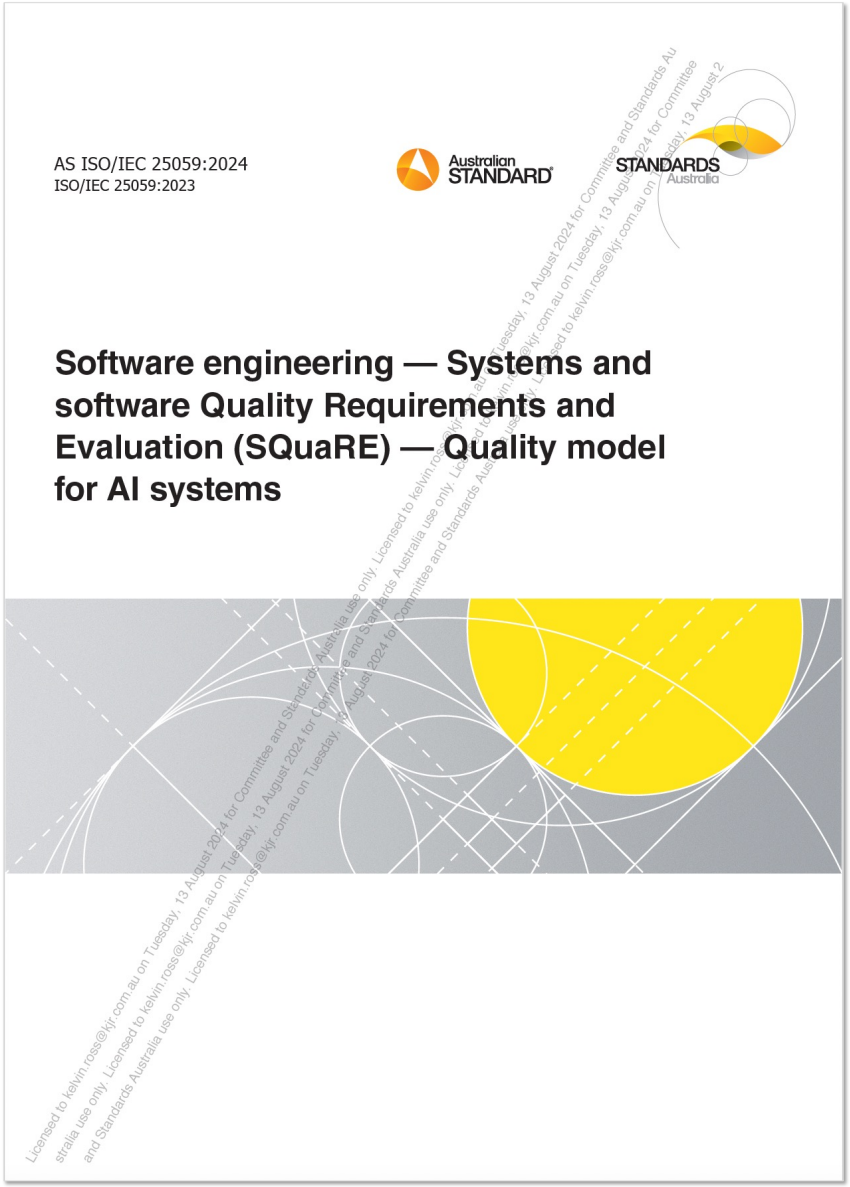
Explain components of AI solution

[A] AI Solution
[B] HMI
[C] AI use inputs
[D] AI use outputs
[E] Object of AI Action
[F] Design inputs
[G] Sector and Context
[H] Broader governance
[I] Evaluation



- Human, societal, and environmental wellbeing
- Human-Centred Values
- Fairness
- Privacy Protection and Security
- Reliability and Safety
- Transparency and Explainability
- Contestability
- Accountability

Explain effect of AI solution on values alignment



AI Quality Characteristics

KJR est 1997

ISO/IEC 25010:2023 - System and software quality models

SOFTWARE PRODUCT QUALITY								
FUNCTIONAL SUITABILITY	**PERFORMANCE EFFICIENCY**	**COMPATIBILITY**	**INTERACTION CAPABILITY**	**RELIABILITY**	**SECURITY**	**MAINTAINABILITY**	**FLEXIBILITY**	**SAFETY**
FUNCTIONAL COMPLETENESS	TIME BEHAVIOUR	CO-EXISTENCE	APPROPRIATENESS RECOGNIZABILITY	FAULTLESSNESS	CONFIDENTIALITY	MODULARITY	ADAPTABILITY	OPERATIONAL CONSTRAINT
FUNCTIONAL CORRECTNESS	RESOURCE UTILIZATION	INTEROPERABILITY	LEARNABILITY	AVAILABILITY	INTEGRITY	REUSABILITY	SCALABILITY	RISK IDENTIFICATION
FUNCTIONAL APPROPRIATENESS	CAPACITY		OPERABILITY	FAULT TOLERANCE	NON-REPUDIATION	ANALYSABILITY	INSTALLABILITY	FAIL SAFE
FUNCTIONAL ADAPTABILITY			USER ERROR PROTECTION	RECOVERABILITY	ACCOUNTABILITY	MODIFIABILITY	REPLACEABILITY	HAZARD WARNING
iso25000.com			USER ENGAGEMENT	ROBUSTNESS	AUTHENTICITY	TESTABILITY		SAFE INTEGRATION
			INCLUSIVITY		RESISTANCE			
			USER ASSISTANCE		INTERVENABILITY			
			SELF-DESCRIPTIVENESS					
			USER CONTROLLABILITY					
			TRANSPARENCY					

ISO/IEC 25059:2024 - Quality model for AI systems

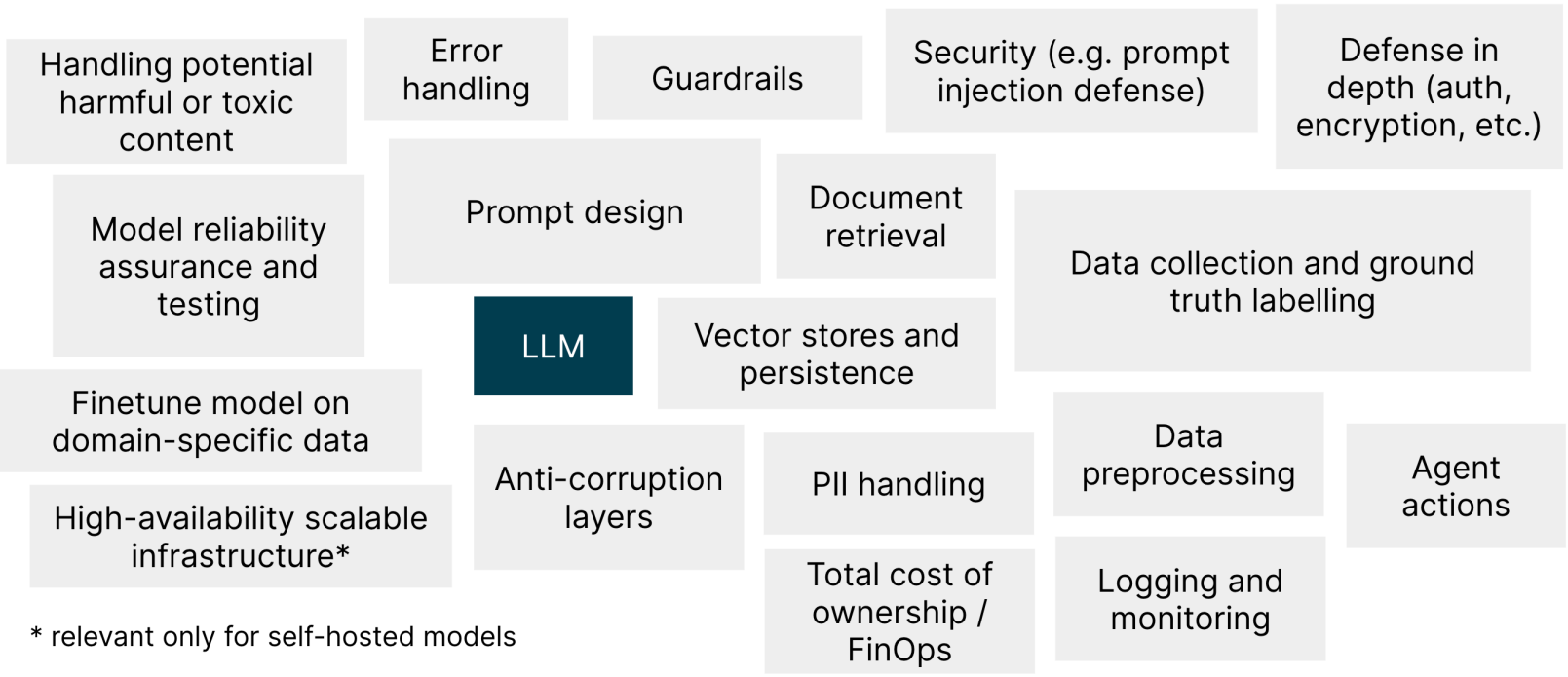
KJR est 1997

ISO/IEC 25010:2023 - System and software quality models

ISO/IEC 25059:2024 - Quality model for AI systems

Architecting LLM applications

The language model is just one part of the technical architecture



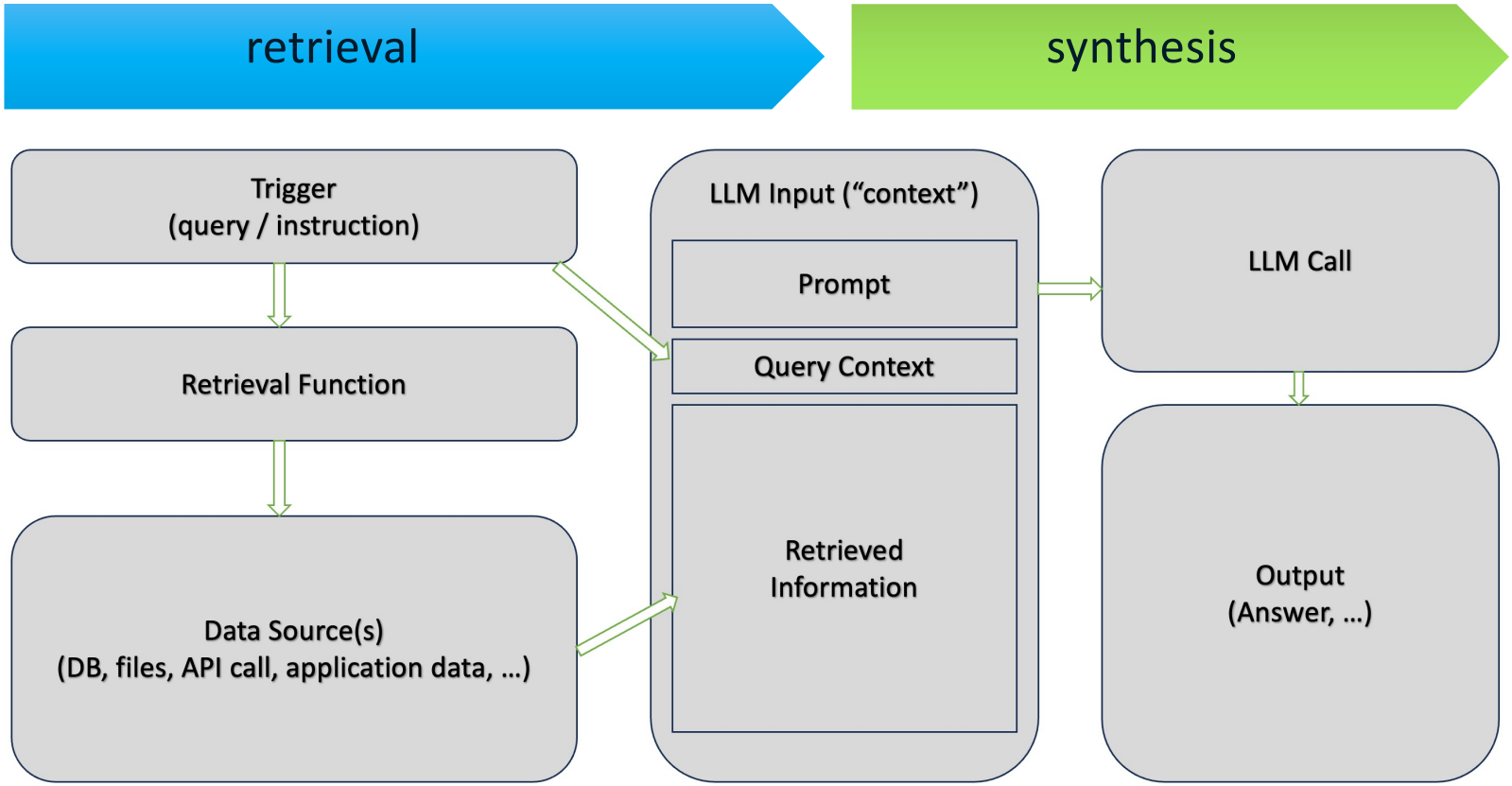
* relevant only for self-hosted models

Adapted from: [Machine Learning: The High Interest Credit Card of Technical Debt (Google)]

https://martinfowler.com/articles/engineering-practices-llm.html

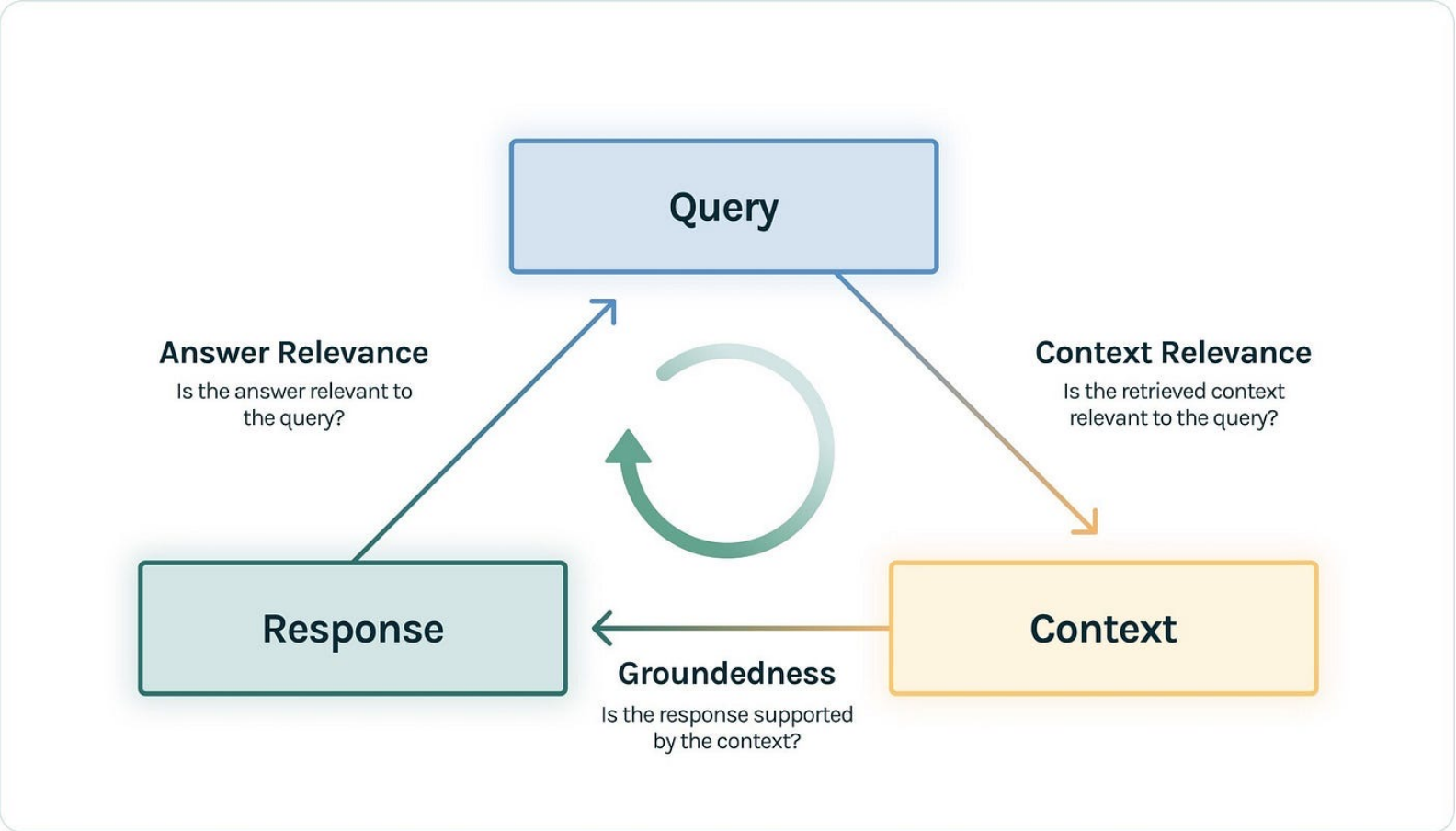
KJR est 1997

RAG Workflow



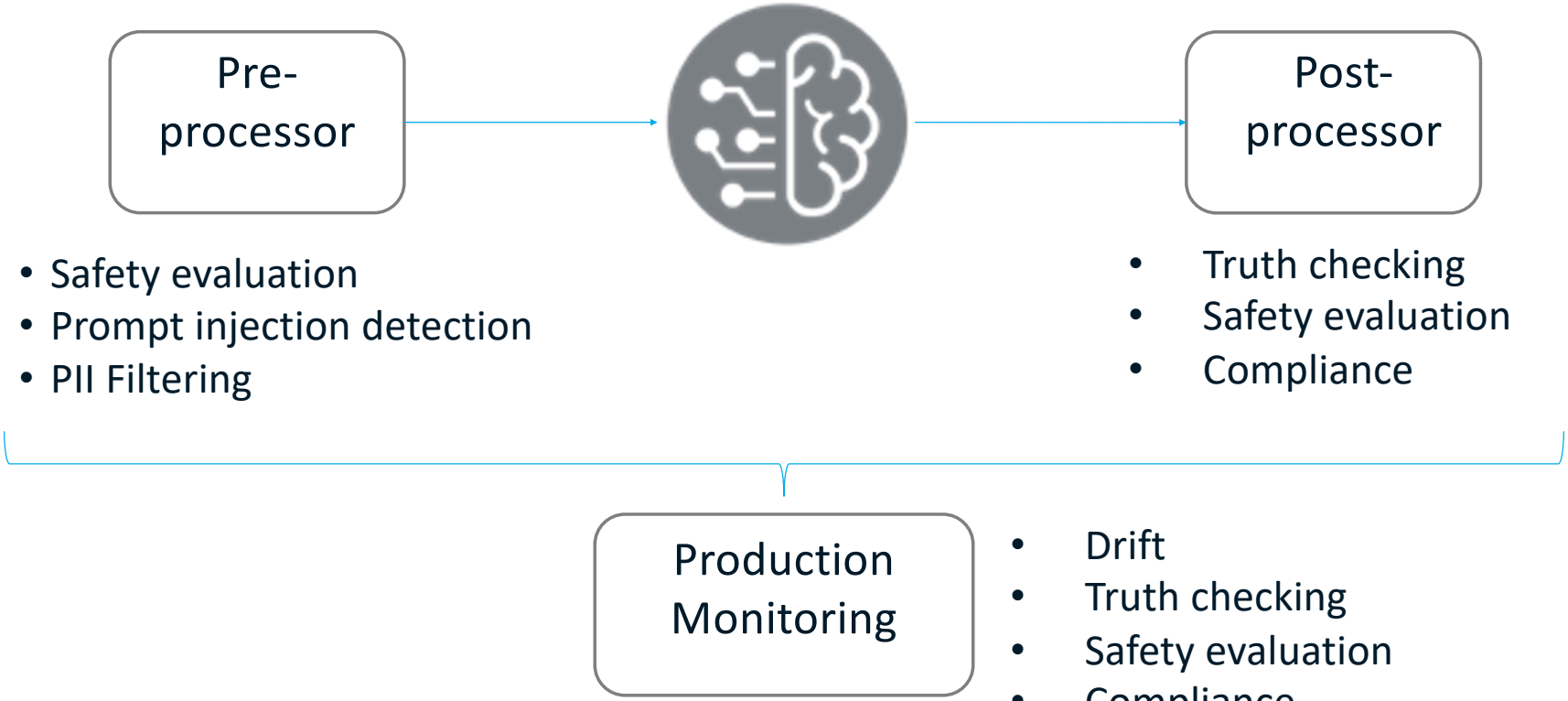
Source: https://techcommunity.microsoft.com/t5/fasttrack-for-azure/grounding-llms/ba-p/3843857

The RAG Triad



Source : https://www.trulens.org/trulens_eval/core_concepts_rag_triad/

Guardrails


Pre-processor
Post-processor
Production Monitoring
• Safety evaluation
• Prompt injection detection
• PII Filtering
• Truth checking
• Safety evaluation
• Compliance
• Drift
• Truth checking
• Safety evaluation
• Compliance

ISO TS 29119-11 : Testing of AI-based Systems

ISO/IEC JTC 1/SC 42/JWG 2 **N 244**

ISO/IEC JTC 1/SC 42/JWG 2 "Joint Working Group ISO/IEC JTC1/SC 42-ISO/IEC JTC1/SC 7 : Testing of AI-based systems"
Convenorship: **BSI**
Convenors: **Reid Stuart Dr, Smith Adam Leon Mr**

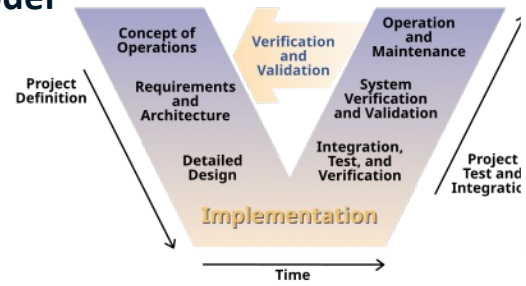
ISO TS 29119-11 V0.23

ISO/IEC JTC 1/SC 7
Software & Systems Engineering

ISO/IEC JTC 1/SC 42
Artificial Intelligence

Risk-Based Testing

V-Model



Specific Test Techniques
Evals / Test Oracles

Data Sampling

Data Quality Testing / Model Testing

Validation Data Leakage

Performance Metrics

Uncertainty

Coverage

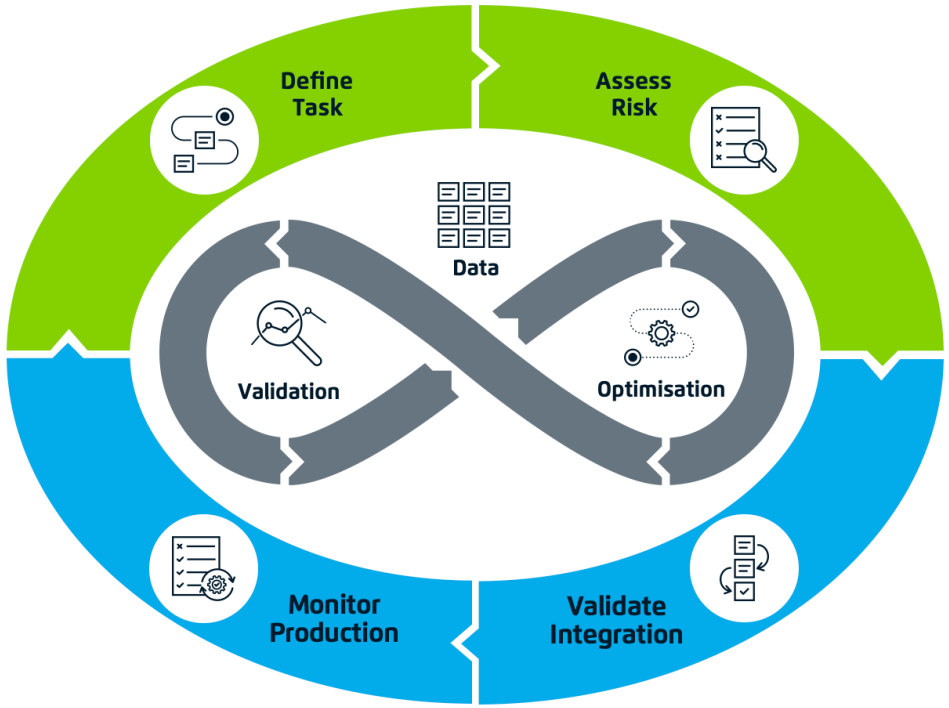
Data Science

Statistics

KJR est 1997

VDML™

Validation Driven Machine Learning (VDML) is a methodology developed by KJR to guide development of robust and reliable Machine Learning (ML) models. VDML emphasises understanding the risks inherent within the application context and the limitations that arise from the available data and model building processes applying iterative validation and optimisation methods to deliver an acceptable solution which can be integrated and governed within a real-world context.



DEFINE CONTEXT

Define the goals of applying machine learning to a specific problem area, being sure to include the data being used, the context of use (historical analytics vs live decision support) and expected benefits from a range of different stakeholders. Given this context, assess risks, including the impacts of potential failure, the required governance processes.

RESOLVE LIMITATIONS

Direct use of pre-built models or naïve approaches to machine learning can lead to unreliable performance. Key to validating and optimising model performance is the selection of training and testing data sets which are close to real world usage, and detailed error analysis which can uncovers underlying faults and limitations.

GOVERN BEHAVIOUR

Track the integrity of the model through build, deploy and operation, monitoring for residual risks, model drift / sabotage, identifying opportunities for further optimisation and risk reduction.

KJR est 1997

PODCAST SERIES

TRUST YOUR AI WITH VDML®

PRACTICAL CONVERSATIONS ABOUT AI ASSURANCE

LISTEN NOW

KJR est. 1997

Other ISO Standards


INTERNATIONAL STANDARD

ISO/IEC 42001

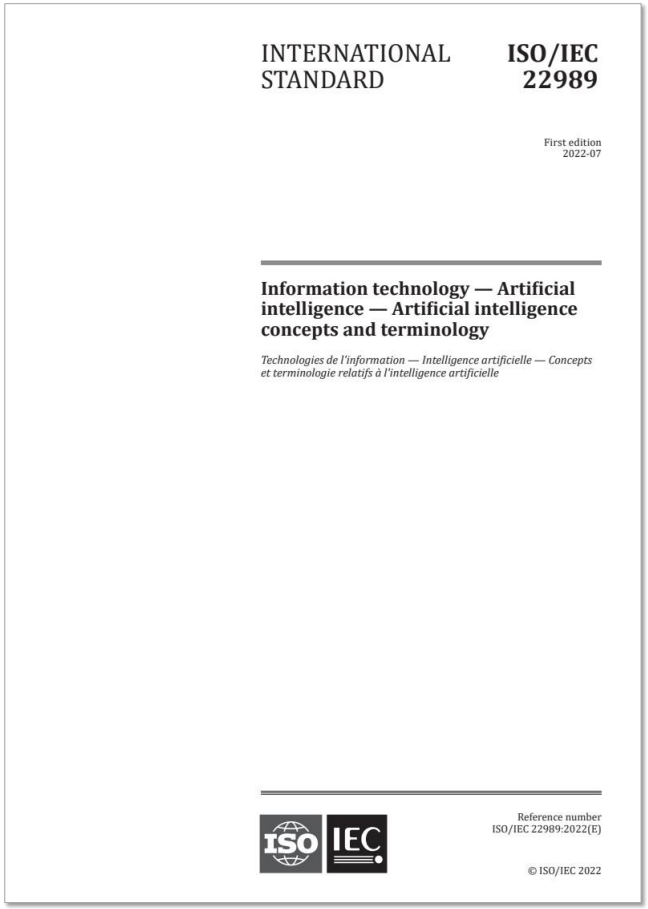
First edition
2023-12

Information technology — Artificial intelligence — Management system

Technologies de l'information — Intelligence artificielle — Système de management

Reference number
ISO/IEC 42001:2023(E)

© ISO/IEC 2023


INTERNATIONAL STANDARD

ISO/IEC 22989

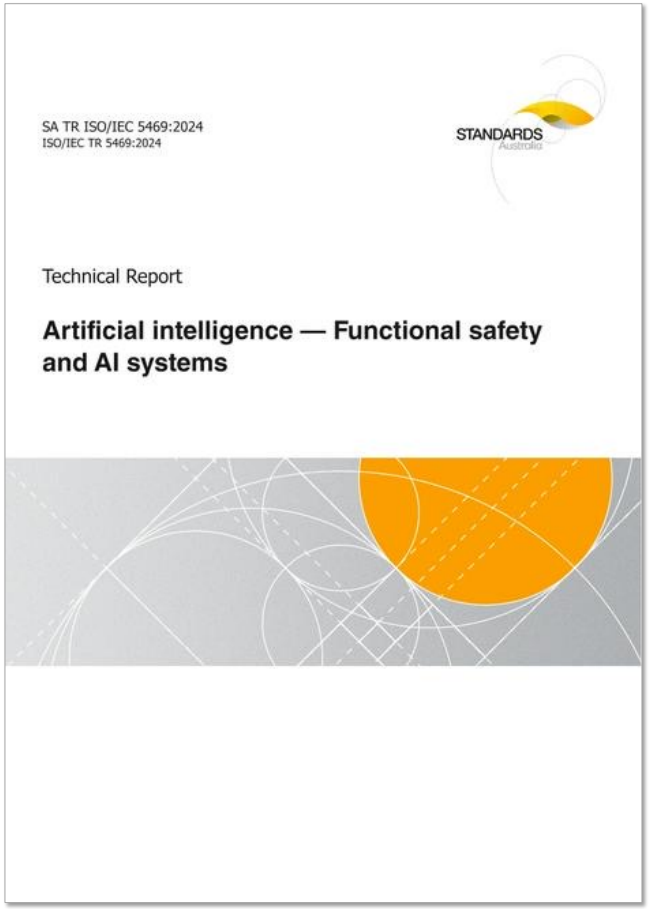
First edition
2022-07

Information technology — Artificial intelligence — Artificial intelligence concepts and terminology

Technologies de l'information — Intelligence artificielle — Concepts et terminologie relatifs à l'intelligence artificielle

Reference number
ISO/IEC 22989:2022(E)

© ISO/IEC 2022

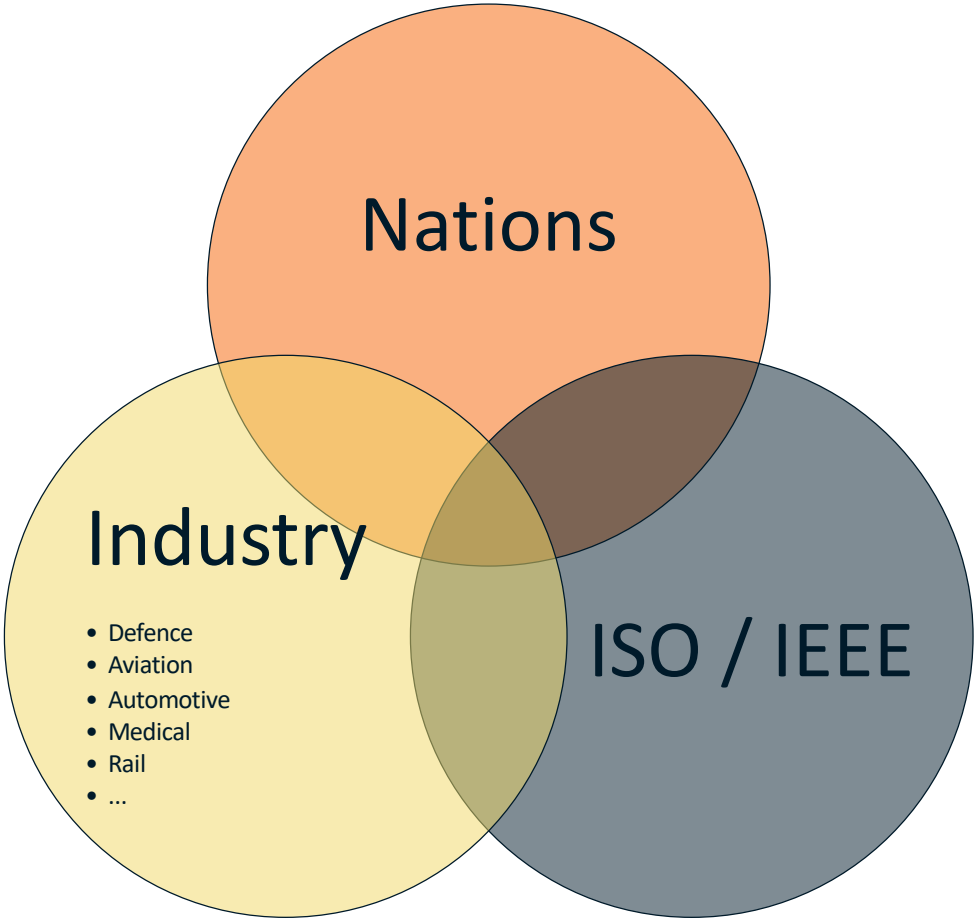

SA TR ISO/IEC 5469:2024
ISO/IEC TR 5469:2024

STANDARDS Australia

Technical Report

Artificial intelligence — Functional safety and AI systems

Policy / Standards Development



Key Takeaways

- Emerging Obligations
  - Do we have the approach, skills and capacity to provide the requested assurance?

- Overall governance is not that unfamiliar for systems assurance folk
  - What role will the Systems Safety Assurance people play in this expanded AI Safety space?

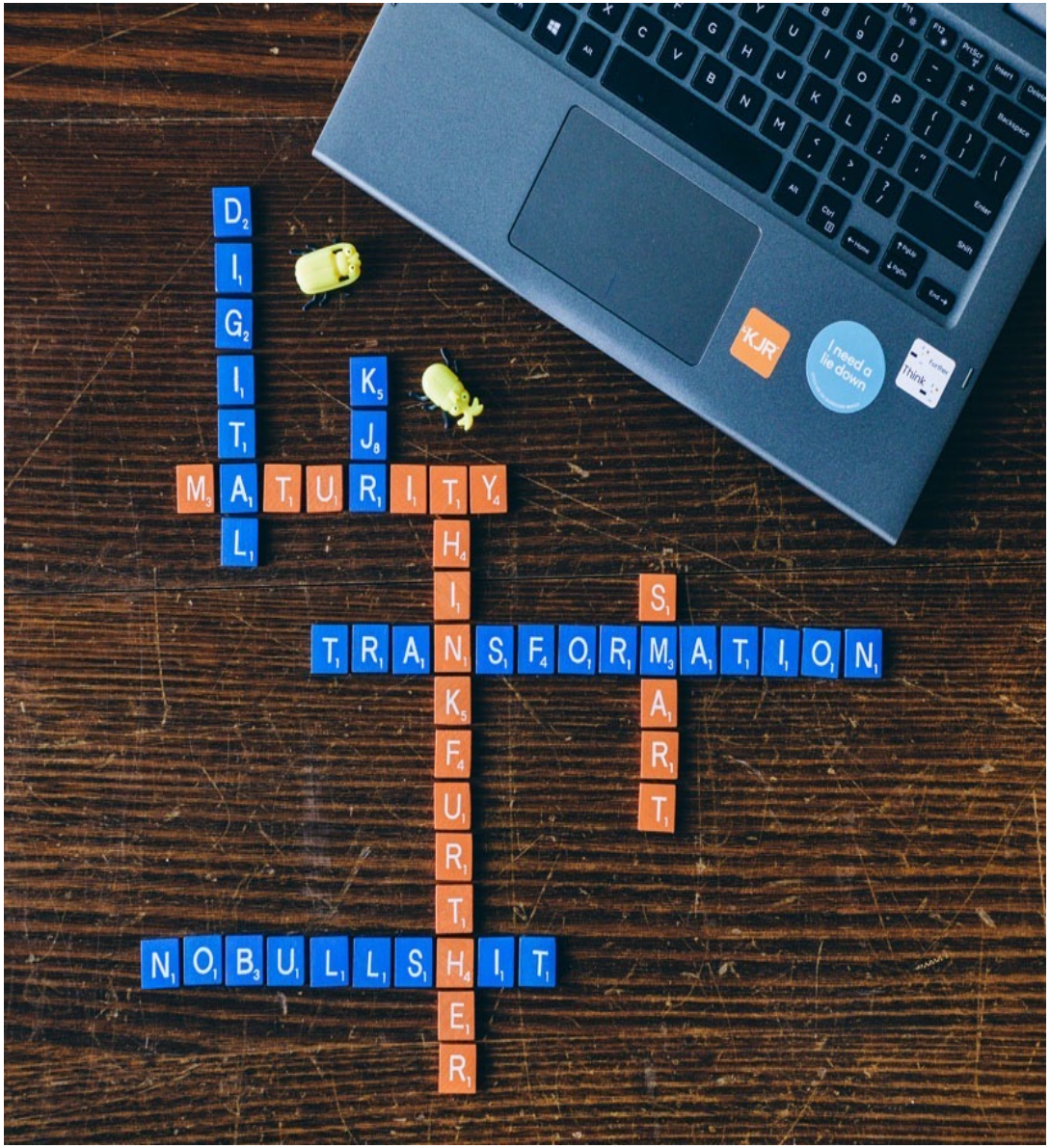
- What is different about AI?
  - Definition of scope – what is AI?
  - Understanding of risks
  - Mitigations and controls

KJR est 1997

Trust your AI

Kelvin Ross
kelvin.ross@kjr.com.au

Mark Pedersen
mark.pedersen@kjr.com.au



kjr.com.au