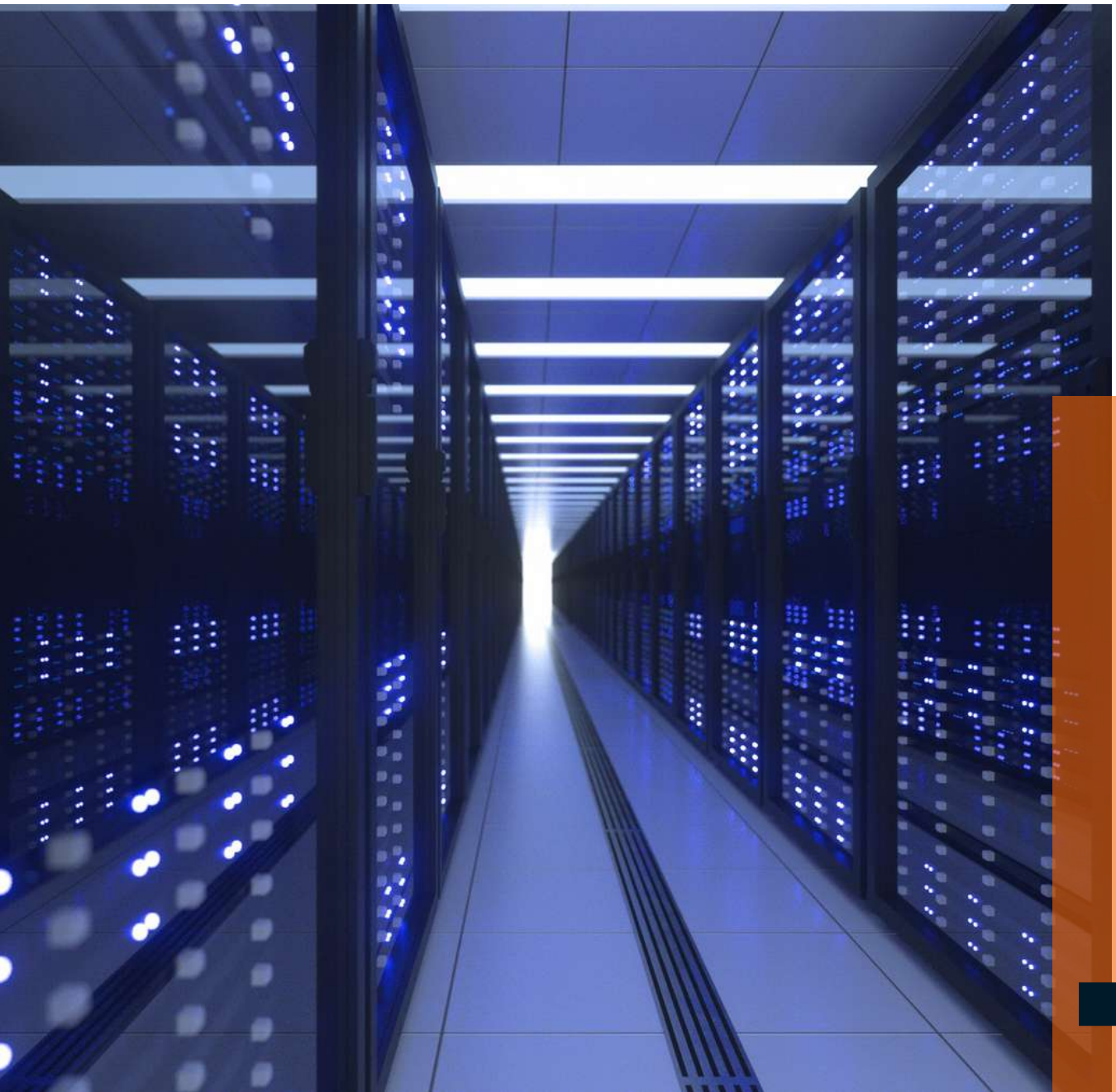
PRACTICAL STEPS TOWARD RESPONSIBLE GOVERNANCE OF AI-ENABLED SYSTEMS

WHITE PAPER

DR KELVIN ROSS
DR MARK PEDERSEN

kjr.com.au



January 2024

Contents

AI SAFETY REGULATIONS

As AI applications grow in capability and accessibility, various governance standards and frameworks are emerging across different jurisdictions. A number of technical and legislative frameworks have emerged globally, with a new international standard on the management of Artificial Intelligence systems (ISO/IEC 42001) published in December 2023. In addition to these formal industry standards and governance mechanisms, there has been a global shift toward establishing legislation to regulate the way in which AI-enabled systems are developed and used. This includes the Bletchley Declaration arising from the 2023 AI Safety Summit in the UK, the US Executive Order on AI Safety and the EU AI Act, all of which emphasise the need for strong governance frameworks to ensure that AI's benefits are maximized whilst minimizing risks.

Technical & Legislative Frameworks

ISO/IEC 42001
Bletchley Declaration
US Executive Order on AI Safety
EU AI Act

AI safety regulations apply just as much to organisations deploying AI as those developing core Machine Learning models and platforms. In contrast to traditional software applications which essentially encode a set of fixed business rules, the power of AI-enabled systems arises from their ability to be trained on and adapt and respond to specific local data sets.

This also means that AI-enabled systems can fail in ways that are significantly different from traditional software systems:

- When target data differs from the data that an ML model was trained on, its performance can degrade significantly. For example, a model designed to detect crop diseases may have been trained on images of North American crops, and perform very well in that context but fail to recognise similar diseases when given images of Australian crops. There is typically a need to test and fine-tune any ML model to ensure it will function at an acceptable level in a local context.
- ML models may perform poorly in edge cases, for which there may be a sparse amount of data. For example, computer vision models which support self-driving and lane-assist features in cars may have been trained on a large amount of "normal traffic" imagery, and perform well in those contexts, but struggle to process situations such as road-works, for which there may be comparatively much less training data as well as much greater variability.
- ML models can draw unexpected and incorrect correlations, based on the training data provided. A famous example is a model which appeared to perform excellently at distinguishing between wolves and domestic dogs with similar features, but all it was really doing was detecting the presence of snow in the background of the image.

Technical & Procedural Compliance Requirements

Organisations deploying AI-enabled systems have a heightened duty of care with regard to how these systems behave in the specific context in which they are deployed: it is not reasonable to expect these systems to work "off-the-shelf" without testing them. In terms of governance frameworks, the typical technical controls required include:

Accuracy	ensuring that the AI behaves predictably and correctly in the target context.
Robustness	ensuring that system is resilient against unexpected input, including adversarial attacks.
Privacy and Security	ensuring the confidentiality and security of data against unauthorized access and breaches.
Data Quality	ensuring the data used to train and feed into the AI is accurate and representative.
AI system bias and decision-making	evaluating and correcting biases inherent in data and models.

In addition to checking that an AI-enabled system meets its intended performance requirements at a technical level, organisations deploying AI need to put in place a range of procedural compliance practices, including:

Transparency

ensuring users and stakeholders understand how the AI functions and makes decisions.

Accountability

determining responsibility for AI decisions, especially when they are wrong or cause harm.

Fairness

ensuring AI does not perpetuate or amplify existing biases, leading to discrimination.

Risk Evaluation

understanding potential harms and operationalizing risk management.

Human Oversight

ensuring there is human review and the ability to override AI decisions.

Community Benefits

assessing the wider societal advantages and ensuring AI provides broad benefits.

Organisations looking to deploy AI-enabled systems must develop the capability to manage these systems in line with the principles of responsible AI and, where appropriate, the specific requirements of both ISO 42001 and any specific governance frameworks and legislation that apply to their local jurisdiction. Most organisations with enterprise-scale software deployments will have a set of formal governance processes (e.g. ITIL) to control the quality of their software deployments. AI-enabled systems are no different in as much as they will also need to fit within a typical IT governance process. However, unlike traditional software systems which can typically be deployed by system integrator, configured in a limited number of specific ways, and then tested for user acceptance, and then follow a defined maintenance pathway, AI-enabled systems require much more extensive evaluation, fine-tuning and continual monitoring to ensure that the quality of their performance is maintained.

VALIDATION DRIVEN MACHINE LEARNING

Validation Driven Machine Learning (VDML) is a methodology developed by KJR to guide organisations in deploying robust and reliable AI solutions. The VDML process covers five key stages in AI assurance:

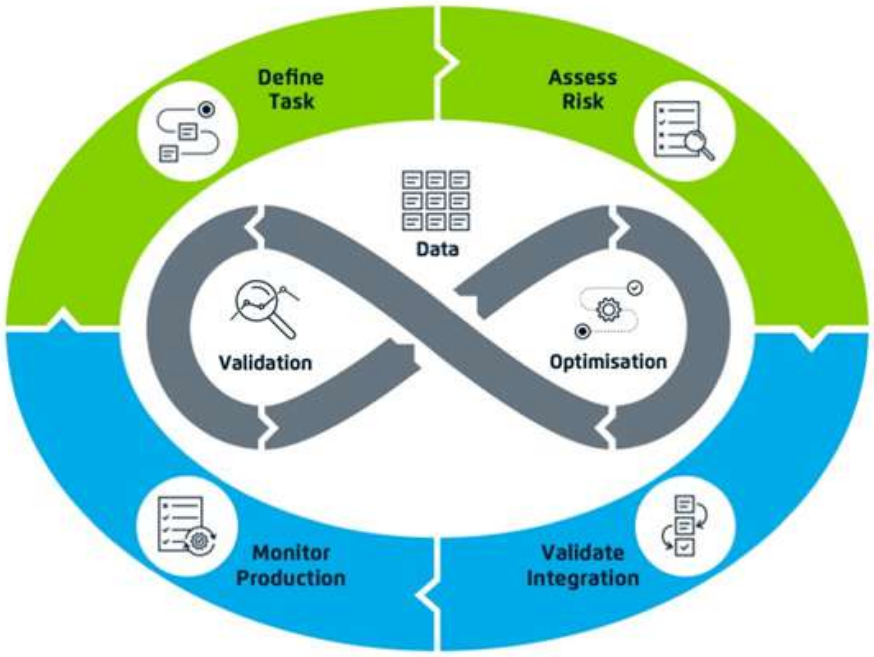
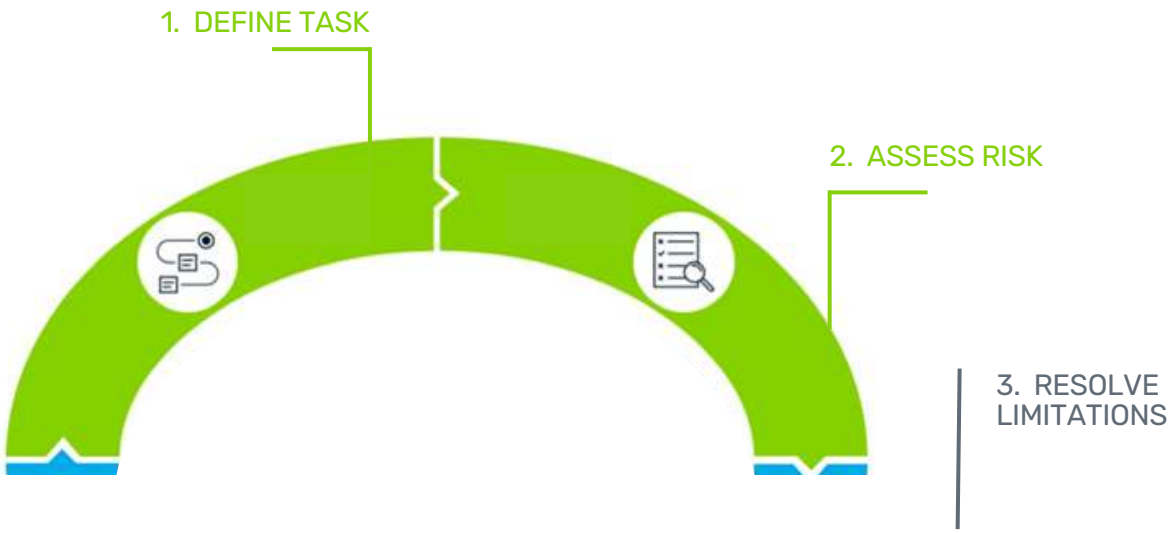


Figure 1 VDML Lifecycle

1. DEFINE TASK

By clearly defining the benefits the AI-enabled system is expected to deliver, stakeholders develop a clear understanding of the context in which the system is being used. This sets a baseline for assessing risk. For example, applying AI to assist with information retrieval and summarisation of large bodies of text is a significantly different context to applying AI to support clinical diagnosis of x-rays.



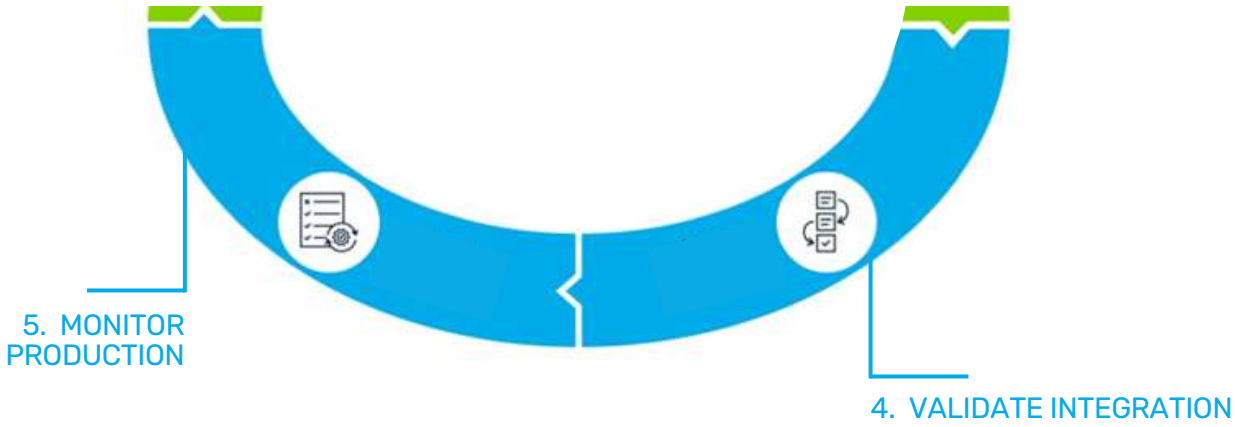
2. ASSESS RISK

Based on the task, and the relevant legislation and industry compliance requirements that may apply, a risk assessment helps establish the required governance practices that need to be put in place. The risk assessment process includes understanding the data being processed by the system and any data being used to develop or fine tune a machine learning model and the likely impact of any errors the system may make.

3. RESOLVE LIMITATIONS

Most organisations conduct User Acceptance Testing to validate that a generic software solution has been configured appropriately to support their specific business processes, as there is a need to validate machine learning models to ensure they will work within the intended context. Direct use of pre-built models or naive approaches to machine learning can lead to unreliable performance. As discussed, it is typical that any machine learning component needs to be tested against the target data and fine-tuned if performance does not meet expectations.

VDML provides a number of specific techniques by which KJR can help organisations test and tune machine learning components of their system to be confident that they will perform as expected. By selecting test data sets which are close to real world usage, and carrying out detailed error analysis, KJR can help organisations uncover underlying faults and limitations and put appropriate risk mitigations in place. In situations where tuning has already been performed by a solution provider, independent validation of the model's performance is key to a responsible approach to AI-enablement.



4. VALIDATE INTEGRATION

An AI-enabled system is much more than just a bare machine-learning model. ML components need to be integrated into a service delivery pipeline which enables appropriate interaction with the host organisation's data and users. This can include ensuring that ML components have access only to the appropriate data (e.g. ensuring that client privacy is not being breached), and conversely that only the appropriate users have access to the ML components (e.g. to ensure that models are not tampered with, or subject to unauthorised access). A key element of validation at this stage is ensuring that any processes which are being used to enable the ML components to learn from, and respond to live production data work as expected, have the appropriate governance controls in place.

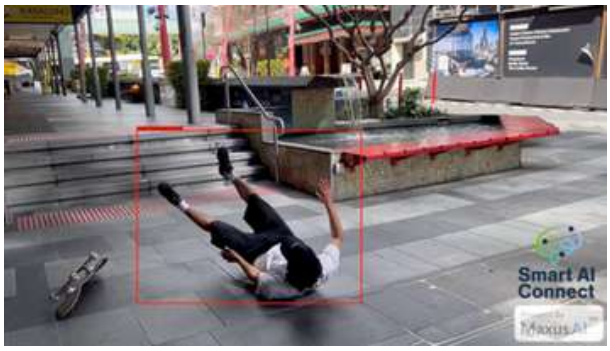
5. MONITOR PRODUCTION

By helping our clients to track the performance and integrity of their AI-enabled solutions from deployment, operation and maintenance, we can put in place the practical implementation of the governance controls identified as part of the risk assessment process. By monitoring residual risks, detecting model drift / sabotage, or simply measuring performance and cost, the hosting organisation can identify opportunities for further optimization and risk reduction.

While VDML provides an overall approach to AI governance, setting up the necessary compliance processes, ensuring that issues are detected and resolved promptly, and maintaining all of the essential evidence to support compliance reporting can be an overwhelming burden. Retrofitting the specific needs of an AI governance process into existing IT governance tools is also not ideal. This is where SmartAIConnect's Responsible AI (RAI) Framework fits in to enable safe, rapid and secure deployment and governance of AI-enabled systems at an enterprise scale.

CASE STUDY

As a case study, let's consider an organisation, Big B, that wants to improve occupational workplace health and safety by deploying AI-enabled cameras to detect any slip and fall incidents on its premises.



These cameras come configured with a machine learning model that can detect a slip and fall incident and report the occurrence back to base where a local security team can respond. The cameras also contain other algorithms which may provide additional features that are not required and should not be enabled. The governance challenge for embedded AI solutions is significant:deploying and maintaining an approved set of ML models to a network of potentially hundreds of AI-enabled cameras is not feasible without automated support.

Governance Context and Model Risk Assessment

To get started, the Big B team fill out an Organisation Questionnaire to ensure their company has thought through the considerations of deploying AI applications on their cameras. The Organisation Questionnaire contains 18 questions divided into 5 sections:

1. Privacy
2. Transparency and Explainability
3. Contestability
4. Accountability
5. Human Agency and Oversight

Following on from the organisation questionnaire, the Big B team also need to assess the specific risks associated with the slip and fall detection model by completing a **Model Questionnaire**. The purpose of the Model Questionnaire is to ensure that those responsible for the deployment of a given AI model have considered the impact of using that AI model on both the general public, and the organisation. The Model Questionnaire contains 25 questions divided into 6 sections:

1. Installation
2. Configuration
3. Fairness and Bias
4. Human Agency and Oversight
5. Transparency and Explainability
6. Accuracy

Many of the questions in the Model Questionnaire require the Big B team to review information provided by the model supplier. SmartAIConnect provide the model suppliers with a template model card to fill in, so that relevant information about each model can be tracked easily within the governance platform. It is up to the Big B team to review that information and accept any risks that the model introduces. One of the purposes of the model questionnaire is to guide the team to the important model information that should be reviewed.

Model Evaluation and Fine Tuning

Every AI Model in the Model Library has a Risk rating. SmartAIConnect calculate the risk rating using the results of a model evaluations based standard set of test data relevant to that model. Big B have a very diverse workforce and while completing the model questionnaire, the Big B team identified a need to ensure that there is no bias in the way the slip and fall model operates, so they elect to get some additional model performance evaluation carried out on their specific test data set. The model did indeed show some degraded performance with certain classes of input, and as a result Big B worked with the model supplier to improve the performance of the model until it met their required performance levels. There were still some limitations to the model's performance in low light situations, and those details were included in the updated model card.

Deployment and Governance

With overall context and risk assessment in place, the next step towards AI compliance for Big B is knowing what AI features are running on their cameras, when it got there, and who put it there. This is achieved through auditing the AI apps on the cameras, and managing the AI app deployment process. In order for a camera to be compliant, the Big B team need to either:

- ensure that there are no AI Apps on the camera, or
- ensure the AI apps on the camera have been deployed through the RAI Framework

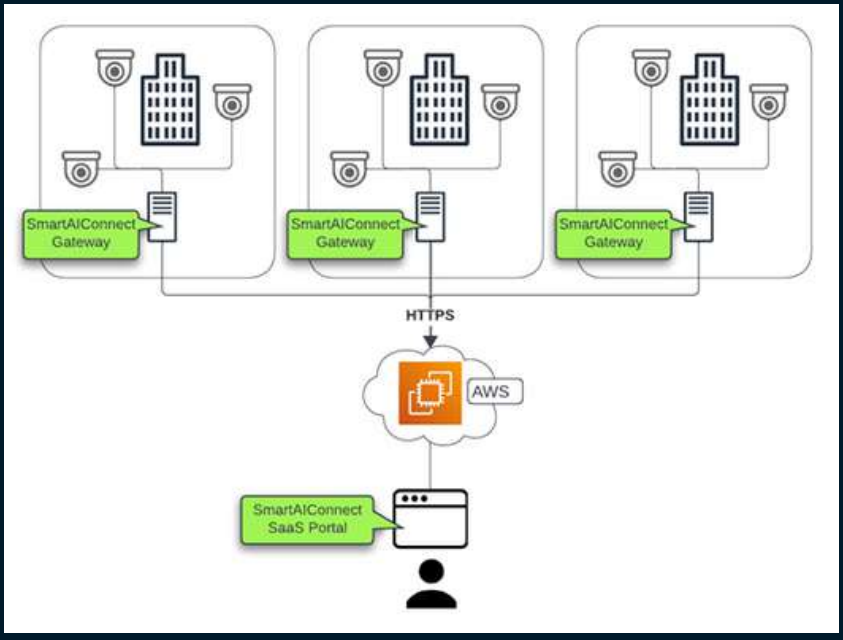


Figure 2 Camera governance network

Ensuring that only approved AI applications are running on the camera network is a key element in the governance of the system: all too often there can be a communication gap between technical teams and the business stakeholders who are ultimately accountable. By using the ModelOps features of the RAI Framework to deploy AI apps automatically from the Model Library to the specified cameras, only those apps which have been evaluated and approved for release are deployed. Similarly, the business is able to directly request installed AI apps to be removed from the camera network if needed. For all of these actions, an audit trail is recorded in the RAI Framework. In this way, the Big B team have been able to take their automated slip and fall detection system from concept to reality and roll it out across multiple sites in a responsible and well governed way.

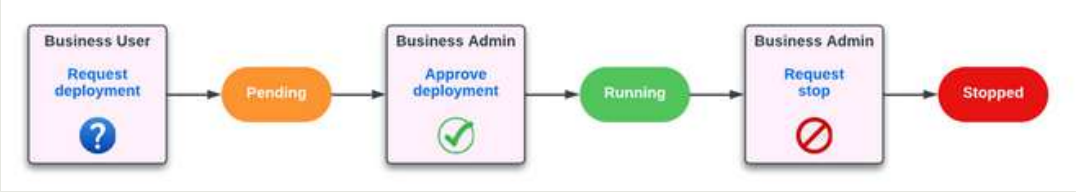


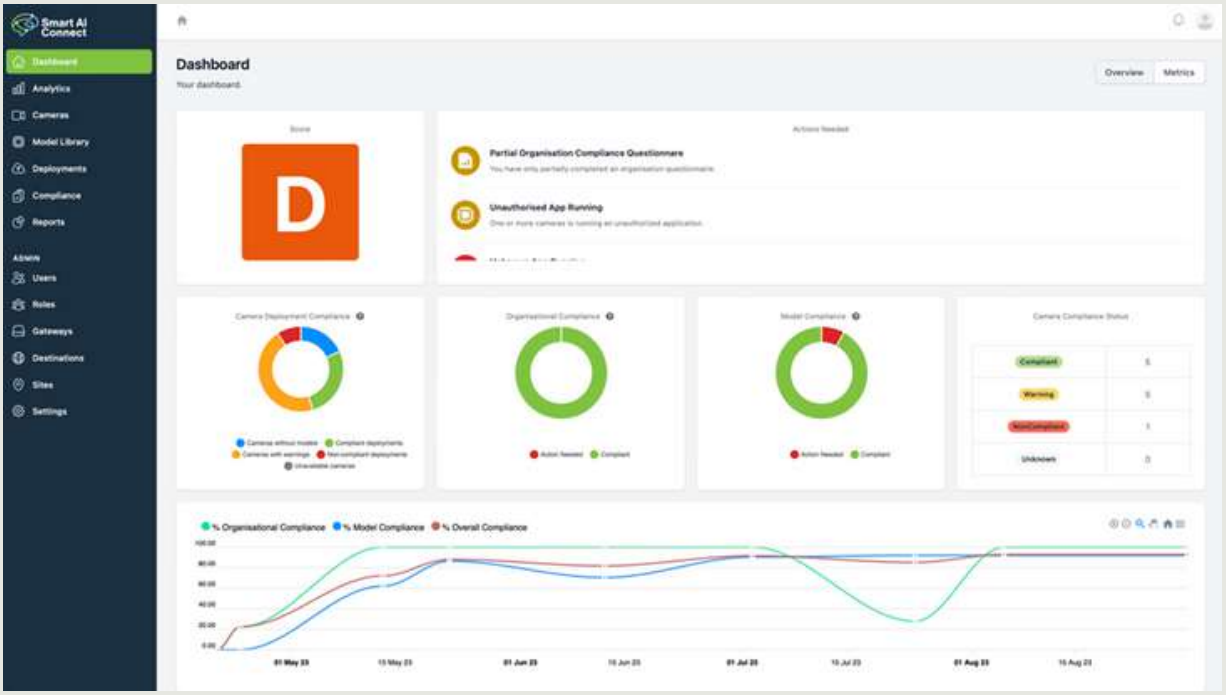
Figure 3 Model Deployment Process

Compliance Reports

The Big B team also need to prepare quarterly reports on their usage and deployment of AI. The platform can provide a compliance report as a point in time snapshot of the state of compliance of the Big B camera network, including:

- a snapshot of AI app deployments
- a snapshot of the Big B camera network
- the current state of the last completed Organisation Questionnaire
- the current state of the last completed Model Questionnaires
- a score based on the positive completeness of the compliance questionnaires

It is important to note that within the RAI Framework, compliance monitoring is a continual process. Changes to the camera network, including updates to installed AI apps, changes to the risk context or other process and organisational changes will trigger re-evaluation of the current compliance status and identify any items which need remediation. In this way, the RAI Framework functions not just as a passive reporting tool but as an active element of the day-to-day governance, risk and compliance process.



WRAPPING UP

By using the VDML process to guide their overall approach to adopting an AI-enabled system, specifically supported by the RAI Framework, the Big B team were able to provide confidence to their senior executives that the solution is delivering the desired work-place health and safety benefits while ensuring that the AI elements of the solution are operating safely and in compliance with the relevant legislation. The team were able to achieve this in two ways:

COMPLIANCE

1. Big B completed an **Organisational Compliance Questionnaire** to understand the risks of running AI within their business.
2. The camera (where models are deployed) have been deemed by Big B as **permitted to run AI**.
3. The AI model that they want to run has only been accepted into the model store because it contains a **model card** which details the purposes of the model, expected usage scenarios, any potential biases, limitations etc.
4. The person requesting the AI model has specified **where the data will be sent**.
5. The person who is authorizing the AI model has reviewed the above model card and accepted the risks of deploying the model by completing a **Model Questionnaire**.
6. Big B have chosen to provide the camera targets (ie. people walking in front of the camera) with details about the AI that is running by providing signage with a link to the **public model card**.
7. Privacy is preserved because **detections are done at the edge** - video footage with personally identifiable information is not sent to the cloud.
8. Regular **compliance reports** can be run to review the ongoing compliance status as new AI models are introduced and business practices change.

GOVERNANCE

- ·All **deployments** of AI to the camera are **recorded**: who/when/why/where.
- ·All **unauthorised installations** of AI to the camera are detected: what/when.

As the application of AI technology accelerates, good governance of these systems will be essential to building and maintaining trust with the public and meeting the regulatory requirements of the jurisdictions in which the technology is being used. By taking a validation driven approach to the use of machine learning technology, and leveraging a practical AI governance platform, organisations will be able to maximise the benefits of this technology while minimising the risk of unexpected or adverse outcomes.

K.J. Ross Associates Pty Ltd

https://www.kjr.com.au

Email: info@kjr.com.au

Dr Kelvin Ross

KJR Founder

Dr Mark Pedersen

KJR CTO

Smart AI Connect

is for Responsible AI

SmartAIConnect Pty Ltd

https://www.smartaiconnect.com

Email: info@smartaiconnect.com