A. RESPONSIBLE AI FOR DEFENCE (RAID) TOOLKIT

Guide for Defence Industry

Version 1.1
Consultation 2023

Preamble to the Toolkit

Defence industry plays a crucial role in the development and transfer of AI capability to Defence.

Defence and the ADF are responsible to the Australian government and the Australian people to conduct their activities in a lawful and ethical manner. This responsibility includes ensuring appropriate consideration of the risks arising from the design, development, acquisition and use of AI capability by Defence.

The novel nature of AI means that Defence will be required to develop and implement new control measures to ensure AI capabilities remain lawful, ethical, and safe. Control measures implemented in the design phase will be critical to ensuring AI capabilities are capable of performing their functions lawfully, ethically and safely. In short, AI capabilities will need to be legal and ethical by design.

This Toolkit provides a suite of Tools to aid Defence industry become more literate in the needs of Defence in fielding responsible AI.

Disclaimer

All inquiries can be addressed to info@tasdcrc.com.au, Trusted Autonomous Systems, Queensland, Australia.

TRUSTED AUTONOMOUS SYSTEMS

CONTENTS

INTRODUCTION

The purpose of this document is to introduce the Responsible Artificial Intelligence in Defence (RAID) Toolkit and provide information about how it is intended to be used, and to anticipate AI governance requirements and recommendations, with a particular focus on legal and ethical compliance that Australian Defence Industry can apply during the design and development of their AI capability.

WHAT IS THE TOOLKIT?

The Responsible AI for Defence Toolkit has three parts, with accompanying information, that explains how to complete the RAID Tools:

- RAID Checklist;
- RAID Risk Register; and
- Legal and Ethical Assurance Program Plan (LEAPP).

RAID Checklist

The RAID Checklist (the 'Checklist'): this is the entry point to ensuring anticipated AI governance and assurance requirements of Defence could be met. The Checklist provides a method to assist Industry in identifying and assessing AI risks and AI risk treatment frameworks likely to be relevant to acquisition of an AI capability by Defence.[1]

It requires that you answer a series of preliminary questions about the AI capability; which will prompt further actions associated with the legal and ethical risk presented by the capability. Depending on the answers to the Checklist questions, it will direct you to complete an AI Risk Register to record how minor risks have been resolved; or prompt completion of an AI Risk Register supplemented with a more detailed risk identification and management plan (a Legal and Ethical Assurance Program Plan or LEAPP).

RAID Risk Register

The RAID Risk Register:[2] describes any identified legal and ethical risks specific to an AI capability and the proposed treatment of those risks. It provides for the tracking of risks throughout the life-cycle of the capability. It tracks and records the 'how'; whereas the Checklist and LEAPP identify the 'when' in treatment and mitigation processes for complex RAS-AI and their accompanying systems risks.

RAID Legal and Ethical Assurance Program Plan (LEAPP)

The Legal and Ethical Assurance Program Plan ('LEAPP'): this is an iterative document that identifies and manages risks (focussed on legal and ethical risks) across the capability acquisition cycle for AI capabilities meeting certain criteria, as set out in the Checklist. The LEAPP will cover those risks that require deeper analysis, when compared to what is entered into the RAID Risk Register.

---

[1] Note: the definitions adopted in this Toolkit are derived from publicly available Defence frameworks and strategies; definitions are derived from best fit utilised in industry or likeminded states or organisations, and are attributed accordingly.

[2] Note: the Ethical Risk Matrix that appeared in the Method for Ethical AI in Defence has been changed to reflect the development in AI governance and assurance frameworks since the adoption of this Report; and has been nuanced to address Defence's broader acquisition risks for AI capabilities, and incorporates updated industry best practice in this field.

Reference material

This RAID Guide provides a general overview of how the Toolkit operates; information detailing how the Toolkit was developed. The sources of the principles in the framework can be found in the RAID Framework (PowerPoint). High level reference materials are also appended to the Framework.

The RAID Checklist and LEAPP include built-in guidance on how to complete the documents. External and pinpoint references are also contained within these template documents.

HOW DOES IT WORK?

At a glance

The Checklist asks a series of yes or no questions relating to your AI capability, its composition and its proposed operating environment. The answers to these questions guide you to the next appropriate Tool in the Toolkit: either an AI Risk Register; or an AI Risk Register accompanied by a LEAPP.

Once the RAID Checklist is completed, one of two options will apply:

- The qualities, characteristics and proposed operating environment of the AI warrant only a basic level of risk identification and management – complete the AI Risk Register; or

- There are qualities and characteristics of your AI capability and its proposed operating environment that require a detailed analysis of the risks and their management – complete an AI Risk Register supplemented by a LEAPP.

The Toolkit can be used to identify AI risks and their thresholds and how those risks have been or are to be managed to enable it to be capable of operating within Defence's system of control. It does not identify whether Defence should acquire an AI capability – that is a matter for Defence. However, using the Toolkit will enable you to demonstrate your methodology in risk identification and risk treatment and management for your AI capability, potentially providing a competitive advantage if your product is under consideration for acquisition by Defence.

RAS-AI systems in Defence are comprised of several components, which are the parts of the AI system that are required for it to function within a specific Defence environment.

Answering some basic questions about these components – through the RAID Checklist – provides an initial assessment about the complexity of the risk posed by the AI. Depending on the answers to the Checklist questions, the AI may only require the completion of a RAID Risk Register to record how the low legal and ethical risks associated with the design, development and use of the AI are handled. If any of the risks are considered complex, the Checklist will indicate that a LEAPP is required, either for discrete components of the AI, or for all components of the AI.

Legal and ethical risks associated with Responsible AI have been translated into a number of measurable elements. Measuring an AI capability against these elements enable an assessment as to the levels of legal and ethical risks and mitigation measures taken. The LEAPP poses a series of questions to enable an assessment to be made on how an AI capability performs against relevant measurable elements. It is used to inform assessments on any residual levels of risk by addressing questions posed by the LEAPP and proposes frameworks suitable to appropriately resolve or sufficiently mitigate those identified risks. The LEAPP is completed as an iterative document, recording how those risks are treated during the design and development process of the capability, for each stage of the One Defence Capability System (ODCS) processes and risk mitigation requirements.

In summary:

- Any capability being proposed for acquisition should consider completing a RAID Checklist at the earliest possible stage. This identifies where additional risk for components of the AI require treating to mitigate or resolve the identified risks.

- Identified legal and ethical risk related to the project are recorded in the RAID Risk Register, and updated as those risks are treated, mitigated or resolved.

- The LEAPP is completed for those higher risk components identified in the RAID Checklist, and mitigated during the design and development of the capability; by assessing whether the measurable elements have been appropriately addressed by the developer or whether further work needs be undertaken by the developer or Defence to properly mitigate those risks. It represents an iterative governance and risk mitigation process.

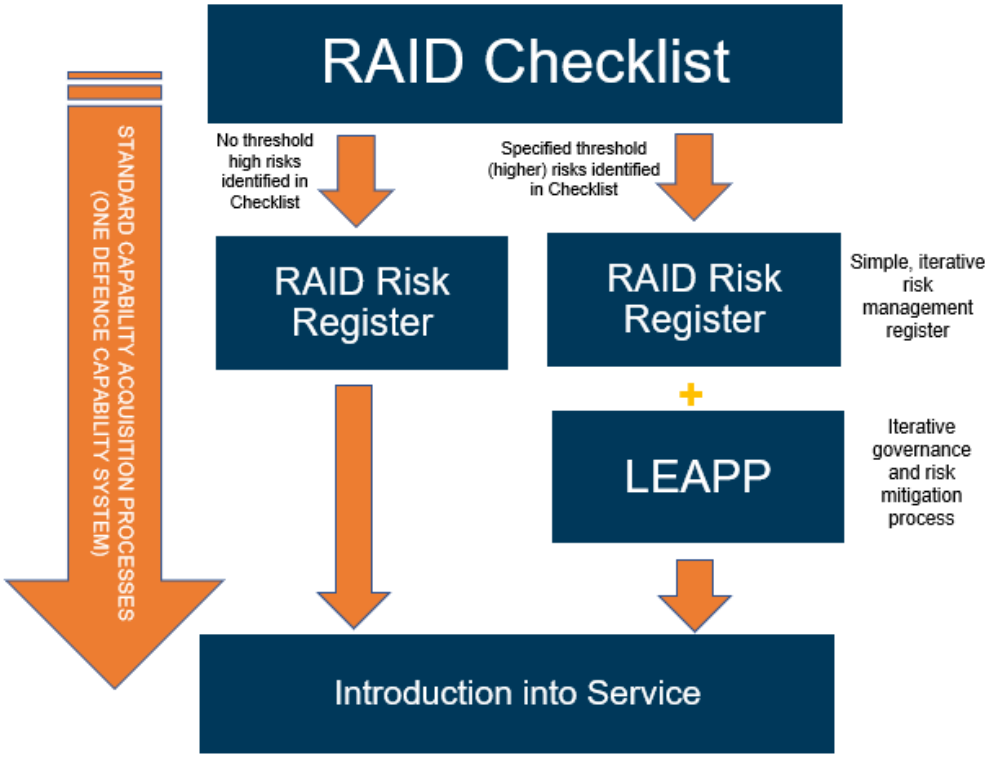

Diagram 1. How the Toolkit works

Elements and Components

There are many different principles and frameworks that have been created to address the risk, assurance, governance, and compliance obligations relating to the design, development, and acquisition of AI capabilities by governments. A common aspect of all is the intent to guide the design and development of AI capabilities in a manner that meets societal values and standards. From a design and development perspective (industry) and an end-user perspective (government), these theoretical concepts require translation into measurable quantities that can be: applied to the technology to inform assessments relating to its performance (compliance and assurance); and assist in decision-making relating to its acquisition for use.

The method adopted by the RAID, which builds upon the *Method for Ethical AI in Defence Report*,[3] is to use elements to articulate the overarching principles or characteristics that drive the adoption of legal and ethical AI for a military context. In the absence of a specific Australian Defence Framework, these elements have been derived from the principles and frameworks articulated in existing Australian AI frameworks; and those released by other nations with military-specific AI frameworks (such as the UK and US), and organisations (such as NATO) to ensure that all relevant principles in an Australian context are addressed; and in turn, converted into measurable outputs to enable them to be operationalised. These elements are then applied to the components that make up any AI system. For designers and developers of AI capability, adopting the elements articulated in this RAID Toolkit

[3] See DSTG, A Method for Ethical AI in Defence Report, 2020

can enable them to be confident that they can demonstrate how their capability will conform with relevant AI frameworks.

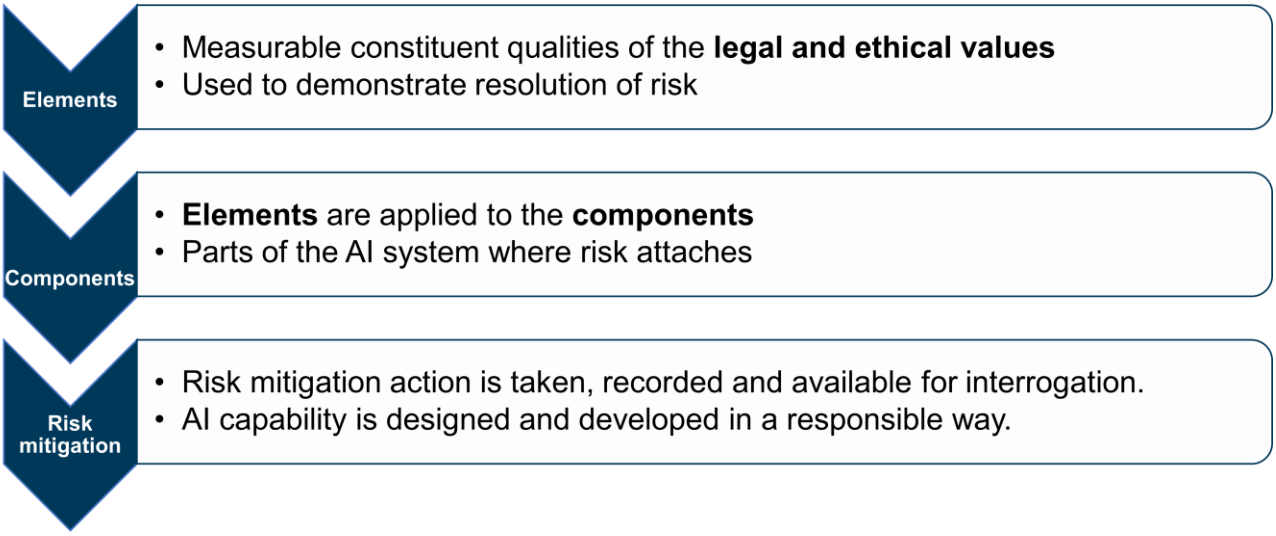

Diagram 2. Interactions between the Elements and Components

Elements

The 12 elements are the measurable parts of the risk mitigation process that ensure that the legal and ethical issues relevant to the life-cycle of the AI capability are addressed. They operationalise and categorise the ethical and legal issues identified within the components to be assessed and relate to the constituent qualities of the characteristics that are necessary for AI to be employed responsibly. They allow these legal and ethical risks to be addressed in a measurable, repeatable and recordable way.

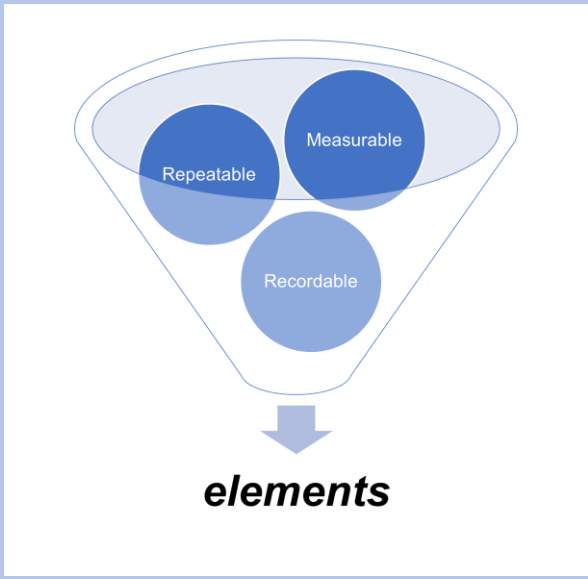

Diagram 3. Make-up of the elements

These elements are:

1. Responsible
2. Accountable
3. Understandable
4. Explainable
5. Reviewable
6. Reliable
7. Predictable
8. Compliant
9. Controllable
10. Integrated
11. Safe
12. Secure

Components

The components are the parts that make up an AI system, the system it operates in, and those things that are affected by the AI when used by Defence.

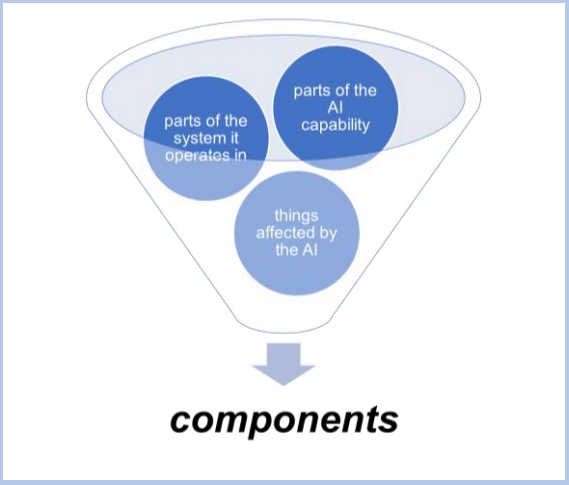

Diagram 4. Make-up of the components

They include:

- A. the **AI** itself;
- B. **design inputs** to the AI, which include things like data and values and standards;
- C. the **Human Machine Interaction (HMI),** which relates to the way in which the human controls, directs and interacts with the AI;
- D. **AI inputs** for the use of the AI, which includes things like sensor feeds or external stimulus;
- E. **AI outputs** when using the AI, which includes the effects generated by the AI;
- F. the **object of AI action**, which is the target of the AI outputs, and which is the thing or things that are effected by the AI system;
- G. the specific **use case environment** proposed for the capability; and
- H. the broader **system of controls** within which the AI system operates in the Defence environment.

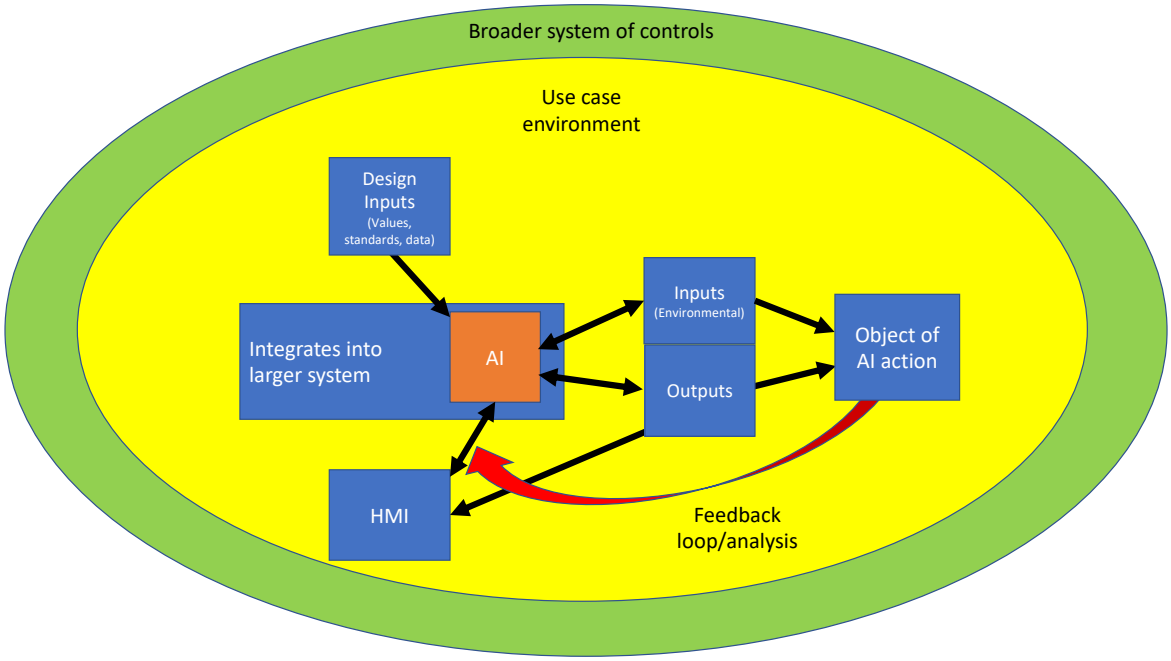

Diagram 5. How the components fit together

The above diagram demonstrates how the components of the RAS-AI capability fit together. Every element can be attached to an action or interaction between components, or a component themselves.

Elements comparison to Australian and international frameworks												
	Responsible	*Accountable*	*Explainable*	*Reliable*	*Understandable*	*Controllable*	*Secure*	*Compliant*	*Predictable*	*Safe*	*Integrated*	*Reviewable*
Australian AI Ethic's Principles [4]	*Fairness#*	*Accountability*	*Transparency and Explainability*	*Reliability and Safety*			*Privacy Protection and Security*	*Human, Societal and Environmental Wellbeing#*		*Reliability and Safety*	*Human-Centred Values*	*Contestability*
MEAID[5]	*Responsibility, Governance, Trust, Law, Traceability*	*Responsibility, Governance, Trust, Law, Traceability*	*Responsibility, Governance, Trust Law, Traceability*	*Governance, Trust, Law*	*Responsibility, Trust, Traceability*	*Responsibility, Governance, Trust, Law*	*Governance, Trust*	*Governance Trust, Law*	*Governance, Trust, Law*	*Governance, Trust, Law*	*Governance, Trust*	*Responsibility Governance, Trust, Law, Traceability,*
UK[6]	*Responsibility*	*Bias and Harm Mitigation*		*Reliability*	*Understanding*				*Bias and Harm Mitigation*		*Human-Centricity*	
US[7]	*Responsible*		*Traceable.*	*Reliable.*		*Governable*			*Equitable.*			
NATO[8]	*Responsibility and Accountability*	*Responsibility and Accountability*	*Explainability and Traceability*	*Reliability*		*Governability*		*Lawfulness*	*Bias Mitigation*			

Table 1. Elements comparison to international frameworks.

#Note: The first two principles of the Australian AI Ethics Principles apply to the extent that they are not displaced by the *lex specialis* of LOAC (that is, in armed conflict, it is permissible and in some cases necessary to utilise AI that will have an unfair outcome, provided, for example that outcome is authorised by law, if it is deemed militarily necessary and proportionate to the concrete and direct military advantage achieved.

[4] Department of Industry, Science and Resources, Australia's AI Ethics Principles, https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles
[5] DSTG, A Method for Ethical AI in Defence Report, 2020
[6] UK Ministry of Defence, P*olicy paper - Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence*, 15 Jun 22
[7] US DoD, *Ethical Principles for Artificial Intelligence*, 24 Feb 20
[8] NATO, Summary of the NATO Artificial Intelligence Strategy, 22 Oct 21, https://www.nato.int/cps/en/natohq/official_texts_187617.htm

WHEN CAN THE TOOLKIT BE USED?

The Toolkit should be used as early as possible during the design and development phase, for any AI capability that is intended to be sold to Defence. It is recognised that this decision will be made at various stages during the design cycle, however, the earlier potential Defence required governance frameworks are identified and risks mitigated, the less onerous certification requirements will be when introducing an AI capability into service.

The Toolkit moves beyond the Method for Ethical AI in Defence Report, to operationalise non-tangible concepts into practical guidance to assist Defence Industry in the design and development of AI capabilities intended for use by Defence. The Toolkit has been developed in consultation with Defence stakeholders, but it is not an endorsed Defence product. It is also not intended to replace any existing Defence capability acquisition or test and evaluation processes.

WHY RESPONSIBLE AI?

Defence and the ADF are responsible to the Australian government and the Australian people to conduct their activities in a lawful and ethical manner. This responsibility includes ensuring appropriate consideration of lawful, ethical and safety risks arising from the design, development, acquisition and use of AI capability by Defence and the ADF.

The novel nature of AI means that Defence will be required to develop and implement new control measures to ensure AI capabilities remain lawful and ethical. Control measures implemented in the design phase will be critical to ensuring AI capabilities can perform their functions lawfully and ethically. In short, AI capabilities will need to be legal and ethical by design.

Defence's acquisition and employment of AI that functions lawfully and ethically will also enhance the ADF's warfighting effectiveness. Acquisition and employment of AI capability that either cannot or might not function in this way will generate legal and ethical risk to Australia and the individuals involved in its use.

Likeminded States have also established strategies, principles, and in the case of Canada, an assessment framework to identify ethical risk. States are working on the integration of these new practices into existing design, development and acquisition processes to enable the early identification and treatment of legal and ethical risks associated with AI capabilities. It is anticipated that any framework developed by Defence will consider Defence's interoperability requirements with allies and partners, which includes the requirement for AI to be responsible.

CASE STUDY FRAMEWORK AND NARRATIVES

During the creation of the Toolkit a case study from each of the air, land and maritime domains will be undertaken and added to the resources when available.

The intent is these case studies will present vignettes using future, fictional technologies based upon a potential Defence RAS-AI capability, with excerpts of the Tools completed for the capability. The example case study is a truncated concept demonstrator only given the use of a fictional AI capability. Tools used for actual AI capabilities will be more detailed and specific.