D. RESPONSIBLE AI FOR DEFENCE CHECKLIST

Version 1.2 Consultation Draft 2023

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 can perform their functions lawfully, ethically and safely. In short, AI capabilities will need to be legal and ethical by design.

The Responsible AI for Defence Toolkit is aimed to provide AI literacy in the mitigation of legal and ethical risk anticipated during the design and development of AI capabilities to be acquired by the ADF.

Disclaimer

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

TRUSTED AUTONOMOUS SYSTEMS

CONTENTS

SECTION ONE: INTRODUCTION

This Checklist has been designed as the practical first step in the Responsible Artificial Intelligence (AI) in Defence (RAID) Toolkit. The RAID Toolkit is designed to help Defence industry determine if their AI product or those aspects of their product which includes AI capabilities requires additional analysis in the form of a Legal and Ethical Assurance Program Plan ('LEAPP'), to identify and address issues relating to responsibility, governance, trust, the law and traceability, in order to meet anticipated Defence acquisition and use requirements.

In addition to answering general questions about the risk attached to the AI capability and its functionality, the Checklist prompts you to answer a list of 'trigger' questions that will direct you to specific parts of the LEAPP in order to prompt more detailed consideration of risk-related issues identified in this Checklist and how they might be either treated or resolved.

The Responsible AI for Defence Toolkit has three parts:

1. **The Responsible AI for Defence Checklist (the 'RAID Checklist' or 'Checklist')**: this is the entry point to ensuring anticipated AI governance and assurance requirements of Defence are met. The Checklist provides a method to determine whether additional governance and assurance is likely to be required for relevant AI risks necessary for Defence to acquire an AI capability:[1] an AI Risk Register; or an AI Risk Register supplemented with a more detailed risk identification and management plan (a LEAPP).

2. **The RAID Risk Register**:[2] this describes any identified risks specific to an AI capability and its functionality and their proposed treatment.

3. **The Legal and Ethical Assurance Program Plan ('LEAPP'):** this is an iterative document that manages risks (focussed on legal and ethical risks) across the design and development phases of AI capabilities, meeting certain criteria, as set out in the Checklist. The LEAPP will cover those risks that require deeper analysis, when compared to the RAID Risk Register and as identified in the RAID Checklist. In the event your AI capability is being actively considered for acquisition by Defence, there is an option to update this document in a manner consistent with the process used by Defence's One Defence Capability System.

[1] Note: the definitions adopted in this Toolkit are derived from publicly available Defence frameworks and strategies; where definitions are not available, (as in the case of 'AI', for example), 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 anticipated, broader acquisition risks for AI capabilities, incorporating updated industry best practice in this field.

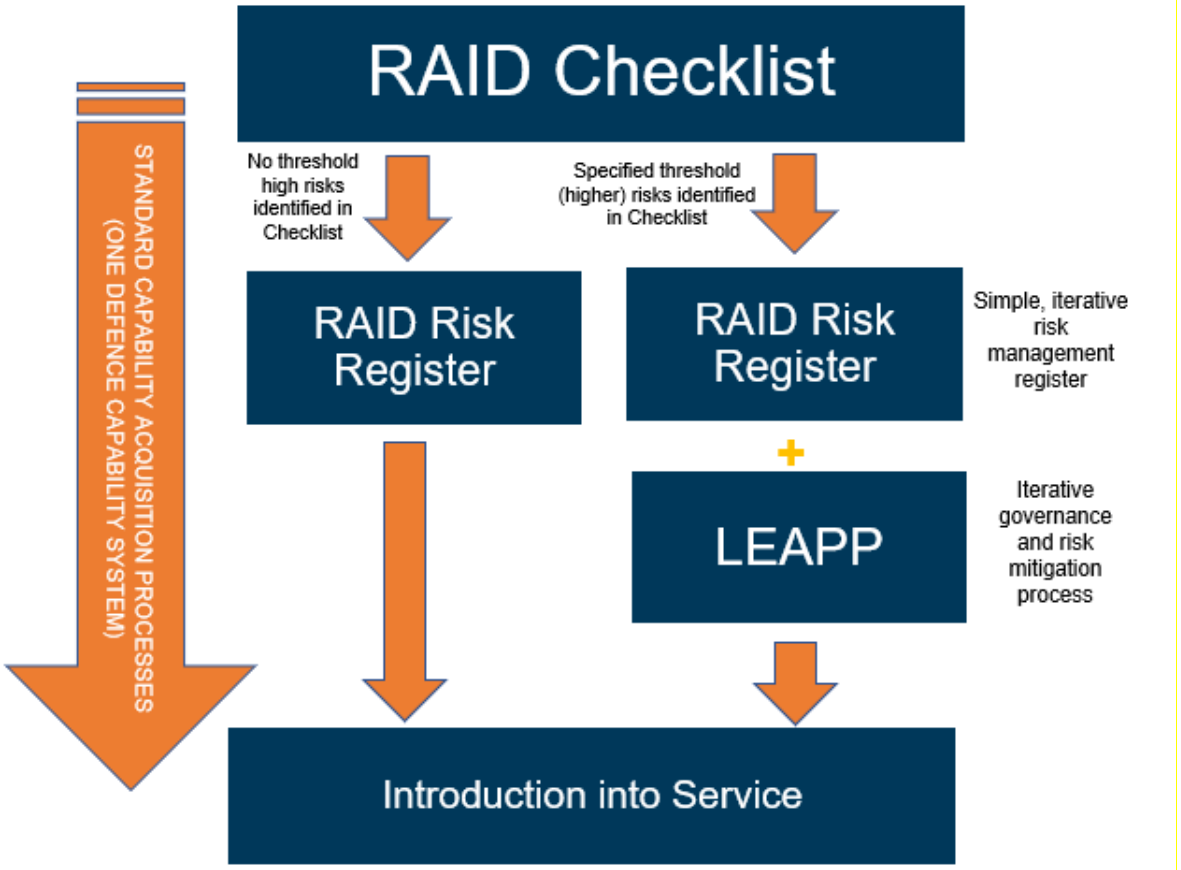

Diagram 1. How the Toolkit works

The Checklist is the first step for Defence industry and sovereign AI capability designers and developers to assess their product against the anticipated legal and ethical governance and assurance requirements of Defence as part of its consideration for acquisition and use of AI capabilities.

'AI' is a broad term used to describe a collection of technologies able to solve problems and perform tasks without explicit human guidance.[3]

For the purposes of the RAID Checklist:

- *'AI functionality' refers to the computational operations that the AI is designed or expected to undertake.*
- *An 'AI capability' refers to a product that comprises of or includes an element of AI functionality.*[4]

[3] CSIRO, Artificial Intelligence, *CSIRO Website*, available at < https://www.csiro.au/en/research/technology-space/ai >.
[4] Department of Defence, *ADF Concept for Robotics and Autonomous Systems*, Commonwealth of Australia, 2020.

Purpose of the Checklist

The Checklist is a self-assessment tool to assist Defence Industry in the development of AI capabilities intended for acquisition by Defence. The Checklist directs users to answer questions to identify considerations, processes and frameworks that may be required to ensure that any AI capability acquired by Defence is 'responsible'. Aspects of some AI capabilities will trigger consideration of more complex issues. These triggers are identified in the Checklist which then prompts users to complete relevant parts of the LEAPP, which in turn guides the consideration of those complex issues. All questions are focused on the identification, consideration and treatment of legal and ethical risks relevant to the use of AI for a Defence environment.

The early identification of risks and the concurrent development of risk mitigation strategies, through identifying frameworks, such as appropriate data management strategies, will avoid any need to re-engineer or repeat development processes to ensure Defence requirements are met. In this way, Defence can ensure it retains its technological edge through fast acquisition of assured technologies, while industry can confidently invest in the ongoing development of their leading-edge technology to meet Defence requirements in a cost-effective way.

Aligned to Defence's risk-based approach to new technologies, this Checklist is designed to alert Defence industry to the governance and assurance requirements needed to support the rapid design and development of AI capabilities to maintain Defence's technological edge, while balancing the need for new AI capabilities to be trusted to meet the Defence's values. In this way, Defence's AI capabilities comply with Australia's legal and ethical obligations, while representing best practice in the management of those capabilities.

When to use this Checklist

The Checklist should be completed when you are designing or developing any AI capability – either in conjunction with Defence or that you propose to sell to Defence.

It applies to any proposal that entails the use of an AI capability by Defence; it is not limited to those proposals involving the use of AI functionality in weapons, weapons systems or other means or methods of warfare.

It 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 anticipated Defence governance and assurance considerations are identified, and risks mitigated, the greater the competitive advantage and the less onerous certification requirements will be when introducing an AI capability into Defence service.

How to complete this Checklist

To complete this Checklist, fill out the details in ANNEX A – Responsible AI for Defence Checklist. The content required to be completed is explained in **Error! Reference source not found.**.

Additional guidance (and a PowerPoint version of this Checklist) can be found in the Responsible AI for Defence Framework that also forms part of this Toolkit.

The requirements of the Responsible AI for Defence Checklist[5]

ANNEX A – Responsible AI for Defence Checklist is a template designed to be completed for every AI capability intended for acquisition by Defence. Details as to how to complete this Checklist are contained in Error! Reference source not found..

The information requested in the Checklist will be used to inform your answers which will then identify if a LEAPP is required, and if so, which parts should be completed. It also functions as a record of the features and characteristics of the AI capability and its functionality, so that the acquisition agency can ascertain whether the risk resolution or mitigation undertaken is adequate. Further, developers and designers can use this information to assist in shaping their design and product development with legal and ethical compliance in mind; and the records of these actions can be provided to the acquisition agency, offering both a competitive advantage and easing the certification and integration into service of the capability.

Answering the questions in the Checklist will guide you in determining whether you should proceed to complete parts of, or a whole LEAPP to further consider and treat more complex legal and ethical risks relevant to your AI capability. A summary of these questions is contained at **ANNEX A – Responsible AI for Defence** Checklist.

Checklist questions, and the summary of the questions that trigger a LEAPP, are contained at ANNEX A – Responsible AI for Defence Checklist, below.

Other responsible AI documents required for AI capabilities

The Checklist will assist in preparing your AI capability for Defence acquisition through the identification, consideration and treatment of legal and ethical risks recorded by completion of an AI Risk Matrix or, for more complex issues, development of a more detailed LEAPP. This Toolkit provides templates for these documents.

[5] The Responsible AI for Defence Checklist is an evolution of the Ethical AI Checklist (Kate Devitt, Michael Gan, Jason Scholz and Robert Bolia, 'A Method for Ethical AI in Defence' (EAID), Defence Science and Technology Group, DSTG-TR-3786, January 2021). Note, the EAID Checklist was the product of a first-in-kind workshop looking at ethical (and legal) risks introduced by AI. The key difference between the MEAID and the RAID Checklists is that the latter addresses all governance (or control) risks, whereas the former is limited to primarily ethical risks. This is not to say that the RAID Checklist reduces the importance of addressing ethical risks, rather it accepts that ethical risks are important, but they are one of a number of control risks that need to be considered as part of Defence's relevant system of control for managing its capabilities.

SECTION TWO: GUIDANCE MATERIAL

The information below will assist you complete the Checklist. It is intended as a guide only.

Governance and assurance frameworks for AI are evolving, as is the Defence framework for AI systems. This means the Checklist will likely be updated on a regular basis to reflect contemporary requirements. You should ensure you are using the latest version of this Toolkit, by accessing it through the TAS website.

How the Checklist works

The Checklist asks you to answer questions relating to your AI capability. Your answers will assist you in deciding whether you only need to complete an AI Risk Matrix or whether the qualities and characteristics of your AI capability mean that you should complete the more detailed and iterative LEAPP.

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

- *The qualities or characteristics of the AI capability warrant only a basic level of risk identification and management – complete the AI Risk Matrix.*
- *There are qualities and characteristics of your AI capability that require a detailed analysis of the risks and their management– complete an AI Risk Matrix supplemented by a LEAPP.*

The Toolkit can be used to identify AI risks thresholds and how they have been, or are to be managed, so that your AI capability could be operated by Defence within Defence's system of control (see Diagram 1). Failure to identify the risks attached to your capability, and/or a failure to identify appropriate risk-management measures may erode your competitive advantage in the event your AI capability is under consideration for acquisition by Defence.

The basis of the threshold questions that require a LEAPP

By completing the Checklist, a requirement for a LEAPP to be completed will be identified based on identification of complex legal and ethical issues associated with your AI capability and a heightened risk for legal and ethical compliance associated with those issues.

The underpinning issues requiring a LEAPP include:

- **Risk to life.** By its nature there are higher risks to life – i.e., the AI capability is the control device for the combat functionality of a weapon or means of warfare or the AI capability is the control device for a life support or critical safety system.

- **Risk to rights.** By the inherent risk to people, their rights (including property rights), from the potential consequences of the normal or anticipated use of the AI capability – i.e., the AI capability supports targeting decisions; the AI capability undertakes network defence; the AI capability manages critical logistic programs.

- **Able to understand and explain how the AI works.** The way the AI capability undertakes its computational functionality involves higher risk – i.e., the AI capability functions as a 'black box'; the AI uses complex algorithmic processing that is not immediately transparent, understandable or explainable to humans, the AI functionality requires specialist expertise to operate.

- **Data inputs.** The AI capability's data input or interaction requirements generates increased risk – i.e., the AI capability relies on unsanitised data; the AI capability can implement decisions directly without human input; the AI capability can receive multiple different sources of internal, external and/or uncontrolled data.

- **AI undertaking judgment functions.** The anticipated operational position of the AI capability within Defence's decision-making framework – i.e., the AI capability implements decisions that can either make a human accountable for the decision, or conversely the use of the AI capability could potentially impede human accountability; the AI requires the human activating the capability to have specific levels of knowledge and expertise.

CHECKLIST CONTENT

Below are the Checklist questions, and an explanation of how to complete them. This section provides some examples of the content to be included and explains why these questions are relevant to assessing risk for an AI capability. Importantly, this section flags the types of answers that will trigger the need for a LEAPP to be conducted for the AI capability, rather than just an AI Risk Register.

Additionally, the completion of the Checklist will assist in identifying whether the use of an AI capability warrants additional detailed assessment. For example, in the event a specific LEAPP trigger is not met, the totality of multiple lower risks of the AI capability may warrant further governance and assurance considerations prompted by development of a LEAPP.

A. AI: What is the AI capability and how does the AI component function?

The purpose of questions in this section is to assist explaining what the AI capability comprises, and how the specific AI element will function. AI capabilities may be limited to software that requires human (directed) input to allow the AI capability to compute an output. Alternatively, an AI capability may be a conglomerate capability of which the AI component is one small element – such as a data fusion system on a tank. For conglomerate capabilities, it is important to identify where the AI functionality exists and its intrinsic characteristics.

A.1 What is the AI designed to do?

The purpose of this question is to identify the threshold issue that relates to the anticipated normal use of the weapon.

The consequences of this question will require a specific level of legal analysis and certification be Defence prior to being able to utilise the capability. Specifically, whether an Article 36 Weapon Review of the system will be required to be conducted.

All AI functionality:

- *A.1.1 is designed to enable combat functionality of a weapon[6] or means[7] of warfare,*
- *A.1.2 is designed to undertake safety critical functions,*

requires a LEAPP to be conducted. → Complete ALL LEAPP components.

[6] For Australian purposes, weapon is defined as: 'a device, whether tangible or intangible, designed or intended to be used in warfare to cause: a. injury to, or death of, persons; or b. damage to, or destruction of, objects.'
[7] A means includes sub-systems that enable the weapon functionality include AI decision support tools, AI enhances sensor and communications networks.

A.2 What decisions are addressed by the AI functionality?

The purpose of this question is to identify the risk that flows from the scope and type of decisions that the AI functionality is designed to undertake.

Is the AI:

- *using simple computational processes?*
- *undertaking binary decisions undertaken by mechanical operations?*

If so, complete the Risk Register. A LEAPP is not triggered by these features.

If the algorithm is:

- *A.2.1 is designed to replicate human judgement and discretion in decision making,*

- *A.2.2 undertaking novel decisions only made possible by complex algorithmic processing,*

- *A.2.3 making substantive or complex decisions,*

  a LEAPP is required. → Complete components: AI and AI Object.

A.3 If the AI functionality includes an ability to learn or modify some of its goals – what Test & Evaluation, Verification and Validation (TEVV) is required to ensure ongoing fitness for purpose?

Defence certification processes will require AI functionality to be capable of meeting particular standards during its TEVV. In some cases, the ability of AI functionality to learn or modify goals will require processes that will require ongoing TEVV over any changing functionality.

This question is focused upon the machine-learning/self-changing aspects of the AI functionality. TEVV is addressed in detail in H. System of control: control measures, system integration and AI frameworks.

AI functionality that:

- *A.3.1 can learn or modify its own goals triggers an ongoing requirement for TEVV,*

requires a LEAPP to be conducted. → Complete ALL LEAPP components.

A.4 Can the AI capability convert decisions into action? If so, is this subject to direct human intervention?

The purpose of this question is to identify if and when the exercise of the AI functionality to make a decision can be implemented as an action without direct human intervention. Consider if sub-components of the effect of the AI is implemented within a system without direct supervision, or if the AI's purpose and intended outcome can be affected without a human in direct supervision.

For example, targeting software based on probabilistic algorithms that achieves 99.9% accuracy in its targeting recommendations is kept on an air-gapped stand-alone computer, with inputs limited to human keyboard entry, and outputs limited to a recommendation on the computer screen. The AI functionality is not permitted to, or capable of, converting a decision into action.

If the AI:

- *A.4.1 permits decisions to be converted into action,*
-
- *A.4.2 implements decisions without direct human intervention,*

a LEAPP is required. → Complete components: AI, HMI, AI Outputs, and AI Object.

Direct human intervention includes the term direct human manipulation; and means that the human is capable of intervening directly in the operation of the AI, as compared to an AI that may operate based solely on other control measures established under the system of control. This includes previously established controls, such as coding, soft-locks, hard-locks, operating parameters.

A.5 What form of AI technique, machine learning technique or algorithmic processing is used?

The purpose of this question is to identify the specific type or characteristics of AI technology that are used in the AI functionality (i.e., deep neural processing etc.).

AI functionality that:

- *A.5.1 utilises probabilistic methods to compute a decision based upon incomplete or uncertain information,*

- *A.5.2 operates using an AI model computational processing that cannot be immediately understood or explained – for example, black box functionality, or through complex machine learning such as neural networks, or deep neural processing etc*

- *A.5.3 operates using an AI model or computational processing that is not reviewable,*

- *A.5.4 has embedded values and standards to produce its output,*

requires a LEAPP to be conducted. → Complete ALL LEAPP components.

The purpose of this questions is to identify the origins of the source code driving the AI capability. For example, is it open source, does it have propriety interests attached to it; or other considerations in relation to its genesis, such as the model was originally designed for another purpose and was adjusted to suit the Defence use case proposed. Depending on the answer to this question, there may, for example be impacts of adaption of open-source code in terms of safety, security and accountability that require further analysis.

A.6 What is the source of the AI functionality's code or model?

The purpose of this question is to ascertain if there are any proprietary concerns, secondary data provenance issues or programming history that creates legal risk for the use of the AI capability as described.

AI functionality that:

- *A.6.1 is derived from open-source, proprietary/commercial, bespoke, self- or third-party managed code,*

requires a LEAPP to be conducted. → Complete ALL LEAPP components.

A.7 What mathematical model is it based upon?

The purpose of this question is to identify the mathematical model, and any issues with its adequacy, used to represent the reality for the AI functionality. Any model used by AI functionality will only provide an approximation of the environment that is designed to operate in. Consequently, as the model affects the precision of the AI functionalities estimations, it is important to identify the adequacy (measure of precision or confidence level) of the model it uses. Issues with the adequacy of the model, including control measures to account for this imprecision, should be identified here.

AI functionality that:

- *A.7.1. relies on a mathematical model that is imprecise (requiring control measures to account for the imprecision)*

requires a LEAPP to be conducted. → Complete components: AI, Development inputs, HMI, System of control.

B. Development inputs: what is the composition of the AI functionality?

The purpose of this section is to understand the complexity and structure of the algorithms, and interlinked hardware that constitute the AI functionality. The more complex the computational systems, the more opaque the computational processing is, the more likely a LEAPP is required.

This issue is addressed by answering the following sub-questions:

B.1 What are the AI capability data sources?

The purpose of this question is to confirm whether data sets, external to Defence, were used in the development input to the AI functionality.

Data provided to you by Defence should be assumed to have been internally validated and scrubbed for privacy and accepted for use in AI development.

When answering questions in the LEAPP, this will confirm whether unhygienic, poor quality, unstructured, insecure, rights rich or regulated data was used in the development of the AI capability.

AI functionality that:

- *B.1.1 uses data that was not provided by Defence, for development, training, or certification,*

requires a LEAPP. → Complete component: Development Inputs.

B.2 What is the AI capability's data quality?

The purpose of this question is to identify if unhygienic, poor quality, unstructured, insecure, rights rich or regulated data was used in the development of the AI capability.

When answering this question, designers and developers are encouraged to remain objective about their capability and consider referring to the Defence data standard.

An AI capability for which the developer:

- *B.2.1 cannot describe its data structure, cleaning and bias mitigation process, original data owner, data steward, storage access and security and data rights,*

requires a LEAPP. → Complete component: Development Inputs.

C. Human Machine Interaction

The purpose of this section is to identify how the AI capability (and more specifically the AI functionality) interacts with the human operators across the spectrum of human involvement. It addresses what type or types of human-AI-interfaces exist for the AI functionality or for the AI capability.

HMI is the how the capability is used by humans. It incorporates the processes and procedures that direct the

While there will be overlap between this component and the H. System of control: control measures, system integration and AI frameworks component, this component is focused upon the physical and virtual human-machine connection between the operator(s) of the AI and the capability.

This issue is addressed by answering the following sub-questions:

C.1 What is the AI capability interface?

The purpose of this question is to identify the AI-human interface AI proposed for the capability. An interface is the point at which the system is controlled by, or effected by, a human operator.

AI functionality that:

- *C.1.1 does not have a direct human interface during operation of the AI capability,*
- *C.1.2 has a temporal or geographical dislocation between its interface and effect caused by the AI,*

requires a LEAPP to be conducted. → Complete component: HMI, System of control.

For the purposes of this question, a direct human interface is a physical system whereby a human can directly provide feedback and interact with the AI functionality. This may allow data input or may only allow observation of the operation of the AI functionality.

D. AI Use Inputs

The purpose of this section is to identify the foreseeable inputs for the AI capability to operate when in use. Input is the data that is required for the AI functionality to operate.

These considerations are separate, and additional to, considerations in relation to data inputted for the purpose of training or designing the AI model, which are addressed in Section

B. Development inputs: what is the composition of the AI functionality? However, in considering the AI Use Inputs, there is a need to analyse how the AI model has been trained during its design and thus all questions triggering a review for AI Use Inputs will also require review for Development inputs.

These inputs include the environmental data inputs, the user inputs (that is, what they are inputting into the system).

This issue is addressed by answering the following sub-questions:

D.1 Does the AI require user inputs (from humans) in order to operate?

The purpose of this question is to address the risk that is created as a result of a human providing input that will impact the operation of the AI capability. It is focused upon the point of operation because of the risk of that data being unverified, or incorrectly inputted. Because of the nature of AI, any incorrect input at the point of data entry is likely to result in amplified (or at least, copied) errors when the system undertakes analysis of that data. The risk relates to how the human error or source data error can be mitigated.

AI capability that:

- *D.1.1 requires a human operator to input instructions or data for it to operate*

requires a LEAPP to be conducted. → Complete components: AI, Development Inputs, AI Use Inputs, HMI.

D.2 Does the AI require data from the environment of its designed or intended use?

The purpose of this question is to identify the inputs (data and sources) that are available to the AI capability. An input is any form of data that is capable of being submitted to the AI capability as instructions or as a data source for exercise of the AI functionality. Input sources range from the human-AI interface (or AI capability interface) through to input from environmental sensors or sources that the AI capability can access (whether as part of the AI capability system or though links to external sources or sensors).

AI capability that:

- *D.2.1 requires data from the environment to operate as intended,*

requires a LEAPP to be conducted. → Complete components: AI, Development Inputs, AI Use Inputs, HMI.

E. AI Use Outputs

The purpose of this component is to identify the foreseeable outputs from the AI capability. Output is the data resulting from the execution of the AI functionality, or the ways that people or things receive the data resulting from the execution of the AI functionality (in whatever form the data is represented). In most cases the output will be limited to the provision of information to the human-AI interface (such as for decision support software) however, where an AI is designed (or inherently is required) to be connected to external outputs that results in independent action of effect, such as the use of force, greater consideration will need to be given to the risks that arise from data sharing and data use.

This issue is addressed by answering the following sub-questions:

E.1 What are the AI capability data outputs?

This question addresses the fundamental issue associated with this component: what are the data outputs?

AI functionality that:

- *E.1.1 sends output to external sources without being checked by a human first,*
- *E.1.2 produces an output involving data that is regulated by the law,*
- *E.1.3 is designed to (or consequentially) provides output that directly contributed to independent action of effect that is regulated by the law,*

requires a LEAPP to be conducted. → Complete component: AI Use inputs.

F. AI Object

The purpose of this section is to ascertain on what or whom the impact of the AI action will be. It requires consideration of the external actors that will be influenced or affected by the AI's actions within the environment of its anticipated use case.

It should be interpreted expansively and encompass consideration of who may be affected, and whether the AI will have a direct impact upon the rights of others, including property rights. The AI Object can also be an objects or data.

This issue is addressed by answering the following sub-questions:

F.1 Does the AI interact with humans?

Interaction includes providing direct inputs to the system, providing direct supervision to the system, or engaging with the system as a whole.

AI capability that:

- *F.1.1 interacts with humans as the object of the AI action,*

requires a LEAPP → Complete components: AI Object, AI Data Outputs and System of Control.

F.2 Can the AI generate effects that can directly affect third parties?

AI capability that:

- *F.2.1 directly affects the rights or obligations of persons or things not operating the system,*

requires a LEAPP → Complete components: AI Object, AI Data Outputs and System of Control.

G. Use Case Environment: Describe the military context in which the AI will be employed

While the AI capability may have more functions available for use, the use case will direct and refine the inputs to the AI functionality throughout the design and development process.

For example, an AI capability designed to enable the autonomous operation of an uncrewed underwater vessel that does not regularly communicate with human operators will generate different risks to an AI capability designed to assist a contract manager identify contracts predicted to go over budget.

Identification of the purpose of the AI capability and its use case will assist in identifying whether any war fighting functions are engaged.

The LEAPP contains more detailed questions and descriptors relating to the anticipated use environment of the AI capability. These questions are linked to risks already captured by other threshold questions regarding use of the AI capability or functionality. Accordingly, there is only one threshold question related to this component as a stand-alone trigger to conduct a LEAPP.

G.1 Is the AI intended to be used as a method of warfare?

Methods[8] of warfare is a broad term used to describe processes and applications that will result in the conduct of military operations during armed conflict.

Capabilities intended to be used in armed conflict trigger consideration of additional legal and ethical issues. While this question has been foreshadowed at A.2, in relation to use as a weapon or means or warfare, assessment as to whether an AI capability will undertake a method of warfare requires consideration of that AI capability in its anticipated normal use case and environment.

AI capability that:

- *G.1.1 is intended to enable a method of warfare,*

requires a LEAPP. → Complete ALL LEAPP components.

[8] A method is defined as Tactics, Techniques and Procedures applied to military operations…

H. System of control: control measures, system integration and AI frameworks

The purpose of this section is to identify how the AI capability fits within the broader system of control applied to military operations and activities. Questions focus on how the AI capability fits into the broader ecosystem of conducting military activities or operations, rather than the human-AI interface. Answers to the questions should consider the span of the relevant system of control which could range from factors such as education, training, and command and control requirements to maintain and use the AI capability through to managing distribution of the output from the capability.

This issue is addressed by answering the following sub-questions:

H.1 Explain how the AI capability (or AI functionality) integrates with other systems.

How does the AI capability nest within its proposed system of operation? For example, describing how a decision-support AI operates within the targeting cycle could include a description of who the analysed date is provided to, and which steps of the targeting cycle it is intended to influence or effects (as far as the designer/developer knows).

For example, an AI capability intended to assist Defence in personnel management processes by monitoring personnel leave credits and prioritising management engagement with personnel with excessive leave credits would need to be nested (integrated) into Defence's existing personnel management systems. It would support, but not perform, human decision-making but would be governed by prohibitive domestic law relating to privacy of personal information.

If the AI capability is:

- *H.1.1 is integrated within, or as part of, a larger system and sends output to that system without it being checked by a human first,*

a LEAPP is required. → Complete LEAPP components: AI, HMI, System of control.

H.2 What control measures are required for the capability to operate in its designed or intended use

The system of control is ordinarily directed by Defence, based upon the use case context, and the legal and ethical obligations surrounding the AI capabilities use in that instance. There may, however, be circumstances where particular processes of systems of control are required to be applied to the capability in every instance, to enable its use as envisaged by the designers and developers.

If the AI capability:

- *H.2.1 requires specific practice, process, procedure or intervention to restrict, limit or alter its functionality so that it can perform as intended,*

a LEAPP is required. → Complete LEAPP components: AI, HMI, System of control.

H.3 Has the AI capability or AI functionality been subject to independent review?

Building AI capabilities requires TEVV of the capability as it is developed. Defence processes are anticipated to require detailed scrutiny and TEVV by Defence (or on behalf of Defence), depending on the capability and its associated risk, to enable that capability to be certified for use. It is therefore important to identify if the system has been subject to independent TEVV.

If the AI is capability that:

- *H.3.1 has been subject to TEVV and has not been independently verified, or*
- *H.3.2 cannot be subject to an independent TEVV,*

a LEAPP is required. → Complete ALL LEAPP components.

If independent TEVV has been or is being conducted, further information in the AI Risk Register is required.

Note: The standard of independence for TEVV will be set by Defence based on factors including adequacy and relationship to designer/builders. Standards may also vary according to the nature of the AI capability created.

H.4 Do you need an expert to operate this AI capability?

The purpose of this question is to identify if the AI is complex in its operation, requiring particular experts to operate or input to the operation of, the capability. These questions related to risk as it is necessary to consider the risk imposed by non-experts utilising the capability; or the training burden associated with adopting a capability that requires training to be specialised or a limited cohort of operators. The majority of this issue will be addressed through the FIC acquisition process, however, the concomitant legal and ethical risks associated with this particular issue are flagged by the LEAPP.

If the AI capability:

- *H.4.1 cannot be operated without developer or contractor assistance (i.e. contracted specialist),*

- *H.4.2 cannot be designed or developed or operated without expert assistance,*

a LEAPP is required. → Complete ALL LEAPP components.

Experts for these purposes of this question include people with specialist expertise relating to AI and:

- *Military ethics,*
- *Decision science (including psychologists),*
- *Humanities and Social Sciences,*
- *Economics,*
- *Sociology,*
- *Anthropology,*
- *Law relevant to military operations,*
- *Human factors,*
- *Data science.*

I. LEAPP Summary: Is a specific LEAPP trigger met, or considering all the above, should a LEAPP be completed?

This section is designed to summarise the Checklist outputs, and collate whether a LEAPP is required, in addition to the AI Risk Register, and if so, which components of the LEAPP are required.

See Annex A, Part I for the Summary Table of the RAID Checklist Summary.

A copy of a RAID Checklist Summary can be found in a standalone document on the TAS website.

The summary of the questions from the Checklist that prompted completion of the LEAPP components are required to be translated into the LEAPP document, at the commencement of each Component.

By answering the questions of the Checklist, designers will understand some of the additional governance and assurance requirement for their proposed capability, if a specific LEAPP trigger is not met.

There are specified triggers throughout the Checklist that require a LEAPP to be conducted. It may also be that the totality of complexities associated with the AI capability generate risks that warrant the more detailed consideration provided by a LEAPP.

ANNEX A – Responsible AI for Defence Checklist

The below table includes the Checklist questions, and criteria to prompt completion of a LEAPP.

A	A. AI: What is the AI and how does the AI component function? **A.1 What is the AI designed to do?** • *A.1.1 is designed to enable combat functionality of a weapon[9] or means[10] of warfare* • *A.1.2 is designed to undertake safety critical functions* **A.2 What decisions are addressed by the AI functionality?** • *A.2.1 is designed to replicate human judgement and discretion in decision making* • *A.2.2 is undertaking novel decisions only made possible by complex algorithmic processing* • *A.2.3 is making substantive or complex decisions* **A.3 If the AI functionality includes an ability to learn or modify some of its goals – what TEVV is required to ensure ongoing fitness for purpose?** • *A.3.1 can learn or modify its own goals triggers an ongoing requirement for TEVV* **A.4 Can the AI capability convert decisions into action? If so, is this subject to direct human intervention?** • *A.4.1 permits decisions to be converted into action* • *A.4.2 implements decisions without direct human intervention* **A.5 What form of AI technique, machine learning technique or algorithmic processing is used?** • *A.5.1 utilises probabilistic methods to compute a decision based upon incomplete or uncertain information* • *A.5.2 operates using an AI model computational processing that cannot be immediately understood or explained – for example, black box functionality, or through complex machine learning such as neural networks, or deep neural processing etc* • *A.5.3 operates using an AI model or computational processing that is not reviewable* • *A.5.3 has embedded values and standards to produce its output*

[9] For Australian purposes, weapon is defined as: 'a device, whether tangible or intangible, designed or intended to be used in warfare to cause: a. injury to, or death of, persons; or b. damage to, or destruction of, objects.'
[10] A means includes sub-systems that enable the weapon functionality include AI decision support tools, AI enhances sensor and communications networks.

	A.6 What is the source of the AI functionality's code or model? • *A.6.1 derived from open-source, proprietary/commercial, bespoke, self or third-party managed code* **A.7 What mathematical model is it based upon?** • *A.7.1. relies on a mathematical model that is imprecise (requiring control measures to account for the imprecision)*
B	B. Development inputs: what is the composition of the AI functionality? **B.1 What are the AI capability data sources?** • *B.1.1 uses data that was not provided by Defence, for development, training, or certification* **B.2 What is the AI capability's data quality?** • *B.2.1 cannot describe its data structure, cleaning and bias mitigation process, original data owner, data steward, storage access and security and data rights*
C	C. Human Machine Interaction **C.1 What is the AI capability interface?** • *C.1.1 does not have a direct human interface during operation of the AI capability* • *C.1.2 has a temporal or geographical dislocation between its interface and effect caused by the AI*
D	D. AI Use Inputs **D.1 Does the AI require user inputs (from humans) in order to operate?** • *D.1.1 requires a human operator to input instructions or data for it to operate* **D.2 Does the AI require data from the environment of its designed or intended use?** • *D.2.1 requires data from the environment to operate as intended*
E	E. AI Use Outputs **E.1 What are the AI capability data outputs?** • *E.1.1 sends output to external sources without being checked by a human first* • *E.1.2 produces an output involving data that is regulated by the law* • *E.1.3 is designed to (or consequentially) provides output that directly contributed to independent action of effect that is regulated by the law*

	•
F	F. AI Object **F.1 Does the AI interact with humans?** • *F.1.1 interacts with humans as the object of the AI action* **F.2 Can the AI generate effects that can directly affect third parties?** • *F.2.1 directly affects the rights or obligations of persons or things not operating the system*
G	G. AI Use Case **G.1 Is the AI intended to be used as a method of warfare?** • *G.1.1 is intended to enable a method of warfare*
H	H. System of control: control measures, system integration and AI frameworks **H.1 Explain how the AI capability (or AI functionality) integrates with other systems.** • *H.1.1 is integrated within, or as part of, a larger system and sends output to that system without it being checked by a human first,* **H.2 What control measures are required for the capability to operate in its designed or intended use** • *H.2.1 requires specific practice, process, procedure or intervention to restrict, limit or alter its functionality so that it can perform as intended* **H.3 Has the AI capability or AI functionality been subject to independent review?** • *H.3.1 has been subject to TEVV and has not been independently verified* • *H.3.2 cannot be subject to an independent TEVV* **H.4 Do you need an expert to operate this AI capability?** • *H.4.1 cannot be operated without developer or contractor assistance (i.e. contracted specialist)* • *H.4.2 cannot be designed or developed without expert assistance*

LEAPP Requirement Summary	
I	I. Is a specific LEAPP trigger met, or considering all the above, should a LEAPP be completed? Note, LEAPP trigger questions are contained in the summary Checklist and at the end of each section. This section is a repeat of content derived in the sections above.

The AI functionality….

Serial	Threshold Question	Yes, or No?	LEAPP Component to be completed							
			AI	Development Inputs	HMI	AI Use Inputs	AI Use Outputs	AI Objects	Use Case	System of Control
A. AI: What is the AI and how does the AI component function?										
A.1.1	…is designed to enable combat functionality of a weapon[11] or means[12] of warfare									
A.1.2	…is designed to undertake safety critical functions									
A.2.1	…is designed to replicate human judgement and discretion in decision making									
A.2.2	…undertakes novel decisions only made possible by complex algorithmic processing									
A.2.3	… makes substantive or complex decisions									
A.3.1	…can learn or modify its own goals triggers an ongoing requirement for TEVV									
A.4.1	…permits decisions to be converted into action									

* If the answer to any of the Threshold Questions is 'Yes', then the white components of the LEAPP are required to be completed.

[11] For Australian purposes, weapon is defined as: 'a device, whether tangible or intangible, designed or intended to be used in warfare to cause: a. injury to, or death of, persons; or b. damage to, or destruction of, objects.'

[12] A means includes sub-systems that enable the weapon functionality include AI decision support tools, AI enhances sensor and communications networks.

Serial	Threshold Question	Yes, or No?	LEAPP Component to be completed							
			AI	Development Inputs	HMI	AI Use Inputs	AI Use Outputs	AI Objects	Use Case	System of Control
A.4.2	…implements decisions without direct human intervention									
A.5.1	…utilises probabilistic methods to compute a decision based upon incomplete or uncertain information									
A.5.2	…operates using an AI model computational processing that cannot be immediately understood or explained – for example, black box functionality, or through complex machine learning such as neural networks, or deep neural processing etc									
A.5.3	…operates using an AI model or computational processing that is not reviewable									
A.5.4	…has embedded values and standards to produce its output									
A.6.1	…is derived from open-source, proprietary/commercial, bespoke, self or third-party managed code									
A.7.1	…relies on a mathematical model that is imprecise (requiring control measures to account for the imprecision)									
B. Development inputs: what is the composition of the AI functionality?										
B.1.1	…uses data that was not provided by Defence, for development, training, or certification									
B.2.1	…cannot describe its data structure, cleaning and bias mitigation process, original data owner, data steward, storage access and security and data rights									
C. Human Machine Interaction (HMI)										
C.1.1	…does not have a direct human interface during operation of the AI capability									

Serial	Threshold Question	Yes, or No?	LEAPP Component to be completed							
			AI	Development Inputs	HMI	AI Use Inputs	AI Use Outputs	AI Objects	Use Case	System of Control
C.1.2	…has a temporal or geographical dislocation between its interface and effect caused by the AI									
D. AI Use Inputs										
D.1.1	…requires a human operator to input instructions or data for it to operate									
D.2.1	… is susceptible to uncontrolled input – including using data from AI capability sensors									
E. AI Use Outputs										
E.1.1	…sends output to external sources without being checked by a human first									
E.1.2	…produces an output involving data that is regulated by the law									
E.1.3	…is designed to (or consequentially) provides output that directly contributed to independent action of effect that is regulated by the law									
F. AI Object										
F.1.1	…interacts with humans as the object of the AI action									
F.2.1	…directly affects the rights or obligations of persons or things not operating the system									
G. AI Use Case										
G.2.1	… is intended to enable a method of warfare									
H. System of control: control measures, system integration and AI frameworks										

Serial	Threshold Question	Yes, or No?	LEAPP Component to be completed							
			AI	Development Inputs	HMI	AI Use Inputs	AI Use Outputs	AI Objects	Use Case	System of Control
H.1.1	…is integrated within, or as part of, a larger system and sends output to that system without it being checked by a human first									
H.2.1	…requires specific practice, process, procedure or intervention to restrict, limit or alter its functionality so that it can perform as intended									
H.3.1	…has been subject to TEVV and has not been independently verified									
H.3.2	…cannot be subject to an independent TEVV									
H.4.1	…cannot be operated without developer or contractor assistance (i.e. contracted specialist)									
H.4.2	…cannot be designed or developed or operated without expert assistance									
			AI	Development Inputs	HMI	AI Use Inputs	AI Use Outputs	AI Objects	Use Case	System of Control
I	LEAPP Components requirement summary: (Complete by indicating which components must be completed in the LEAPP, as triggered by the questions in at A-H, above.		☐	☐	☐	☐	☐	☐	☐	☐

ANNEX B – Additional resources

Domestic Legal Frameworks

The domestic legal framework applicable to the AI will be generally determined by the physical location of the operator(s) of the AI and the functions that the AI will be performing.

Specific legal advice should always be obtained to ensure lawful compliance during the development, testing and employment of an AI.

The list below is indicative of the types of laws that may govern some of the functions performed by an AI in Australia. It is a non-exhaustive list and does not include regulatory instruments that are also relevant to the legal use of AI capabilities. Notably, many of these pieces of legislation interact across domains and functions of the ADF; this list is only general in nature.

Commonwealth Legislation

Information Management Legislation

Australia has several Acts that regulate how information should be stored and handled; some of which impose specific obligations on how that information can be used or disclosed to other parties.

This legislation includes:

- *Privacy Act 1988*
- *Archives Act 1983*
- *Freedom of Information Act 1982*

Communication Legislation

Federal legislation which regulates telecommunications and radio communications including transmissions, access and interception includes:

- *Telecommunications Act 1997*
- *Telecommunications (Interception and Access) Act 1979*
- *Radio Communications Act 1992*

Legislation Governing the Conduct of ADF Activities (generally)

Legislation exists to give the ADF is statutory powers to conduct both its defence and general administrative functions as an element of government. This legislation has a wide and varied scope, ranging from powers to conduct domestic and offshore military operations, to limiting public access to certain areas of Australia to more mundane but necessary day-to-day requirements such as public accountability and management responsibilities.

This legislation includes:

- *Defence Act 1903*

- *Control of Naval Waters Act 1918*
- *Public Governance, Performance and Accountability Act 2013*
- *Work Health and Safety Act 2011*
- *Maritime Powers Act 2013*
- *Seas and Submerged Lands Act 1973*
- *Underwater Cultural Heritage Act 2018*
- *Environment Protection and Biodiversity Conservation Act 1999*
- *Environment Protection (Sea Dumping) Act 1981*

Accountability Legislation

There are numerous Federal acts that regulate conduct during armed conflict and criminalise conduct that amounts to breaches of Australia's international law obligations in terms of the laws of armed conflict. These laws require understanding and application to autonomous capabilities conducting any type of warfighting function. Separately, there are legislative requirement relating to the possession and transport of certain goods. Other legislation regulates, the use of and development of kinds of weapons, explosives and capabilities for a range of purposes including to align with Australia's international law obligations.

Legislation relevant to the conduct of armed conflict includes:

- *Geneva Conventions Act 1957*
- *Genocide Convention Act 1949*
- *International Criminal Court Act 2002*
- *Defence Force Discipline Act 1982*
- *Criminal Code Act 1995*

The weapons/capability specific legislation includes:

- *Explosives Act 1961 (and associated regulations)*
- *Space (Launches and Returns) Act 2018*
- *Submarine Cables and Pipelines Protection Act 1963*
- *Chemical Weapons (Prohibition) Act 1994*

Export Control Legislation

The regime controlling the import and export of specified goods that are limited because of their capacity to be used in armed conflict is coordinated through the Defence Export Control Office (DECO). DECO has separate guides and toolkits to assist Defence industry identify export control requirements for sovereign capabilities. The relevant legislation is also listed on their website. Noting that these import and export control obligations can also apply to 'dual use' technologies, particularly related to software, it is recommended that designers and developers be familiar with the obligations required of this regime.

This legislation includes:

- *Autonomous Sanctions Act 2011*
- *Defence Trade Controls Act 2012*
- *Customs Act 1901*

- *Weapons of Mass Destruction (Prevention of Proliferation) Act 1995*

State or Territory legislation

State or Territory legislation will also apply; the application of relevant legislation will be determined by physical location and the functions intended to be performed by the AI. For example, in Queensland additional legislation likely to be relevant for an 'identification' function include *Criminal Code Act 1899*, *Invasion of Privacy Act 1971*.

Specific legal advice should be obtained on laws applicable to the State or Territory of Australia in which the AI is being developed, tested or used.

International Legal Frameworks

International law relevant to the AI will generally be determined by the nation State employing the AI, the functions performed and the circumstances in which it is employed. While the specific international laws that apply will be determined by the government of that nation State, it is important for government acquisition purposes that AI is capable of complying with the international laws that may apply.

Specific legal advice should be obtained on international laws likely to apply to a specific nation State during armed conflict or during peacetime.

The following list is indicative of treaties that could apply to Australia's use of AI for warfighting purposes during an international armed conflict.

Multi-domain

General

- *Convention relative to the opening of hostilities. The Hague, 18 October 1907*
- *Treaty on the protection of artistic and scientific institutions and historic monuments (Roderick Pact). Washington 15 April 1935*
- *Convention for the protection of cultural property in the event of armed conflict. The Hague,1954*
- *Geneva Convention relative to the treatment of prisoners of war of August 12, 1949.*
- *Geneva Convention relative to the protection of civilian persons in time of war of August 12, 1949.*
- *Protocol I to the Geneva Conventions relating to the protections of victims of international armed conflicts.*
- *Protocol III additional to the Geneva Conventions relating to the adoption of an additional distinctive emblem, 2005.*
- *Convention on the Rights of the Child, 1989.*
- *Optional Protocol to the Convention on the Rights of the Child on the involvement of children in armed conflict, 2000.*

Weapons law treaties

- *Declaration renouncing the use, in time of war, of explosive projectiles under 400 grammes weight. St. Petersburg, 29 November – 11 December 1868.*
- *Declaration concerning expanding bullets. The Hague, 29 July 1899.*
- *Protocol for the prohibition of the use in war of asphyxiating, poisonous or other gases, and of bacteriological methods of warfare. Geneva, 17 June 1925.*
- *Convention on prohibitions or restrictions on the use of certain conventional weapons which may be deemed to be excessively injurious or to have indiscriminate effects. Geneva, 10 October 1980. Inclusive of the following Protocols:*
    - *Protocol I: Protocol on non-detectable fragments.*

  - *Protocol II: Protocol on prohibitions or restrictions on the use of mines, booby traps and other devices.*
  - *Protocol III: Protocol on prohibitions or restrictions on the use of incendiary weapons.*
  - *Protocol IV: Protocol on blinding laser weapons*
  - *Protocol V: Protocol on Explosive Remnants of War*
- *Convention on the Prohibition of Biological Weapons, 1972.*
- *Convention Prohibiting Chemical Weapons, 1993.*
- *Convention on the prohibition of the use, stockpiling, production and transfer of anti-personnel mines and on their destruction, 1997.*
- *Convention on Cluster Munitions, 2008.*
- *Treaty on the Non-Proliferation of Nuclear Weapons, 1968.*

Accountability

- *Convention on the Prevention and Punishment of Genocide, 1948.*
- *Convention against torture and other cruel, inhumane or degrading treatment or punishment, 1985.*
- *The Rome Statute for the International Criminal Court, 1998.*

Environment

- *Convention on the prohibition of military or any hostile use of environmental modification techniques, 1976.*
- *Arms Trade Treaty, 2013.*
- *Antarctic Treaty, 1959.*
- *Protocol on Environment Protection to the Antarctic Treaty, 1991.*

Domain-specific

Land

- *Convention respecting the laws and customs of war on land. The Hague, 18 October 1907*
- *Convention respecting the rights and duties of neutral powers and persons in case of war on land. The Hague, 18 October 1907.*
- *Geneva Convention for the amelioration of the conditions of the wounded and sick in armed forces in the field of August 12, 1949.*

Sea

- *Convention relating to the status of enemy merchant ships at the outbreak of hostilities. The Hague, 18 October 1907.*
- *Convention relating to the conversion of merchant ships into warships. The Hague, 18 October 1907.*
- *Convention concerning bombardment by naval force in time of war. The Hague, 18 October 1907.*

- *Convention relative to certain restrictions with regard to the exercise of the right of capture in naval war. The Hague, 18 October 1907.*
- *Declaration concerning the laws of naval warfare. London, 28 February 1909.*
- *The laws of naval warfare governing the relations between belligerents. Manual adopted by the Institute of International Law (Oxford Manual of naval war). Oxford, 9 August 1913.*
- *Process-verbal relating to the rules of submarine warfare set forth in Part IV of the Treaty of London of 22 April 19030. London, 6 November 1936.*
- *Convention concerning the rights and duties of neutral powers in naval war. The Hague, 18 October 1907.*
- *Convention on maritime neutrality. Havana, 20 February 1928.*
- *Convention relative to the laying of automatic submarine contact mines. The Hague, 18 October 1907.*
- *Geneva Convention for the amelioration of the condition of wounded, sick and shipwrecked members of armed forces at sea of August 12, 1949.*
- *International Convention for the Safety of Life at Sea, 1974.*
- *Convention for the Protection of Submarine Telegraph Cables, 1884.*

Air

- *Rules air warfare. Drafted by a Commission of jurists at The Hague, December 1922 – February 1923.*
- *Declaration prohibiting the discharge of projectiles and explosives from balloons. The Hague, 18 October 1907.*

Space

- *Treaty on Principles Governing the Activities of States in the Exploration and Use of Outer space, including the Moon and Other Celestial Bodies, 1967.*
- *Agreement on the Rescue of Astronauts, the Return of Astronauts and the Return of Objects Launched into Outer Space, 1968.*
- *Convention on International Liability for Damage Caused by Space Objects, 1972.*
- *Convention on Registration of Objects Launched into Outer Space, 1976.*
- *Agreement Governing the Activities of States on the Moon and other Celestial Bodies, 1984.*

Cyber

There are no cyber-specific instruments that guide the use of the cyber domain in military operations. Legal frameworks are derived from legal obligations in other domains.

Additional reference materials

In the absence of an Australian Government policy framework for the use of AI for Defence, this content provides examples of other AI frameworks or guidelines that are being implemented in both Australia and in other countries. While they are suitable for many Defence AI applications they may be unsuitable for weaponised AI, or for the design, development, acquisition, or use of AI supporting combat functionality. Accordingly, the Responsible AI Toolkit has incorporated the most appropriate elements of these frameworks and has translated them into both an Australian and defence context.

When under consideration, these frameworks must also take into account relevant Defence doctrine, such as ADF-P-0 *Military Ethics* and ADDP 06.4 *Laws of Armed Conflict*.

Australian Frameworks:
The following Australia AI Frameworks are relevant to the design, development and use of AI for Defence.

Responsible AI in the Military Domain (REAIM) Summit 2023 Call to Action Call to Action[13]

60 nations including Australia are signatories to the Responsible AI in the Military Domain Call to Action. Nations are committed to good practices, responsibility, accountability, inclusivity, security, stability, reliability, human involvement, human oversight and data quality. Nations agree to establishing normative ways of working including using risk assessments, ensuring an ongoing multi-stakeholder and international dialogue, working holistically and comprehensively, and sharing good practices and lessons learnt. Nations are expected to establish and employ norms, frameworks, policies, commit to and fulfil legal obligations, establish governance, national strategies, AI principles, data standards, and conduct research, testing and assurance. Education and training are expected to accomplish the aims and objectives of responsible AI and to avoid unintended harms and consequences of AI systems in the military domain. Nations are committed to ensuring that humans remain responsible and accountable for decisions when using AI in the military domain. Nations will achieve this through multi-stakeholder dialogue on best practices to guide the development, deployment and use of AI in the military domain to ensure an interdisciplinary discussion throughout of good practices and policies on responsible use of AI in the military domain.

Australia's Artificial Intelligence Ethics Framework[14].

Since 2019, the Commonwealth Department of Industry, Science, Energy and Resources has led the Australian development of an ethical framework. This framework describes eight *voluntary* AI Ethics Principles to be applied at each phase

[13] https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action
[14] https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework

of the AI system life cycle. These principles are intended to ensure AI is safe, secure and reliable. It is intended to reduce the risk of negative impacts of AI and ensure its use is supported by good governance standards:

- ***Human, societal, and environmental wellbeing:*** *AI systems should benefit individuals, society, and the environment.*
- ***Human-centred values:*** *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 protection 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.*
- ***Transparency and 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.*
- ***Contestability:*** *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:*** *People responsible for the different phases of the AI system life cycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.*

Defence Data Strategy 2021-2023[15]

Defence's Data Strategy is built upon five pillars:

- *Govern,*
- *Trust,*
- *Discover,*
- *Use and*
- *Share.*

These pillars are intended to guide data management across Defence. Effective, safe, and secure data is critical to the development and functioning of AI capabilities. The *Defence Data Strategy* indicates that, '[g]uidelines around the ethical use of data will be developed to ensure we have a shared understanding of our legislative and ethical responsibilities'.

A Method for Ethical AI in Defence[16]

A Method for Ethical AI in Defence (MEAID) technical report published by DSTG provides the following five 'facets' of Ethical AI in Defence and corresponding questions to provide a broad framework for defining legal and ethical requirements by AI stakeholders.

[15] https://www.defence.gov.au/about/strategic-planning/defence-data-strategy-2021-2023
[16] https://www.dst.defence.gov.au/publication/ethical-ai

- *Responsibility – who is responsible for AI?*
- *Governance – how is it controlled?*
- *Trust – how can AI be trusted?*
- *Law – how can AI be used lawfully?*
- *Traceability – How are the actions of AI recorded?*

Australia's *AI Ethics Framework* is reflective of and complementary to the facets of ethical AI described in MEAID.

NSW AI Assurance Framework[17]
NSW has published their own AI Assurance Framework that mandates AI projects to undergo ethical risk assessment.

Overseas AI Frameworks

United States: The US Department of Defense (DoD) are a world leader in adoption of ethical AI practices in Defence. In 2018 the US Government published its *AI Strategy,* which directed the DoD to create guiding principles for lawful and ethical AI. In March 2020 this led to the DoD adopting *Ethical Principles for AI*: responsible, equitable, traceable, reliable, and governable; accompanied by research into how to integrate them into DoD commercial prototyping and acquisitions programs. In 2022 DoD released Responsible Artificial Intelligence Strategy and Implementation Pathway.[18]

UK: On 23 October 2020, the UK Defence Science and Technology Laboratory (DSTL) published a 'Biscuit Book' titled *Building Blocks for AI and Autonomy.*[6] The book describes the nine Building Blocks of AI and autonomy. This was followed on 22 September 2021 by the UK Government's *National AI Strategy* which creates a 10-year plan to ensure that the UK keeps up with evolving AI technology. In 2022 Ministry of Defence released the Defence Artificial Intelligence Strategy policy paper [19].

Canada: Like the UK, Canada's Department of National Defense does not have a defence AI strategy, however, has published *Identifying Ethical Issues of Human Enhancement Technologies in the Military.* This publication creates a Military Ethics Assessment Framework to identify potential ethical risks in technology. This ethics assessment framework is like MEAID in that it is a technology-agnostic, risk assessment tool that emphasises legality, safety, accountability and trust as factors to be considered in identifying potential ethical issues.

[17] https://www.digital.nsw.gov.au/policy/artificial-intelligence/nsw-artificial-intelligence-assurance-framework
[18] Department of Defense Responsible AI Working Council. (2022, June). Responsible Artificial Intelligence Strategy and Implementation Pathway. U.S. Department of Defense. https://media.defense.gov/2022/Jun/22/2003022604/-1/-1/0/Department-of-Defense-Responsible-Artificial-Intelligence-Strategy-and-Implementation-Pathway.PDF
[19] Ministry of Defence. (2022, 15 June). Policy Paper: Defence Artificial Intelligence Strategy. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy

France: In 2019, the French Ministry of Armed Forces published their AI Task Force's *AI in Support of Defence Strategy.* This was the first military AI strategy published in Europe. It emphasises ethics and responsibility as essential elements of 'controlled AI' under the guidelines of 'trustworthy, controlled and responsible AI.' The French *AI Strategy* also creates a ministerial Defence Ethics Committee to oversee and advise on the adoption of AI.

NATO: On 22 October 2021, the North Atlantic Treaty Organisation (NATO) released its *Strategy for AI*.[7] It is intended to provide a foundation for NATO and its Allies to develop responsible AI, accelerate AI adoption, enhance interoperability, and protect and monitor AI technologies. While technology development occurs primarily at the national or bi-lateral levels, NATO emphasises that legal, ethical and policy differences could endanger interoperability. NATO's strategy includes six Principles for Responsible AI in Defence: Lawfulness; Responsibility and Accountability; Explainability and Traceability; Reliability; Governability and Bias Mitigation.

OECD: In 2019, the Organization for Economic Co-operation and Development (OECD) published their *Principles on AI*. Their principles were among the first to address and promote AI that is innovative, trustworthy and that respects human rights and democratic values. Australia's *AI Ethics Framework* specifically affirms the Government's commitment to the OECD Principles and evidence Australia's decision to become a founding member of the Global Partnership on Artificial Intelligence (GPAI). Widespread adoption of Australia's *AI Ethics Framework's* principles among business, government and academia will build trust in AI systems.[20]

UNESCO: On 24 November 2021, 193 countries adopted the Recommendation on the Ethics of Artificial Intelligence by UNESCO's General Conference at its 41st session. This agreement represents the largest global agreement on the ethics of AI to date[21].

[20] https://oecd.ai/en/
[21] https://www.ohchr.org/sites/default/files/2022-03/UNESCO.pdf

ANNEX C – Comparison of AI Ethics Frameworks

The below table demonstrates how international military AI frameworks, and the Australian civilian AI ethics frameworks have been integrated into the measurable elements adopted by the RAID.

Elements comparison to Australian, and international frameworks.												
	Responsible	*Accountable*	*Explainable*	*Reliable*	*Understandable*	*Controllable*	*Secure*	*Compliant*	*Predictable*	*Safe*	*Integrated*	*Reviewable*
Australian AI Ethic's Principles [22]	*Fairness#*	*Accountability*	*Transparency and Explainability*	*Reliability and Safety*			*Privacy Protection and Security*	*Human, Societal and Environmental Wellbeing#*		*Reliability and Safety*	*Human-Centred Values*	*Contestability*
UK[23]	*Responsibility*	*Bias and Harm Mitigation*		*Reliability*	*Understanding*				*Bias and Harm Mitigation*		*Human-Centricity*	
US[24]	*Responsible*		*Traceable.*	*Reliable.*		*Governable*			*Equitable.*			
NATO[25]	*Responsibility and Accountability*	*Responsibility and Accountability*	*Explainability and Traceability*	*Reliability*		*Governability*		*Lawfulness*	*Bias Mitigation*			

A side-by-side comparison[12] of the key ethical frameworks of MEAID, Australia's *AI Ethics Principles* (see below) and the OECD *Principles on AI* reveal common considerations among each of the frameworks. The comparison also reveals gaps in the areas of law and traceability. These gaps have been addressed, in a Defence context, within the RAID Toolkit.

Facets of Ethical AI in Defence	OECD Principles on AI	Australian Government's AI Ethics Principles
RESPONSIBILITY:		

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

Who is responsible for AI?	Human, social and environmental wellbeing: Throughout their lifecycle, AI systems should benefit individuals, society and the environment. Human-centred values: Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.	1. AI should benefit people and the planet by driving inclusive growth, sustainable development and well-being. 2. AI systems should be designed in a way that respects … human rights, democratic values and diversity.
GOVERNANCE: How is AI controlled?	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. Transparency and explainability: There should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system and can find out when an AI system is engaging with them. Contestability: 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 output of the AI system.	5. Organisations and individuals developing, deploying or operating AI systems should be held accountable for their proper functioning in line with the above principles 3. There should be transparency and responsible disclosure around AI systems to ensure that people understand AI-based outcomes and can challenge them.

TRUST: How can AI be trusted?	Reliability and safety: Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose Fairness: Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups Privacy protection and security: Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection and ensure the security of data.	2. [AI systems] should include appropriate safeguards – for example, enabling human intervention where necessary – to ensure a fair and just society. 4. AI systems must function in a robust, secure and safe way throughout their life cycles and potential risks should be continually assessed and managed.
LAW: How can AI be used lawfully?	No equivalent.	2. AI systems should be designed in a way that respects the rule of law.
TRACEABLILITY: How are the actions of AI recorded?	No equivalent (but implied).	No equivalent (but implied).

A side-by-side comparison[12] of the key ethical frameworks of MEAID and the NATO *Principles on AI* reveal common considerations among each of these frameworks as well. The comparison demonstrates a closer alignment than the OECD *Principles*.

NATO Principles of Responsible Use of AI	MEAID

Lawfulness: AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable.	**Law:** How can AI be lawfully used?
Responsibility and Accountability: AI applications will be developed and used with appropriate levels of judgment and care; clear human responsibility shall apply in order to ensure accountability.	**Responsibility:** Who is responsible for the AI? Addressed by education and Command.
Explainability and Traceability: AI applications will be appropriately understandable and transparent, including through the use of review methodologies, sources, and procedures. This includes verification, assessment and validation mechanisms at either a NATO and/or national level.	**Traceability:** How are the actions of AI recorded? Addressed by explainability and accountability.
Reliability: AI applications will have explicit, well-defined use cases. The safety, security, and robustness of such capabilities will be subject to testing and assurance within those use cases across their entire life cycle, including through established NATO and/or national certification procedures.	**Trust:** How can AI be trusted? Addressed by competency and integrity.
Governability: AI applications will be developed and used according to their intended functions and will allow for: appropriate human-machine interaction; the ability to detect and avoid unintended consequences; and the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour.	**Governance:** How is the AI controlled? Addressed by effectiveness, integration, transparency, human factors, scoping, confidence and resilience.

Bias Mitigation: Proactive steps will be taken to minimise any unintended bias in the development and use of AI applications and in data sets.	**Trust:** How can AI be trusted? Addressed by competency and integrity.