TRUSTED AUTONOMOUS SYSTEMS

RAID Toolkit

# F. RESPONSIBLE AI FOR DEFENCE (RAID) TOOLKIT

## Legal and Ethical Assurance Plan (LEAPP) – Template

Consultation 2023

**Disclaimer**
The Trusted Autonomous Systems Defence Cooperative Research Centre accepts no liability for the accuracy of the information nor its use or the reliance placed on it.

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

TRUSTED AUTONOMOUS SYSTEMS

## INTRODUCTION

The RAID LEAPP Template provides a framework for actions industry can take to implement responsible AI through the incorporation of legal and ethical considerations in the design, development, use, and evaluation of AI systems.

The unique nature of military operations and the laws of war mean that the legal and ethical compliance requirements for use of AI capabilities in armed conflict must be integrated into an AI capability from its inception, rather than added on once the capability has been acquired for use by Defence.

The complexities associated with integrating legal obligations into AI capabilities and ethical considerations of the use of AI capabilities in military operations means that there will be a need to develop a detailed and considered plan as to how legal and ethical risks associated with AI capabilities have been addressed during the development of an AI capability.

This document provides a template to assist in the development of a Legal and Ethical Assurance Program Plan (LEAPP). It contains nested guidance materials on how to complete a LEAPP, including information on how to complete each section.

The LEAPP provides a framework for addressing the specific legal or ethical risk(s) associated with your AI capability by guiding you to relevant measurable elements, that if addressed, will mitigate or resolve identified risks. It also offers suggestions for providing quantifiable output to aid further assessment of each risk.

It does not replace extant Defence acquisition documents or contract requirements. Rather, this document will foreshadow and prompt you to complete legal and ethical assurance and risk mitigation tasks, with prompts to engage with Defence on those processes, in order to meet eventual Defence certification and introduction into service requirements.

This document is intended to be iterative: you will be required to revisit your risk mitigation strategies as the design and development of your AI capability progresses. In the event your capability is considered for acquisition by Defence, it also prompts reviews and updates of the LEAPP at trigger points consistent with Defence's acquisition process using the One Defence Capability System. Using this approach which provides detailed, recorded consideration and treatment of legal and ethical risks relevant to Defence, should provide your capability with a competitive advantage during Defence's acquisition process.

This document is a live document. It will continue to be updated as improvements occur to AI design, assurance and governance practices and Defence's framework for the acquisition, governance and use of AI capabilities evolves. Accordingly, TAS welcomes feedback on this document, which can be directed to info@tasdcrc.com.au

# HOW TO COMPLETE THE LEAPP TEMPLATE

## The LEAPP as part of the RAID Toolkit

LEAPP Guidance Materials are contained within this document **in blue bold.**

The LEAPP should be completed only if identified as necessary after completion of the RAID Checklist. Completion of the RAID Checklist identifies which parts of the LEAPP require completion, as determined by the nature and intended use of your AI capability.

Record the identified risks and the steps taken to resolve or mitigate them in the project RAID Risk Register. Depending on the AI capability, this may require revising as the capability design and development progresses and the risk mitigation or resolution process can be more refined.

The LEAPP is an iterative document and should be updated as the development of an AI capability progresses.  In the event Defence is considering your AI capability for acquisition, there is scope to use the LEAPP to support the progression of your AI capability through the ODCS Gates or as it reaches specified or significant project milestones.[1]

## The LEAPP as a Risk Mitigation Tool

The LEAPP articulates the risks related to the legal and ethical compliance of AI capabilities in a Defence context. It does so by anticipating a hybrid performance-based and management-based regulatory method for Defence AI; specifically, the mitigation of risk will be consequent upon a combination of regulatory approaches to adequately address the entity of the risk of an AI capability operating within a defined environment.  Performance-based management 'imposes outcomes objectives on the targets of regulation rather than telling them exactly what actions they must take or technologies they must adopt'; whereas management-based regulation directs 'attention to risks and mandating the establishment of internal processes will reduce the probability of failures, even if that reduction may not be provable empirically'.[2]

The LEAPP specifies which risks are identified for mitigation, rather than simply assigning an overall risk to the system. Once the risks are identified by answering the questions outlined in the LEAPP, the accepted level of risk, mitigation action and the residual risk are entered into the Risk Register.  This allows for iterative risk reduction as the design and development progresses. It also allows separate risk assessment in the event different use cases for the AI capability are contemplated.

[1] Note: if iterations of the LEAPP are required in addition to the standard ODSC Gates, then the Defence Project sponsor or appropriate Defence representative will identify when those iterations will be required.

[2] This hybrid approach mirrors current cyber risk mitigation and regulation processes, see: Dempsey, 'Cybersecurity Regulation: It's Not 'Performance-Based' If Outcomes Can't Be Measured', *Lawfare*, October 2022.

It further allows for the articulation of risk to reflect the kind of risk being mitigation in a complex system: whether management or performance based. This LEAPP does not specify what the risk reduction methodology for the identified risk is. This is in part because there are myriad different risk mitigation methodologies; but more relevantly because the LEAPP contemplates a risk mitigation methodology for all AI capabilities intended for acquisition by Defence. This therefore requires a capability-by-capability; context-by-context; and risk-by-risk approach to risk.  It will also be a matter for the acquisition agency/user to determine if residual risk after risk mitigation/treatment is acceptable for the lawful and ethical adoption of the AI capability withing the specified use environment. This LEAPP process allows for the designers and developers to engage with Defence during the acquisition process to iteratively mitigate the risk of the capability, which will support the rapid certification of the capability prior to its introduction into service.

## Procedure for using this Template

1. Enter your name, company and date in the author field document control panel.
2. Replace [🤖 *bracketed text*] in any empty sections with your project information and/or document content.

Note that the bracketed text provides examples of the type of information to be provided within those brackets.

The document headers are set within the document style and will appear differently than the italicised text; and can be updated using the "Update Field" functionality.

3. Each section contains guidance materials, **in blue bold** next to the ℹ information icon.
4. Each section will be preceded with a copy of the RAID Checklist content relevant to that section.
5. Some sections also contain boilerplate in a standard, non-italicized text to use as a starting point. Review and modify any existing boilerplate content and add additional content as necessary to fulfill the requirement of each section.
6. Use the Styles *H1 – H5* for section headers, *Figure Caption* for captions below figures, and *Table Caption* for captions above tables so that the Table of Contents, List of Figures, and List of Tables can be automatically updated.
7. Define acronyms at the first usage in parenthesis after the expanded term and add to the Glossary.
8. To complete the LEAPP for submission/completion of Defence acquisition gateway:
    a. Update the acquisition gateway on the title page.
    b. Delete the template front page, these instructions pages, all instructions, and the detailed instruction notes and examples that

are identified with the information icon ℹ within the document sections that you are working on (that is, the relevant Component page required for completion).

c. Update the filename and file location in the document control panel by right clicking the field, then clicking “Update Field.”

d. Update the Headers and Footers to have the appropriate document title and version.

e. Delete the List of Tables or List of Figures if they do not contain any items.

f. Update the Table of Contents, List of Tables, and List of Figures by right-clicking and selecting “Update Field,” then “Update entire table.”

g. Have the document modified and reviewed as appropriate, and have each reviewer and modifier enter their name, organization, and date in the document control panel.

h. Submit the document for approval and go through the review/revision needed to obtain approval to finalize the document.

i. Repeat the review cycle and resubmit for approval as needed to obtain approval to finalize the document.

j. Enter the approver’s name, organization, and date in the *approved by* section of the document control panel.

k. Enter the approval date on the title page and in the footer throughout the document and update the revision history at the end of the document.

l. Remove the DRAFT watermark on the title page and the content pages by entering the Edit Header and Footer mode of the document and deleting the DRAFT image.

m. Print the document to PDF and review it outside of the Microsoft Word application.

n. Submit the Word and PDF versions of the document as final (for that particular version of the LEAPP, corresponding to that stage of Defence acquisition).

## Template Revision History

| Version | Date | Name | Description |
|---|---|---|---|
| 1.0 | 31/10/2022 | IWR | Original template of LEAPP. |
| 1.1 | 13/11/2022 | TAS | Initial template review. |
| Consultation | 16/03/2023 | TAS | Website for consultation/feedback |

TRUSTED AUTONOMOUS SYSTEMS

RAID Toolkit

# Legal and Ethical Assurance Program Plan for [🤖 insert project name]

Project Phase: [🤖 insert ODCS Phase or Gateway, if applicable]

Version: [🤖 insert version number]

Approval Date: [🤖 insert Approval Date]

# Part One – Introductory Information

## Overview

In this Part, preliminary questions about the AI capability will provide the basis upon which action can be identified to address the risks articulated in the latter parts of the LEAPP.

## Updates

It is not likely that the introductory information will require updating, unless the use case or anticipated use of the AI capability fundamentally changes. A fundamental change is where changes to the AI capability will result in engagement of different components or engagement of components in a way that would alter the original Checklist inputs.  If there is doubt about whether a fundamental change to the capability has occurred, the capability designer/developer should seek further expert advice on which parts of the LEAPP require revisiting because of those changes.

The information contained in this Part may, however, require updating as details of the AI capability develop during the design/development process. Updates should occur as changes become known.

## Contents of this Part

This Part contains the preliminary AI capability project data that sets out information relevant for a reader (likely to be one of many multidisciplinary experts engaged during the LEAPP process) to understand the project with sufficient detail so as to advise on identified risks and their proposed mitigation actions.  It also contains some basic document management information.

## Document control panel

This Panel is to be used in the event your AI capability is being considered by Defence for acquisition through the ODCS.

| | |
|---|---|
| File name | Legal and Ethical Assurance Program Plan Template.docx |
| File location | [🤖 insert file location/path] |
| Version | [🤖 insert version number] |

| | | |
|---|---|---|
| [🤖 insert Gate number] | | |
| Created by: | [🤖 insert author name, organisation and email contact] | [🤖 insert creation date] |
| | | |
| Reviewed by: | [🤖 insert author name, organisation and email contact] | [🤖 insert creation date] |
| | | |
| | | |
| | | |

| | | |
|---|---|---|
| Modified by: | [🤖 insert author name, organisation and email contact] | [🤖 insert creation date] |
| | | |
| | | |
| | | |
| | | |
| Approved by: | [🤖 insert author name, organisation and email contact] | [🤖 insert creation date] |

## CONTENTS

## List of Tables

[ 🤖 insert list of tables on completion]

## List of Figures

[ 🤖 insert list of figures on completion]

## LEAPP Completion Map

Check which Components and sub-category questions of the LEAPP are required to be completed for this iteration of the LEAPP (as copied from the RAID Checklist in first instance; or as required for updated versions of the LEAPP).

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| RAS-AI Component category | RAS-AI Component sub-category | | | | | | |
| All LEAPP required □ | | | | | | | |
| OR – COMPONENTS OF LEAPP REQUIRED: | | | | | | | |
| A. AI □ | A.1 AI □ | A.2 AI □ | A.3 AI □ | A.4 AI □ | A.5 AI □ | A.6 AI | A.7 AI |
| B. DevInputs □ | B.1 DevInputs □ | B.2 DevInputs □ | | | | | |
| C. HMI □ | C.1 HMI □ | | | | | | |
| D. AI Use In □ | D.1 AIUseIn □ | D.2 AIUseIn □ | | | | | |
| E. AI Use Out □ | AIUseOut-1 □ | | | | | | |
| F. AI Object □ | F.1. AIObject □ | F.2. AIObject □ | | | | | |
| G. AI Use Case □ | G.1 AIUseCase □ | | | | | | |
| H. System of Control (SOC) □ | H.1 SOC □ | H.2 SOC □ | H.3 SOC □ | H.4 SOC □ | | | |
| Annex A (Article 36 information) □ | | | | | | | |
| Annex C (Legal and Ethical Assurance Working Group Terms of Reference) □ | | | | | | | |

## Glossary and definitions

**AI** – Artificial Intelligence; a broad term used to describe a collection of technologies able to solve problems and perform tasks without explicit human guidance.

**AI functionality** – refers to the computational operations that the AI is designed or expected to undertake.

**AI capability** – refers to a product that comprises of or includes an element of AI functionality.

**Component** – a part that makes up an AI system, the system it operates in, and those things that are affected by the AI when used by Defence.

**Data sanitisation** – the deletion, amendment or cleaning of data to remove unwanted bias or error.

**Direct supervision** – means having a human in the loop, in the loop for exception, or on the loop, capable of influencing the outcome of the AI action.

**Element** – a measurable part of the risk mitigation process that ensures that the legal and ethical issues relevant to the life-cycle of the AI capability are addressed

**LEAPP** – Legal and Ethical Assurance Program Plan

**Means of warfare** – weapon or weapons system; a means includes sub-systems that enable the weapon functionality include AI decision support tools, AI enhances sensor and communications networks.

**Method of warfare** – the way or manner in which weapons and weapon systems are to be used.

**ODSC** – One Defence Capability System.

**RAID** – Responsible AI for Defence.

**RAS-AI** – Robotic and Autonomous Systems – Artificial Intelligence.

**Source Code** – Computer program in its original programming language, human readable, before translation into object code usually by a compiler or an interpreter. It consists of algorithms, computer instructions and may include developer's comments.

**Test Environment** – An environment containing hardware, instrumentation, simulators, software tools, and other support elements needed to conduct a test.

**Weapon** – 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.

## LEAPP Component Category and Sub-Category

ℹ The below list summarises the LEAPP Component Category, and each Sub-category related to that Component.

Within each Component page in the LEAPP, guidance materials for each sub-category will include:

- The Checklist trigger questions that prompt completion of the LEAPP component.
- Explanation of the legal and ethical risk being addressed by the questions
- Questions to identify the risks associated with that component, grouped by elements. If these questions identify risk, a risk mitigation strategy should be selected by the designed/developer and entered into the Risk Register. Note: Refer to Part Four for indicative guidance on types of risk mitigation methodologies that can be applied to each element.

## Part Two - Preliminary Project Information

### Project Overview

ℹ This section provides an overview of the capability to be delivered by this project.

It should describe the capabilities that the project is funded or proposing to deliver. Additional information relevant to the legal and ethical risk assessment of the project should be recorded, such as: any specified materiel solution; capability options; and the dates for the ODSC approvals.

[🤖 Insert project description here]

### Need for a LEAPP

ℹ This summarises the general reason that a LEAPP is required, as summarised at the end of the RAID Checklist.

[🤖 Insert Checklist Outcomes here – which summarised the need for a LEAPP.

For example:
As the AI Capability is a means or warfare, it requires all components of the LEAPP completing; OR

As the AI capability used publicly available personal information from the social media application Facebook to train the algorithm, the DevInputs section of the LEAPP requires completion.]

### Document Overview

ℹ Update the LEAPP Completion Map, above, to correspond to the parts that are being completed and summarise those parts here.

This document is the Legal and Ethical Assurance Program Plan (LEAPP) for the [🤖 *insert project name*]. This LEAPP is a plan that is intended to identify the legal and ethical risks related to the Defence procurement and introduction into service of this capability; and identified what mitigations are required during the development, delivery, integration, installation, verification/certification and support for the project as a whole. This is broken into stages of the One Defence Capability Management System (ODCS).

In conjunction with the RAID Risk Register, it documents certain processes and procedures for the technical management, procurement, installation and risk acceptance related to this project.

The LEAPP details are scaled in proportion with the scop, risk and complexity of the project; and the risks that are treated are those that are identified within the RAID Checklist.

The document is organised as follows:
Part One – Introductory Information
Part Two – Preliminary Project Information
Part Three – Legal and Ethical Assurance Program Plan

[🤖 Insert document overview description here, for example:
This LEAPP addresses the following components of the AI system:
AI, Development Inputs, HMI, AI Use Inputs, AI Use Outputs, Use Case Environment, and System of Controls.]

## Project Identification

ℹ Insert any project identification issued by Defence in this section. If Defence has not yet agreed to the purchase of this project, enter N/A.

[🤖 Insert any Project Identifiers issued by Defence. For example, any Defence Innovation Hub reference; or any Project name or sub-Project that the capability will be acquired under, if known.]

## Project Purpose and Scope

ℹ The project purpose should outline if the project includes design and development of a new AI capability, or adoption (and adjustment) of existing AI capabilities for this acquisition. It should include a scope of works in respect of the proposed technology and material solutions that forms the AI capability as described in the Overview.

The section should contain the following key information:

- the proposed technology and its components, including any off-the-shelf components;
- any material difference in options within the proposed AI capability being proposed for and its use case; and
- if relevant, describe the technology differences between the options.

Inclusion of a high-level graphical overview of the system could be helpful. This could be in the form of a physical layout diagram, a functional block diagram, or some other diagram that depicts the system and its environment.

This information could be copied from other documents and should mirror the project purposes and scope found in other documentation (as listed below), such as the Concept of Operations, the Quality Assurance Plan, Operations and Maintenance Plan, Technical Risk Assessment plan, etc.

[ Insert Purpose and Scope outcomes here; and include differing options of technology and material relevant to the proposed or accepted AI capability.]

## Relationship to Other Plans

If this LEAPP is largely repetitious of another required plan (for example, the Data Management or Systems Engineering Management Plan, include these documents as references).

[ Insert lists of the relevant plans here, whether Defence-initiated or drafted as part of the extant product development and design process.

Examples include:
DSTG Technology Risk Management Plan v1.0 dated xxx;
Defence Science and Technology Plan – Project XX, v1.0, dated xx;
Defence Innovation Hub Approval – Project xx, dated
Software Management Plan (SMP);
Integrated Support Plan (ISP);
Configuration Management Plan (CMP);
Verification and Validation Plan (V&VP)]

## Applicable documents

This section allows the listing of relevant engineering, design or acquisition documents to be included. If this LEAPP is largely repetitious of another required plan (for example, the Data Management or Systems Engineering Management Plan), include these documents as references.

Optionally establish a centralised repository to house and archive all of the documentation related to the project and provide the location here:

[ Insert document repository file path here.

Insert document list here, or in the below table:
System Engineering Management Plan, dated xx;
Defence Data Management Plan dated xx.

**Table 1: Referenced Documentation**

| Document Name | Identification number, Revision, Date. | Link or Contact Info to Obtain Document |
|---|---|---|
| Systems Engineering and ITS Architecture Procedure 123-456-123 | 20XX | Mr J. Jay, j.jay@email.com |
| Project Systems Engineering Mangaement Plan: Deliverable 1.0: Technical Memorandum | 06 Mar 2019, Version 2 | www.jjaybusiness.eg |

]

## Project Stakeholders

ℹ This section requires you to list the stakeholders of the Project; it includes those that have been engaged in the design, problem statement, operational use requirement scoping or other acquisition steps.

[🤖 Insert Stakeholders as relevant here. For example:
Service Director General; Service Project Desk; CASG Director General – relevant branch; CASG Project Director; Project Manager; Defence Innovation Hub POC; government funding bodies].

## Legal and Ethical Assurance Working Group

ℹ If acquired by Defence, where the SOW requires the Contractor to establish a LEA Working Group (LEAWG), the LEAPP shall include a plan for the LEAWG, including:

a. objectives and the terms of reference for the LEAWG;

b. the membership and points of contact for the LEAWG; and

c. arrangements for the conduct of LEAWG meetings.

[🤖 Insert if applicable:
Annex C contains the details of the LEAWG, including:
- Terms of Reference
- Objectives
- Membership and points of contact
- Arrangements for LEAWG meeting].

## Legal and Ethical Assurance in System Analysis

ℹ The LEAPP shall describe the participation of LEA in system mission analysis, determination of system functional requirements and capabilities, allocation of system functional requirements to human/hardware/software, development of system functional flows, and performance of system effectiveness studies.

The LEAPP shall describe the methods used by the Contractor to answer the following questions:

a. Who is responsible for the AI?

b. How is the AI controlled?

c. How can the AI be trusted?

d. How can the AI be used lawfully?

e. How are the actions of the AI recorded?

## References

ℹ This section allows for reference to any specific standards or design criteria or processes that have been adopted or references, relevant to legal and ethical compliance applicable to this project.

[🤖 Insert references here.]

## Part Three - LEAPP Requirements

ℹ

### Overview

This Part is where the risks identified in the Checklist will be addressed by application of the measurable elements. The resulting processes will either resolve or mitigate the identified risk and will assist you in explaining how you have reached your assessed residual risk levels (if any) in the event of consideration for acquisition by Defence.

In this way this Part is iterative – the risk identified in the Checklist is articulated in detail by reference to the technical specification of the components, the mitigation action will be proposed having regard to the triggered measurable elements, and the suitability of the proposed action will evolve as the capability design and development progresses. The content of this risk mitigation plan may be dictated or agreed with the appropriate Defence stakeholder if your capability is being acquired by Defence.

The outputs of this section (in terms of actions to be undertaken, or progress on action taken, to address identified risks) is recorded in the RAID Risk Register.

### Updates

The detailed assessment of the LEAPP will require review at each Gate of the ODCS. In the event that there is a material change to the project, acquisition pathway or agreed use case, the document will also require review.  It is recommended that you engage with your Defence project sponsor to discuss to what extent, and when, they may wish to discuss LEA issues associated with your capability. It is anticipated that this is a living document and based upon iterative engagement between Defence and the capability developer/designer.

### Content of this Part

The LEAPP is broken down into Components, which constitute all parts of the AI system, set within its intended military use case, to ensure all legal and ethical requirements are considered across the system’s lifecycle. Within each Component, there is a separate optional section to complete in the event the capability is being considered for acquisition by Defence through the ODCS.

Not all parts of the LEAPP will require completion for every AI capability. The RAID Checklist identifies whether a Component requires completion in this template. The RAID Checklist content is replicated at the commencement of each Component section.

Not all parts of the LEAPP will require completion for every AI capability. The RAID Checklist will identify whether or not a Component requires completion in this template. The RAID Checklist content is replicated at the commencement of each Component section.

RAID Elements will be listed against each specific risk mitigation requirement specific within the LEAPP:

1. Responsible

2. Accountable

3. Understandable

4. Explainable

5. Reviewable

6. Reliable

7. Predictable

8. Compliant

9. Controllable

10. Integrated

11. Safe

12. Secure

Suggested processes to achieve the measurable outputs directed in the LEAPP are contained in the guidance notes and summarised in Part Four of this Template. These methods, strategies or processes are not exhaustive. There are many more available software, hardware and system design methodologies that could be adopted to achieve this requirement, and methodologies suitable for one capability may not suit another. Accordingly, this Template does not dictate how the measurable output must be achieved but flags the requirement for a measurable output to be applied to mitigate the identified risk. Equally, if your capability is being considered for acquisition by Defence, Defence may direct the methodology to be used to mitigate, measure or record a particular risk. These risk mitigation requirements may also be nested within extant Defence acquisition documents or requirements, in which case Defence's information requirements can be extracted from the relevant parts of a completed LEAPP.

## A. The AI

ℹThe AI is the computational component of the system and the parts that make up the AI system. This is the software, hardware and platform that the AI operates.

It refers to the software process by which the information is analysed, whether it be via a convolutional neural network, rules-based, reactive, limited memory or theory of mind or self-aware.

It also includes the broader design of the AI capability, which consists of the design approach to its software, hardware and associated platform. That is, the AI is the computational component, but it is attached to a physical capability – whether that be a computer screen that provides the output data for a human to read, or a platform that is physical moved or changed as a result of the AI output.

### A. RAID Checklist Questions – Prompting AI Component Completion

ℹThis section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| A.1.1 | …is designed to enable combat functionality of a weapon[3] or means[4] 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 | …undertaking novel decisions only made possible by complex algorithmic processing |
| A.2.3 | … making substantive or complex decisions |
| A.3.1 | …can learn or modify its own goals triggers an ongoing requirement for Test & Evaluation Validation and Verification (TEVV) |
| A.4.1 | …permits decisions to be converted into action |
| 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 |

[3] 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.’

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

| Serial | Threshold Question |
|---|---|
| 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) |
| 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 |
| G.2.1 | … is intended to enable a method of warfare |
| 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, developed or operated without expert assistance |

## A.A. Preliminary information

### A.A.i Is your AI capability a weapon, means or method of warfare?

If yes, in addition to other components, complete Annex A – Article 36 Weapons Review requirements.

Completing Annex A, addresses the risk identified in the RAID Checklist, at A.1.1 by addressing the legal certification requirements for weapons and means of warfare as required by extant Australian policy and Australia’s international legal obligation.

### A.A.ii Please describe the composition of your AI functionality or functionalities. In particular identify:

- what AI functionality comprises the AI capability; and
- whether the capability is comprised of one AI functionality (or AI model) or several AI functionalities;
- if there are several AI functionalities – if any of these AI functionalities are interlinked;
- if the AI functionality is statistical, symbolic, or a hybrid combination of both;
- the nature and type of rules, mathematical models or algorithmic processes utilised;

- any ‘learning’ capacity (i.e. data-based machine learning), and if so:
    - what types of learning (i.e. ‘Does the system learn based on human-written rules, from data, through supervised learning, through reinforcement learning?’ (OECD);
    - Where and how the ‘learning’ or training occurs i.e. is the model trained centrally or in a number of local servers or “edge” devices?
    - Whether and how the AI functionality evolves and / or acquire abilities from interacting with data in the field?
- if the AI functionality is ‘generative, discriminative or both?’

Completing the above questions addresses the risk of AI complexities: the more complex the AI the greater the risk that it does not operate as intended.

A.A.iii Please describe the attributes of the model(s) you have chosen, including:

- What design principles were applied to the development of the model; including:
    - What specific design principles were used;
    - How were they used;
    - What tools, techniques or procedures were used to either:
        - build the AI;
        - To implement design principles within the model;
    - What was the source (i.e. open-source or proprietary) and weighting of the tools, techniques or procedures used in implementing the design principles;
    - How were the design principles integrated into the TEVV, or any ongoing performance evaluation;
    - What ongoing tools, techniques, or procedures are required to ensure maintenance of the design principles;
- What considerations were given to developments in model building, such as:
    - Risk-aware AI;
    - Explainable AI;
    - Uncertainty-aware AI;
- What model was chosen, and why, including;
    - What tasks – and how representative the model is;
    - What environments – and how representative the model is;
    - How amenable to rigorous design, testing and monitoring the model is when compared to other relevant models?
- Whether the model is universal, customisable or tailored to Defence’s data;
- What values, principles or subjective parameters (including weightings) were built into the model;
- To what extent/how are you - engaging experts (e.g., through public consultation, expert consultancies, published materials, etc., as appropriate) to integrate diverse perspectives into design practices, particularly when users and third parties’ rights may be impacted by use of the system?

- Expected model robustness;
- How the model was trained;
- Model performance testing and evaluation prior to certification or use, including identifying:
    - the tools, techniques and procedures used, including:
        - What they were,
        - How they were used;
        - Where they are sourced from:
    - what metrics you are using;
    - what is form and structure of the TEVV being used, and how was it implemented;
    - Mitigation measures applied to deviance or unintended and undesired results, including:
        - Qualitative or quantitative improvements in the data (such as acquiring more data or curating the existing data);
        - Qualitative improvement or replacement of the model;
        - Model parameter and weighting adjustments (such as altering the logic of the algorithms, changing the mathematical objectives, or changing the weightings);
        - Qualitative or quantitative changes to the AI outputs or AI objective (i.e. narrowing or altering the goals of the AI, changing the object of the AI output, or changing the intended AI output itself;
    - what performance characteristics are being evaluated (i.e. risk, bias, uncertainty, robustness);
    - How uncertainty is being measured and evaluated;
- Model performance monitoring after certification or acceptance into use – what is it, when did it start and finish, what is still required;

Completing the above questions addresses the risk associated with the accuracy and representativeness of the AI model.

### A.A.iv Please describe the attributes of the algorithms you used for your model, including:

- What was the source of your code (open-source, proprietary);
- How you evaluated your code;
- The parameters of the algorithms, including:
    - how they were chosen and implemented;
    - How the goals were chosen and implemented
- Whether the algorithms themselves (i.e. independent of the AI functionality or Ai capability) were evaluated, and if so how;
- The approach to code that included values, principles or subjective parameters, including:
    - Any known reductionism to convert these values, principles or subjective parameters into code;
    - Any known historical prejudices;
- With respect to algorithmic deviance (i.e. bias, discrimination etc):

  - How you identified it;
  - How you limited its effect in the algorithm;
  - How you ameliorated the effect of the algorithmic deviance in the overall AI capability;
- Whether the code is capable of being reviewed: and if so, how is this achieved (for instance can it be reviewed by a human or other algorithms.

Completing the above questions identifies risks associated with proprietary ownership of source code and risks related to the accuracy and appropriateness of coding.

## A. Elements Analysis

### A.1 Responsible

A.1.i What are the parameters of the decision-making that the AI functionality can undertake, and are those parameters transparent?
If no, assess the risk for lack of transparency (record in Risk Register).

A.1.ii What decisions of the AI functionality are permitted to be implemented without human approval or oversight?

A.1.iii Is the AI functionality limited to objective computational parameters or criteria?

Answering the above questions addresses the risk that the AI functions in a way that is not accurately reflective of the human intention in activating that AI to subsume that previous decision-making process; and the end effect is not the desired effect.

### A.5 Reviewable

A.5.i How are you documenting how issues with reliability and robustness are addressed in model updates and training?

Answering the above question provides a mechanism to enable certification for introduction into service and standard setting when adopted for use.  This ensures that risks with implementation of the AI are accurately recorded to ensure the capability remains fit for purpose.

### A.6 Reliable and A.7 Predictable

A.6.i Please identify the metrics you have used to assess the reliability and predictability of the AI functionality (if different to metrics to assess model performance)

A.6.ii Does the AI capability require more than one type of AI functionality which impact on the overall reliability and/or predictability of the AI capability?

A.6.iii What is the reliability of the operation of the AI capability in its designed or intended operation?

A.6.iv What is the predictability of the AI capability in its production of designed or intended operation?

A.6.v What happens when the AI does not operate reliably?

A.6.vi Is there anything that the AI capability should not be used for (that would otherwise fall within its normal anticipated use)?

A.6.vii What are uses of the AI capability (or its product) that should not occur? If there is anything that the AI capability (or its product) should not be used for – what are they?

A.6.viii What happens when the product of the AI is not sufficiently predictable?

A.6.ix What control measures can be applied to AI capability, or to the elements of the AI functionality, that can remove or reduce the consequences of unreliability or unpredictability?

A.6.x What control measures are being applied to protect against unintentional failures and attacks?

Answering the above questions addresses risk associated with ensuring that the AI functions in a manner that meets performance standards in an unfailing and consistent way; and that when it does operate, it does so in the way that it was intended.

### A.11 Safe and A.12 Secure

A.11.i What safety measure have been included with the AI?

A.11.ii Are there any safety hazards from the operation of the AI capability for the human users?

### A.11.iii What security measure have been included with the AI?

### A.11.iv Please describe how secure the AI capability is, including consideration of:

- Virtual (i.e. if the AI capability secure from hacking or other forms of virtual infiltration or manipulation) and physical (i.e. physical infiltration and access) security threats;
- Any identified vulnerabilities or risks, and how these were or were not addressed
- Independent or self-assessed certification of cyber security, including:
  - Type and nature of testing (i.e. pen testing, red teaming, simulations etc.);
- What virtual and physical security measures are:
  - in place to protect the AI capability across its life-cycle;
  - Are required to be out in place by Defence across its life cycle;

### A.11.v What metrics have been defined for system performance and accuracy as they relate to safety and security?

Answering the above questions will identify if there are any risks to safety of ADF personnel and civilians who are going to rely upon the operation of the AI and the failure of that AI functionality will risk human life, that must be mitigated in the Risk Register.

## B. Design Inputs

ℹ This component describes the processes of design of the AI capability.

It addresses the systems engineering of the AI capability, and the testing, evaluation, validation and verification processes applied to the AI capability as it is designed.

This component includes an assessment of the input data for the design and training of the AI capability (as compared to AI Use Input which deals with the input data when in use).

It also includes the broader design of the AI capability, which consists of the design approach to its software, hardware and associated platform.

### B. RAID Checklist Questions – Prompting Design Inputs Component Completion

ℹ This section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| A.1.1 | …is designed to enable combat functionality of a weapon[5] or means[6] of warfare |
| A.1.2 | …is designed to undertake safety critical functions |
| A.3.1 | …can learn or modify its own goals triggers an ongoing requirement for TEVV |
| 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.7.1 | …relies on a mathematical model that is imprecise (requiring control measures to account for the imprecision) |
| 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 |
| G.2.1 | … is intended to enable a method of warfare |

[5] 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.’

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

| Serial | Threshold Question |
|---|---|
| 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, developed or operated without expert assistance |

### B.A. Preliminary information

B.A.i Development Data Source: Please describe how you used data to develop your AI capability, including identifying:

- The type of data used.
- Can you identify which data was used for training, testing, and validation?
- The provenance of the data used including:
  - Where the data came from:
    - its source (single provider) or multiple sources.
    - If it was open or closed source
  - How the data was collected or created (including whether it was from humans (experts or lay persons), sensors, artificially created (i.e. synthetic) or some combination thereof).
  - If the data was collected consensually and whether the labelling was undertaken without exploitation.
  - Your data quality assurance plans and processes

Answering the above questions will identify if there are any risks in relation to the data source that trigger legal rights or require special handling of the data once inputted into the AI capability or system.

### B. Elements Analysis

### B.6 Reliable and B.7 Predictable

B.6.i Describe how you have ensured the development dataset fit for purpose, including identifying:

- The sample size and quality of the data pool used.
  - On data size:
    - is the sample size of the development data adequate for the AI capability?
  - On the quality of the development data and/or dataset:
    - Please identify how you have validated the data and dataset as being:
      - Representative.
      - Unbiased.
      - Complete.
      - Absent unnecessary noise.

- Consistent.
- Accurate.
- Interrogable (and thus separable or able to be disaggregated)
- Accessible.
- Secure.
- Appropriately marked or labelled (minimum metadata, human versus machine tagging, labelling, sorting, or analysis of data)

B.6.ii Identify what data you have identified as being of insufficient quality, and how you are managing the data deficiencies including measures taken (such as general data sanitisation, general and specific data exclusion, addition of new data, artificial or synthetic data substitution.

Answering the above questions address risk associated with data governance, specifically the risks associated with ensuring that the AI functions in a manner that meets performance standards in an unfailing and consistent way; and that when it does operate, it does so in the way that it was intended.

### B.9 Controllable

B.9.i Please describe the legal and policy status of the data, including identifying:

- Any rights attached to the data such as:
  - Proprietary rights i.e. relating to ownership and associated restrictions or limitations;
  - Personal rights i.e. relating to personal information (information that could identify an individual). If the data contains personal information, then:
    - Does the data contain personal information which creates privacy concerns; and if so, what were these privacy concerns and how were they managed (such as by anonymising or pseudonymising the data)?
    - Does the data impinge on any other rights of the person(s) from who the personal information is sourced?
- Open source or public data with no attached rights and thus restrictions or limitations on use.

B.9.ii Describe any other regulatory requirements applicable to the data, including highlighting:

- What the regulatory requirements are and how they are managed (i.e. what is required to ensure the development data adheres to regulatory requirements and how was this requirement achievement)?
- What are the ethical concerns with the collection and use of the development data in the AI capability, and how are these managed?

B.9.iii Does the AI capability require ongoing access to development data (OECD – are the dynamic, static, dynamic updated from time to time or real-time?

- What happens if the AI capability can no longer access the development data?

B.9.iv Describe the development data format, including identifying:

- How the data (and metadata) is formatted;
- Where data (and metadata) is not standardised, how you manage this; and
- The structure of the data and metadata (i.e. structured, semi-structured, complex structured or unstructured)?

Answering the above questions address risk associated with development data governance and data formatting, namely whether the AI will operate in a way that is unintended because of the input data; or that it handles or manages the input data in a way that is inconsistent with legal rights of the original data owners.

### B.8 Compliant and B.11 Safe and B.7 Predictable

B.8.i Describe how you have you ensured the data was sanitised, including identifying:

- When, where, how and by whom was it sanitised?
- If the development data could not be sanitised how was or is this to be managed?

B.8.ii Identify how you have managed representation, bias, or any other deviance in the development data, including identifying:

- When, where, how and by whom was the data checked for representation, bias, or any other deviance?
- How did you test for and evaluate the representativeness of the data?
    - If the data is representative:
        - What is required to maintain representativeness?
        - What will make it or when will it become unrepresentative?
    - If the data is unrepresentative – what was the lack or representation and how was/is this managed?
- How did you test for and evaluate bias (or lack thereof) in the data?
    - How do you address intentional bias?
    - If the data has no unintentional bias:
        - What is required to maintain lack of unintentional bias?
        - What will make it, or when will it become unintentionally biased?
    - If the data was/is unintentionally biased – what was/is this bias and how was/is this bias managed?
- What other deviances in the data relevant, and how did you test for evaluate these deviances?
    - If the data has no other relevant deviances:
        - What is required to maintain lack of other deviances?
        - What will make it, or when will it display deviances?

  - If the data displays some other form of deviance – what is this deviance and how was this deviance managed?

Answering the above questions address risk associated with development data hygiene, specifically there are no deviances in the input data that will cause the AI to operate in a manner that is not safe or operates in a manner that is biased, or as required under the law.

### B.12 Secure

B.12.i Identify how the data is accessed and secured.

B.12.ii Identify the data custodian for the development data and what controls they have on the data:

- How is the data secured?
- How do you access the data?
- How is the data protected from unauthorised access or data poisoning
- Is the data accompanied by metadata; and what quality is the metadata?

B.12.iii Where data is classified, please identify the type and nature of the security classified data (including the level of classification), how that security classified data is managed and by whom.

B.12.iv Identify what data is required for training and certification, including highlighting:

- the source,
- format,
- quality and appropriateness,
- access and security,
- hygiene, and
- governance.

Answering the above questions address risk associated with development data and security that would render the system capable of being inappropriately corrupted. For example, it could be infiltrated while in use; or that the inputted data is capable of being accessed in a manner inconsistent with the legal and ethical handling requirements associated with that input data.

## C. HMI

ℹThe Human Machine Interaction (HMI) is the component that describes how the human operators engage with the AI capability.

It describes how they can control the capability and the interface system of the AI capability.

This component addresses the physical interaction, the software controls that are available to the human operators, and the physiological and psychological interface between human and machine.

### C. RAID Checklist Questions – Prompting HMI Component Completion

ℹThis section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| A.1.1 | …is designed to enable combat functionality of a weapon[7] or means[8] of warfare |
| A.1.2 | …is designed to undertake safety critical functions |
| 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 |
| 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.7.1 | …relies on a mathematical model that is imprecise (requiring control measures to account for the imprecision) |
| 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.1.1 | …requires a human operator to input instructions or data for it to operate |

[7] 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.’

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

| Serial | Threshold Question |
|---|---|
| D.2.1 | … is susceptible to uncontrolled input – including using data from AI capability sensors |
| G.2.1 | … is intended to enable a method of warfare |
| 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, developed or operated without expert assistance |

### C.A. Preliminary information

C.A.i Explain the required and/or expected human-AI interaction for the AI capability (or for any AI functionality, including identifying:

- How the AI capability operates with human operators, including:
  - How it was designed to operate;
  - How it actually operates;
  - Whether this interaction is subject to change;
  - How HMI effects the designed or expected operation of the AI capability
- If the operator is able to input data?
- If the operator is able to change metrics or other parameters of the AI model?
- What control measures could be applied to the AI functionality by human operators across the spectrum of human involvement?
- Is the AI capability intended (or able) to operate without interaction with a human, or at least interaction prior to implementing decisions that have effects.
- Whether the AI has been designed and developed to protect against undesirable human-AI interactions, and if so, how?
- How is HMI physically achieved:
  - What software and hardware are used to achieve HMI?
  - Is there a temporal or geographical dislocation between its interface and effect caused by the AI,
  - How have you designed the HMI to be usable (including efforts to ensure responsibility, accountability, understandability, explainability, reviewability)

C.A.ii Describe the system ensures continuity of control, including:

- How the AI communicatees with the human interface;
- What shut-off system operates in the event of communication drop-out?

Answering the above questions address risk related to the operating limitations of the AI with respect to the ability of the human operator to interact with it.

## C. Elements Analysis

### C.1 Responsible and C.9 Controllable

#### C.1.i Can you identify the human who is responsible for every action of the AI capability?

- if yes, who are they by position or conduct? In answering this please consider:
    - The responsibilities of humans involved in, inter alia, the design, development, testing, production, certification, training, or use of the AI capability;
    - The required training, education and knowledge of these humans to undertake their responsibilities;
    - The awareness of these humans of their responsibilities and the risks of not fulfilling their responsibilities;
    - Whether such risks have been specifically allocated
- If no, why not and what measures are required to ensure a human or humans are responsible

#### C.1.ii What control measures:

- have been applied to the AI capability or AI functionality against undesirable human-AI interactions?
- could be applied to the AI functionality by human operators across the spectrum of human involvement to ensure responsibility is maintained?
- could be applied to the AI functionality by humans to ensure the AI operates as intended?

Answering the above questions will address risk associated with identifying those measures of human interaction that permit the human to remain responsible through maintenance of control of the AI capability.

### C.1 Responsible and C.2 Accountable

#### C.1.i Upon activation of the AI capability:

- How does a responsible humans interact with the AI capability?
- What is your process for ensuring a responsible human can maintain control of the AI capability, including:
    - is the AI capability designed to implement decisions or create effects without direct human supervision or intervention;
    - do you have processes for controlling (authorising, managing, removing etc.) access to the AI capability;
    - do you have processes for managing unexpected or unintended behaviour
- Is it clear who is responsible for the operation of the AI capability, and how will this responsibility be maintained?

#### C.1.ii Is accountability for the operation of the AI capability the same as responsibility for the AI capability?

C.1.iii Should responsibility of the operator or relevant decision-makers responsible for use of the AI capability (or the system it is integrated) always be aligned with accountability?

Answering the above questions address risk that the interaction permits human accountability and responsibility to be maintained.

### C.3 Understandable

C.3.i Explain the required level of training or experience that a user must have to be able to operate the AI capability.

C.3.ii Can an operator or decision-maker understand the intended operation of the AI capability?

C.3.iii Could a human understand the intended operation of the AI capability?

C.3.iv Will a user be aware when they are interacting with the AI capability; are you designing features into the HMI that ensures a user is aware they are interacting with an AI capability?

C.3.v Describe what is required for a user to operate the AI, including:

- The extent and level of understandability required for the operation of the AI capability, by context of designed or anticipated use, prior to an operator deciding which AI and machine learning techniques to employ in the system.
- What information, processes, education, or training is required to ensure users understand the AI capability outputs?
- What information, processes, education, or training has been prepared to ensure users understand the AI capability outputs?

C.3.vi How will a user know if the AI capability is:

- Operating in accordance with design parameters;
- The extent to which it is operating in accordance with design parameters (such as levels of confidence);
- How the AI capability communicates its level of performance; or
- Capable of unlawful use.

C.3.vii How did you consider human factors in the design or operation of the AI capability?

Answering the above questions addresses the risks posed by a human operator deploying an AI capability in a situation that is beyond its functionality or outside of the risk thresholds set for its use case or it is used in a manner different from the normal anticipated use.

### C.4 Explainable

C.4.i Will the operation of the AI capability, for any given action, be provided to the operator in way that they will understand and be able to explain?

C.4.ii Describe what you have designed or built into the AI to ensure the Ai capability actions can be explained to an operator, including:

- The nature and type of In-built mechanisms to ensure explainability
- The level of competence of the operator required to permit explainability
- The extent of explainability capable.

Answering the above questions address risks of the human deploying the AI in a way that is consistent with their understanding of what the anticipated AI effect of its normal anticipated use will be.

### C.5 Reviewable

C.5.i To what extent/how are you - documenting information about the model to reproduce intended outcomes and prevent undesirable outcomes over the system's lifecycle?

C.5.ii Describe how any decision made or implemented by the AI capability is capable of review, including:

- In-built review mechanisms into the AI capability?
- What aspects of the review mechanisms in the AI capability operation could a human review?
- Could an operator or decision-maker review any given operation of the AI capability; and if so detail where they can review the AI capability (i.e. prior to implementing action, after action but during operation, after completion of operation through data extraction)?

Answering the above questions address the risks related to the operation of the AI that require immediate review, or review during interaction with the human to prevent the operation of the AI continuing in deleterious manner, or outside its normal anticipated use.

### C.8 Compliant

C.8.ii Does the AI operate on the presumption of lawful use, or does it have control measures that can be used to restrict what the operator is permitted to do with the AI?

Answering the above question addresses the risk relating to the interaction between the human and the AI may result in a failure of the AI to operate consistent with legal obligations for its particular use case.

## D. AI Use Inputs

ℹThe AI Use Inputs is the data that is fed into the AI capability for it to undertake its computation process when in operation.

This includes the algorithm itself (as the instructions to the AI) when in use, as well as source data that is analysed by the AI capability to produce its analysis.

### D. RAID Checklist Questions – Prompting AI Use Inputs Component Completion

ℹThis section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| 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.3.1 | …can learn or modify its own goals triggers an ongoing requirement for TEVV |
| 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 |
| 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 |
| G.2.1 | … is intended to enable a method of warfare |
| 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, developed or operated without expert assistance |

[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.

### D.A. Preliminary information

D.A.i Describe how the operator or user of the AI capability is able to:

- Input data;
- Input instructions;
- Alter parameters/standards/metrics;

of the AI functionality or AI model via input?

D.A.ii Describe how the AI capability will use, or is capable of using data to exercise its AI functionality, including identifying:

- The type of data used.
- The provenance of the data used including:
  - Where the data came from (i.e. is it system sensors, external allied sensors, or external data);
  - its source (single provider) or multiple sources;
  - If it was open or closed source.
- The system sensors (or input) nature and type of collection/creation (including whether it was from humans (experts or lay persons), sensors, artificially created or some combination thereof).
- The classification structure that will apply to the data, if any.
- If it is possible to vary the classification structure – how do you do this?
- If the data was collected consensually and whether the labelling was undertaken without exploitation.

Answering the above questions will identify if there are any risks in relation to the environmental data source that trigger legal rights or require special handling of the data once inputted into the AI capability or system.

### D. Elements Analysis

#### D.6 Reliable and D.7 Predictable

D.6.i Describe how you have ensured input data is fit for purpose, including identifying:

- The data is correctly labelled or marked (minimum metadata, human versus machine tagging, labelling, sorting, or analysis of data),
- The data is within acceptable parameters,
- How you address data that is not representative of the context in which the AI is expected to be used,
- The data aligns with the data used to train the AI.

D.6.ii Describe how the data is assured for quality when input from its use environment, including not being subject to:

- bias,
- incompleteness,

- unnecessary noise,
- inconsistency,
- inaccuracy,
- an inability to be interrogated (and thus not separable or able to be disaggregated),
- being inaccessible, and
- insecurity.

D.6.iii Identify how you address:

- un-sanitised data;
- uncontrolled data;
- deficient data; or
- inputs that are incorrect (this includes in-built software locks).

Answering the above questions address risk associated with data governance, specifically the risks associated with ensuring that the AI functions in a manner that meets performance standards in an unfailing and consistent way; and that when it does operate, it does so in the way that it was intended.

### D.8 Compliant and D.11 Safe and D.7 Predictable

D.8.i Describe how the relevant environmental inputs; including

- How you have ensured data hygiene, or included measures for identifying data deviance?
- How have you addressed using data from these sources with the development data?
- How variation from the development data is addressed?

Answering the above questions address risk associated with environmental input data hygiene, specifically there are no deviances in the input data that will cause the AI to operate in a manner that is not safe or operates in a manner that is biased, or as required under the law.

### D.9 Controllable

D.9.i describe the legal and policy status of the data, including identifying:

- Any rights attached to the data such as:
  - Proprietary rights i.e. relating to ownership and associated restrictions or limitations;
  - Personal rights i.e. relating to personal information (information that could identify an individual). If the data contains personal information, then:
    - Does the data contain personal information which creates privacy concerns; and if so, what were these privacy concerns and how were they managed (such as by anonymising or pseudonymising the data)?
    - Does the data impinge on any other rights of the person(s) from who the personal information is sourced?

    - Open source or public data with no attached rights and thus restrictions or limitations on use

D.9.ii Describe any other regulatory requirements applicable to the data, including highlighting:

- What the regulatory requirements are and how they are managed (i.e. what is required to ensure the environmental data adheres to regulatory requirements and how was this requirement achievement)?
- What are the ethical concerns with the collection and use of the environmental data in the AI capability, and how are these managed?

D.9.iii Does the AI capability require ongoing access to development data (OECD – are the dynamic, static, dynamic updated from time to time or real-time?

- What happens if the AI capability can no longer access the environmental data?

D.9.iv Describe the environmental input data format, including identifying:

- How the data (and metadata) is formatted;
- Where data (and metadata) is not standardised, how you manage this; and
- The structure of the data and metadata (i.e. structured, semi-structured, complex structured or unstructured)?

Answering the above questions address risk associated with environmental input data governance and data formatting, namely whether the AI will operate in a way that is unintended because of the input data; or that it handles or manages the input data in a way that is inconsistent with legal rights of the original data owners.

### D.12 Secure

D.12.i Identify the data custodian for the environmental data and what controls they have on the data:

- How is the data secured?
- How do you access the data?
- Is the data accompanied by metadata; and what quality is the metadata?

D.12.ii Where data is classified, please identify the type and nature of the security classified data (including the level of classification), how that security classified data is managed and by whom.

D.12.iii Identify what data is required for training and certification, including highlighting:

- the source,
- format,
- quality and appropriateness,
- access and security,

- hygiene, and
- governance.

Answering the above questions address risk associated with input data and security that would render the system capable of being inappropriately corrupted. For example, it could be infiltrated while in use; or that the inputted data is capable of being accessed in a manner inconsistent with the legal and ethical handling requirements associated with that input data.

## E. AI Use Outputs

ℹThe AI Use Outputs are those outcomes that result from the AI computational process.

This is the result of the data analysis that is undertaken by the AI capability.

Whether a recommendation for a decision maker, or a summary of existing data, this output is then used within the system to influence an action that will affect the object of the AI action.

## E. RAID Checklist Questions – Prompting AI Use Outputs Component Completion

ℹThis section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| 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.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 |
| 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 |
| 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.1.1 | …interacts with humans as the object of the AI action as the object of the AI action |
| F.2.1 | …directly affects the rights or obligations of persons or things not operating the system |

[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.

| | |
|---|---|
| G.2.1 | … is intended to enable a method of warfare |
| 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, developed or operated without expert assistance |

### E.A. Preliminary information

E.A.i describe what the outputs of the AI capability are, including:

- Detailing all of the AI capability outputs, including their purpose;
- Identifying the AI model used to create the output, including
    - If the model creating the output is deterministic and probabilistic, or a combination thereof
- Identifying if any capability outputs are applied without being checked by a human or independent source, and if so how;
- Identifying what the capability output is intended or anticipated to be used for, including
    - Operation of, or integration with, a larger system;
    - To be sent to external people or things
- Identifying if any of the capability outputs are regulated by the law, or will impact on the rights or interests of persons (or things), and if so how;
- Identifying if any of the capability outputs are designed to, or potentially will affect safety-critical or mission-critical functions, and if so how.

E.A.ii Specify the intended or potential context of use of the AI capability resulting in the AI outputs including:

- Any specific military contexts the AI capability is intended to be used, in particular;
    - What, if any, is the operational environment this AI capability is designed for?
    - What is the normal anticipated use of the AI capability?
    - Which military context will be triggered by the use?
    - Will the AI capability, and/or the AI output, require adaptation for the given military context of normal anticipated use?
- Is the training data representative of the context in which the AT capability is expected to be used?

E.A.iii Describe the intended or potential users of the AI output, including:

- Who are the intended or anticipated end user(s) (people or things)?
- What people (or things) will interact with the AI capability (or the system it is integrated into) when operated in its normal anticipated use?
- What people or things will be impacted by the system in its normal anticipated use.

Answering the above questions address risks relating to the product being developed by the AI are; that it, the output analysis resulting from the AI functionality complying with its anticipated normal use case.

## E. Elements Analysis

### E.1 Responsibility and E.2 Accountability

E.1.i Is the AI output designed, intended, or anticipated to impact or harm a person or thing, and if so please:

- describe the relevant harms or impacts, including;
  - where the AI output enables a weapon, means or method warfare, is critical to safety of life functions, or effects the privacy of a person or persons - complete the appropriate annex(es) (Article 36, WHS, and privacy) to this LEAPP.
  - evaluated the relevant harms or impacts against the purpose of the AI capability, or system it is integrated into.
  - How can an impacted or harmed party challenge or seek review of an AI output.

E.1.ii Have you considered if unintended people (or things) may be inadvertently or accidentally impacted by the AI output, and if so:

- what risks have been identified;
- what are the impacts or harms attached to the identified risks;
- how have such risks been managed (including designating risk owners, software and hardware locks, and use restrictions/limitations?

Answering the above questions address risk associated with the output having unintended and/or significant or potentially catastrophic consequences.

### E.4 Explainable

E.4.i Is the AI output explainable by a person (or thing)?

Answering the above questions address risk associated with the ability to interrogate the output of the AI should it not perform as anticipated.

### E.5 Reviewable

E.5.i Is the AI output recorded, and if so please describe:

- where it is stored and in what format;
- how and where it can be accessed;
- who can access it;
- how long it is stored for?

E.5.ii Is the AI output reviewable, and if so:

- How can they be reviewed, in particular including:
  - in what format is the AI output reviewable,
  - by what means or output device can the AI output be reviewed;
  - by what or whom can the AI output be reviewed;

  - are there any standards or methods for the review of the AI output; and
  - at what time can the AI output be reviewed?
- Is there anything relevant to the creating the AI output that is not reviewable, and if so please describe:
- what can’t be reviewed;
- why it cannot be reviewed; and
- what arrangements have been put in place to manage the inability to undertake such a review?

### E.5.iii Are the relevant inputs, user interactions, use context, AI model, or any other relevant considerations to the creation of the AI output reviewable, and if so:

- Are specific inputs etc that are causative of the AI output able to be identified, and if so, how are they identified?
- How can they be reviewed, in particular including:
  - in what format they are reviewable,
  - by what means or output device can they be reviewed;
  - by what or whom can they be reviewed;
  - are there any standards or methods for the review; and
  - at what time can they be reviewed?
- Is there anything relevant to the creating the AI output that is not reviewable, and if so please describe:
  - what can’t be reviewed;
  - why it cannot be reviewed; and
  - what arrangements have been put in place to manage the inability to undertake such a review?

Answering the above questions address risk associated with the ability to interrogate the output of the AI should it not perform as anticipated; or allow interrogation of the AI output to align with legal investigative obligations.

## E.6 Reliable and E.7 Predictable

### E.6.i What control measures can be applied to AI capability (or individual AI elements) to reduce, remove or ameliorate the AI output resulting from the unreliable operation or unpredicted output of the operation of the AI capability?

Answering the above questions address risk that the AI can cause unintended and undesirable effects.

## F. Object of AI Action

ℹThe Object of AI Action is the thing or process that is affected by the AI. It is the object that the AI is designed to influence as a result of its computational process.

It is not the operator of the AI, but rather the person or thing that the decision or algorithmic process is designed to analyse or assess. It may be a person, for example, in the case of an AI capability that is designed to undertake surveillance., it could be property, or it could be data.

### F. RAID Checklist Questions – Prompting Object of AI Action Component Completion

ℹThis section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| A.1.1 | …is designed to enable combat functionality of a weapon[13] or means[14] 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 | …undertaking novel decisions only made possible by complex algorithmic processing |
| A.2.3 | … making 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 |
| 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 |
| 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 |

[13] 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.’

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

| | |
|---|---|
| F.1.1 | …interacts with humans as the object of the AI action 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.2.1 | … is intended to enable a method of warfare |
| 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, developed or operated without expert assistance |

## F.A. Preliminary information

### F.A.i Describe who or what is the intended object of the AI capability, including identifying:

- If the intended object is aware that they are the object of an AI capability;
- Where the intended object is a human – have they consented to the use of the AI capability;
    - If so, how is consent arranged?
    - If not, how is the lack of consent managed?
- Can the object of the AI choose not to be an object of the AI capability and if so, how?
- Can the object of the AI challenge or alter the output of the AI and if so, how?
- Could the operation of the AI capability advantage or disadvantage the interests or rights of a person (or thing):
    - Please list and explain those interests of a person (or thing) that are advantaged;
    - Please list those interests of a person (or thing) that are disadvantaged:
        - Where fundamental human rights are impacted (e.g. human dignity, privacy, freedom of expression, non-discrimination, fair trial, remedy, safety) please identify how these are considered in your design or otherwise managed.
        - Where areas of well-being are impacted (e.g. military justice, job quality, mental or physical health, social interactions, education) please identify how these considered in your design or otherwise managed.
        - Where rights or interests of a person or thing – other than human rights or well-being – are impacted (e.g. financial cost, loss of time, inconvenience, legitimate harms) please identify how these are considered in your design or otherwise managed.

Answering the above questions address the risk of unintended or undesired consequences on humans or things.

## F. Elements Analysis

### F.1 Responsibility

F.1.i If the intended object of the AI is intended to be harmed by the AI (e.g. the AI capability is a, or part of a, weapon, means or method of warfare) how are the rights of the object considered?

Answering the above question addresses risk associated with any legal obligations, such as for example, the AI operating in compliance with specific rules of LOAC that necessitate protection of persons and objects of particular classes.

### F.8 Compliant

F.8.i Are the impacts of the decision of the AI upon the individual person or object reversible?

F.8.ii How long lasting are the impacts of the decision of the AI upon the person or object?

Answering the above questions identifies the magnitude of the risk associated with the AI action impacting upon the AI object.

## G. Use Case Environment

ℹThe use case environment is the location in which the AI operates.

The use case environment sits within the system of controls as a whole.

There may be multiple use case environments for on AI capability, however, individual assessments must be made in respect of each use case environment.

It includes an assessment of the military domain in which the capability will be fielded, given different operating environments result in different legal and ethical considerations in the use of an AI capability.

The purpose of this component is to identify the conceptual, physical or design and development limitations on the system. It is used to identify what the limits on the use of the system should be.

Answers to the following questions will assist in identifying the context in which the AI capability is intended to be fused by Defence. This frames subsequent considerations about risk, noting the context for use will have significant impact upon the risk of use of an AI capability. Equally, its context for use will also aid in determining relevant legal and ethical frameworks and identification of associated risks.

The answers to the following questions may cross multiple contexts, particularly for functions that engage the Command Warfighting function. All relevant contexts should be selected based upon the AI capability's normal anticipated use (which also requires articulation in this step).

The purpose of this question is to identify how the AI capability is intended to be used in order to identify the operational contexts (or ADF warfighting functions) in which the AI capability will be used.

### G. RAID Checklist Questions – Prompting Use Case Environment Component Completion

ℹThis section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| A.1.1 | …is designed to enable combat functionality of a weapon[15] or means[16] of warfare |
| A.1.2 | …is designed to undertake safety critical functions |
| 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 |

[15] 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.'

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

| | |
|---|---|
| | 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 |
| G.2.1 | … is intended to enable a method of warfare |
| 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, developed or operated without expert assistance |

### G.A. Preliminary information

#### G.A.i Describe the intended use case of the AI, including identifying:

- The intended operating environment, including:
  - The operating environment the AI was designed for (land, air, maritime, space, cyber, EMS, or a combination thereof)
  - Are people in the use environment particularly vulnerable?
  - The operating environment the AI capability is being proposed for (if different to designed operating environment
  - If the operating environment can be simulated, and if so how and how accurate was this simulation
  - The military context for proposed use (see below)
- How closely the model can match the operating environment, and what is required to improve the model
- The intended tasks of the AI capability
- The intended operating environment sources of input, and what happens if these data sources are lost
- What happens if the AI capability is used in a different operating environment or for a different task
- What operating environments, inputs, or tasks should the AI capability not be used for?

#### G.A.ii Describe the risks of the normal anticipated use of the AI capability:

- What risks do you estimate to be relevant to the Use Case?
  - For instance, is the AI capability or its intended use subject to (or likely to be subject to) intense public scrutiny (e.g. because of privacy concerns) and/or frequent litigation?
  - Are the consequences of the intended use high or significant?
- How did you determine these risks and corresponding mitigation measures?
- Have you already implemented any of the mitigation measures?
- What evaluation metrics did you use to identify the risks
- How and with what frequency will you evaluate and monitor the success of the risk mitigation measures you intend to implement?

- What other points about the impact and risk of the Use Case are not addressed in this memo but should be included in this assessment?

G.A.iii Describe the effect of the use case on users, including:

- Will the intended use of the AI capability have major impacts on staff, either in terms of their numbers or their roles?

Answering the above question will identify risk associated with the context of the AI use. It will also identify whether the LEAPP is required to be undertaken to address risks for multiple use cases – that is, the normal anticipated use is broad and captures multiple operating environments or combat functions.

## G. Elements Analysis

### G.8 Compliant and G.9 Controllable

G.8.i Describe the normal anticipated military context for use, including:

- What military functions the AI capability is intended to support – use Annex B to identify which Warfighting functions are triggered by the use case for the AI functionality.

| Combat or Warfighting Function | | | | |
|---|---|---|---|---|
| Cmd | FA | FP | FS | SU |
| □ | □ | □ | □ | □ |

| Enterprise Level and Rear Echelon Function | | |
|---|---|---|
| PR | EL | BP |
| □ | □ | □ |

- How will the AI capability support those military functions?
- Any requirements for the operation of the AI capability that would need to be considered in the exercise of that military function?
- Whether the AI capability will replicate work previously undertake by humans?
- What design limitations have been applied to fit the military context?
- Has the normal anticipated use deviated from the design specifications?
- What policy has been developed regarding the application of the AI capability?

## H. System of Control

ℹThe system of control describes the broader Defence system within which the AI will operate.

It describes the systems, processes and authorities that are in place across the lifecycle of use and in-service deployment and disposal of an AI capability.

It is reflective of Australia’s ‘system of controls’ which provide a layered approach to the control mechanisms and processes that control how an AI capability will be used.

It also includes analysis of how the AI capability integrates within the system. System integration considers internal and external connectivity, data pathways, and the physical integration of the AI componentry. This analysis is considered closely with the HMI component when it analyses the human-machine interface.

### H. RAID Checklist Questions – Prompting System of Control AI Component Completion

ℹThis section of the LEAPP is copied directly from the RAID Checklist:

| Serial | Threshold Question |
|---|---|
| A.1.1 | …is designed to enable combat functionality of a weapon[17] or means[18] of warfare |
| A.1.2 | …is designed to undertake safety critical functions |
| 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.7.1 | …relies on a mathematical model that is imprecise (requiring control measures to account for the imprecision) |
| 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 |
| F.1.1 | …interacts with humans as the object of the AI action as the object of the AI action |

[17] 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.’

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

| | |
|---|---|
| F.2.1 | …directly affects the rights or obligations of persons or things not operating the system |
| G.2.1 | … is intended to enable a method of warfare |
| 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, developed or operated without expert assistance |

## H. Elements Analysis

### H.1 Responsible

H.1.i Is it possible to clearly assign responsibility (and where separate accountability) for the use of the AI capability within the system of control? If so, how would you do this?

H.1.ii Are additional subject matters experts required to support the multi-disciplinary certification for use of an AI capability, guide AI governance and assurance development?

If so, what kind of expert and who are those experts (i.e. they may include people with expertise in:

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

Answering the above questions addresses the risks of whether or not there has been an appropriate human operator identified to employ the AI capability; or appropriate human involvement in the design of the AI capability has not been engaged.

### H.4 Understandable

H.4.i Describe what aspects of the design, development, production, or operation of the AI capability you have documented, including:

- TEVV;
- training, re-training, and certification;
- standards or metrics;
- independent evaluation or review.

Answering the above question addresses the risk that there has been insufficient documentation of the AI capability to enable it to be properly understood to support certification and its employment.

### H.7 Predictable

H.7.i Please describe the TEVV that was applied to the AI capability: including:

- What TEVV is, or is planned, to be provided for the AI capability and functionality, including:
    - the type and start date of TEVV undertaken;
    - the results of that TEVV;
    - whether TEVV has been completed and any ongoing requirements for TEVV should the AI capability be acquired by Defence;
    - whether the TEVV is completed internally by the designer and developer, or by an independent agency; and
        - If the TEVV was completed internally was there independent review of the results of your TEVV;
    - whether the same AI system is used to conduct the TEVV, or whether a separate and independent AI capability is used.

- The applied TEVV approach, including:
    - Was the TEVV applied throughout the AI capability development and training;
    - Validation methodology (i.e. of the AI models, inputs, use cases, system integration, testing outputs etc.) applied to ensure fitness for purpose:
    - Validation standards or metrics that were applied;
    - Any other AI capability and system performance evaluation methodologies, data assurance and metrics.

- TEVV documentation, including extent of TEVV for:
    - Robustness of AI capability on normal anticipated use against variability (such as failures, malicious attacks or actions, usage context changes, user error, unexpected object actions etc.);
    - Understandability, explainability and reviewability;
    - Identifying predictability, predictability of unintended outputs or operation, and nature and extent of unpredictable and unintended outputs or operations;
    - Control measures to stop, change or limit unpredictable outputs or operations;
    - AI rules or training processes applied.

### H.7.ii Describe the TEVV is, or is planned, to be provided for the AI capability, including:

validation methodology to be applied (i.e. of the AI models, inputs, use cases, system integration, testing outputs etc.) to ensure fitness for purpose

- Standards or metrics to be applied;
- TEVV documentation, including extent of TEVV for:
    - Robustness of AI capability on normal anticipated use against variability (such as failures, malicious attacks or actions, usage context changes, user error, unexpected object actions etc.);
    - Understandability, explainability and reviewability;
    - Identifying predictability, predictability of unintended outputs or operation, and nature and extent of unpredictable and unintended outputs or operations.
- Control measures to stop, change or limit unpredictable outputs or operations.
- AI rules or training processes to be applied;
- Independent review of the results of the TEVV;
- Any additional subject matters experts required to guide AI governance and assurance development (including kind and type of experts).

### H.7.iii Describe the future or ongoing TEVV requirements for the AI capability, including:

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:

- Validation methodologies that could or should be applied (i.e. of the AI models, inputs, use cases, system integration, testing outputs etc.) to ensure fitness for purpose;
- Standards or metrics that could or should be applied;
- TEVV documentation, including extent of TEVV for:
    - Robustness of AI capability on normal anticipated use against variability (such as failures, malicious attacks or actions, usage context changes, user error, unexpected object actions etc.);
    - Understandability, explainability and reviewability;
    - Identifying predictability, predictability of unintended outputs or operation, and nature and extent of unpredictable and unintended outputs or operations;
- Control measures to stop, change or limit unpredictable outputs or operations;
- AI rules or training processes that could or should be applied;
- Independent review of the results of the TEVV;
- Any additional subject matters experts required to guide AI governance and assurance development (including kind and type of experts).

Answering the above questions address the risk of the AI being incapable of operating as intended through a lack of appropriate TEVV.

### H.8 Compliant

H.8.i If the use of the AI capability (or system in which the AI is integrated) fails to comply with relevant domestic or international legal requirements:

- How does it fail to comply and how do you manage this?
- What control measures are attached to the range of actions that the AI capability can undertake?
- What control measures to ensure compliant use have been developed?
- Do you have control measures to maintain control where the AI functionality fails (is not reliable) or acts unpredictably?
- Where the AI functionality has a learning capacity, will it require the control measures to be updated to maintain control?
- Who (or what) is the AI capability intended to be controlled by?
- What control measures are required to be implemented by Defence to ensure Defence maintains control of the AI capability (or the system it is integrated into)?

Answering the above question addresses the risk that the AI does not meet legal standards when operating.

### H.9 Controllable

H.9.i Describe the design considerations relating to the AI capability being controlled under the system of control applied by the Department of Defence, in particular, including:

- Is the AI capability designed or intended to operate as standalone software, or is it designed or intended to operate within, or as part of a larger system and have the governance and control requirements of both the AI capability and the system it is integrated into been considered?
- How have the governance and control requirements been considered?

H.9.ii Describe the legislative or other regulatory regimes that apply to the AI capability (or the system it is integrated into) including the AI or system output.

H.9.iii Describe any other relevant frameworks that apply, or have been applied, including:

- What frameworks were applied to design and develop the AI capability, and the extent to which they must continue to be applied;
- What frameworks are required to support the adoption of the AI capability, or systems the AI capability is integrated with, and how they are to be applied;
- What frameworks will be required for the future operation of the AI capability, and how they are to be applied, in particular:
    - Are additional, separate governance structures required to support the design, development or use of the AI capability?

H.9.iv Describe what Department of Defence governance measures have been considered in the design and development of the AI capability, including:

- What frameworks apply (i.e. Defence's Data Management Strategy, Defence's privacy framework, Defence's security framework), and how they have been complied with;
- Does it comply with Defence specific:
    - policy,
    - design specifications,
    - technical requirements,
    - tender documents,
    - evaluation or performance metrics,
    - or any other known Defence requirements?
- What relevant frameworks, policies, standards etc. have not been applied, or cannot be applied, and why?

H.9.v Describe the control measures that are (or can be) applied to the AI capability, the system it is integrated into, to improve the safety, security, reliability, predictability, controllability, maintenance of responsibility, or governability (at all stages of its life cycle and for all inputs and outputs), including:

- what control measures are attached to the range of decisions that the AI functionality can make?
- what human, functional and technical control measures are relevant for Defence implementation of the operation of the AI capability.
- what control measures you have put in place to achieve these elements?
- what extra control measures could be applied to enhance these elements?
- what control measures should Defence undertake to minimise the risks relating the development and use of the AI capability?

H.9.vi Describe the oversight, audit or review systems that you have put in place to review the AI capability, including:

- the standards used;
- the results of any such review, audit or oversight;
- corrective action taken to remedy any deficiencies;
- future requirements for further audit or review, or ongoing oversight.

Answering the above questions address the risks relating to the ability of the AI to be certified for use and assured for ongoing use.

### H.10 Integrated

H.10.i When considering the AI capability (or system in which the AI is integrated) including the AI system or output:

- Can, and if so how can, they be used in compliance with relevant domestic or international legal requirements?
- Are there any additional subject matters experts required to guide AI governance and assurance development (including kind and type of experts)?

- What are the AI education, training, and command and control requirements?

Answering the above question addresses the risks of the AI capability being ability to fit within Defence's system of control.

## H.11 Safe

### H.11.i Please describe how the AI capability satisfies Department of Defence safety requirements.

### H.11.ii Describe the standards and processes you utilised to assess the safety (and associated risks) of the AI capability, or the system it is integrated into?

If your AI capability is:

- a component of a weapon or means of warfare;
- is designed to control a weapon, means or method of warfare;
- is part of a life support system;
- is integrated into a system which the unintended actions to the AI capability could cause injury, death, or more than menial damage to property (or other compensable interests);

then please:

- complete Annex A;
- describe the control measures you have applied to your capability to ensure that it meets relevant regulatory and governance requirements;
- identify the exact standards and processes you have utilised to evaluate risks from intended and unintended AI outputs;
- identify measures to improve the safety to users of the AI capability, or the system it is integrated into;
- identify measures to improve the safety to the object of the AI output.

### H.11.iii describe the control measures that are (or can be) applied to the AI capability, or the system it is integrated into) to minimise the risks or hazards from its operation (at all stages of its life cycle), including:

- what control measures you have put in place to ensure risks are minimised?
- what control measures could minimise the risks of the operation of the AI, or otherwise improve the safety, for users and the objects of the AI?
- what control measures should Defence undertake to minimise the risks of the operation of the AI, or otherwise improve the safety for users and objects of the AI?

Answering the above questions address the risks to safety of ADF personnel and civilians who are going to rely upon the operation of the AI and the failure of that AI functionality will risk human life, that must be mitigated in the Risk Register.

### H.12 Secure

H.12.i Describe how the AI capability complies with Department of Defence's security framework?

H.12.ii Describe the security measures the AI capability, or the system it is integrated into:

- is required by regulation (whether legislative or other binding regimes) to comply with;
- Is recommended by codes, best practices, industry standards, or other non-binding to comply with;
- you have included additional to any specific regulatory or voluntary compliance requirements.

H.12.iii Describe the control measures that are (or can be) applied to the AI capability, the system it is integrated into, to improve the security (at all stages of its life cycle and for all inputs and outputs), including:

- what control measures you have put in place for security?
- what extra control measures could be applied to enhance security?
- what control measures should Defence undertake to minimise the security risks relating the development and use of the AI capability?

Answering the above questions address risk associated with control measures and security that would render the system capable of being inappropriately corrupted.

## Part Four – Risk measurement methodologies for the elements

This LEAPP does not mandate which measurement methodology is most suitable to address each identified risk in the LEAPP. However, the below list does identify some available methodologies which could be used to support each measurable element.

There may be a requirement to utilise or adopt multiple methodologies for each element; and the outcomes of a single methodology may not be sufficient to answer the questions related to that measurable element. These measures can be one of three main types, namely:

- human (or educational);
- technical; or
- procedural.

There is no mandatory assurance risk management methodology, either mandated by the Commonwealth Government, or Department of Defence, when it comes to LEA. There are, however, a number of trends that can be identified in risk assessment methodologies to address legal and ethical risk in AI.[19] These are:

- The principles that form AI ethics frameworks can directly inform the AI Risk Assessment process by identifying the risk factors that need to be assessed. The principles that inform the LEAPP risk assessment have been derived from existing Australian ethics and AI frameworks, here, represented as ‘elements’.
- There is a widespread trend towards considering risks arising from violation of AI principles (set, for example, at national, European, international levels). These risks have been identified in the body of the LEAPP, and broken down by
- An AI risk assessment should be an ongoing process throughout the design, development and deployment of AI technologies. This has been addressed by revisiting the LEAPP at the various acquisition gateways.

There are two primary risk methods that can be applied in the approach to risk assessment for AI:

1. Analyse risk and categorise risk of AI overall (i.e., the AI capability in its anticipated use context is, overall, a low, medium of high risk); or
2. Analyse individual risks associated with the AI capability and assign risk status to each of those individual risks.

The LEAPP adopts the second approach.

Further, there are two primary regulatory approaches that can be expected to be adopted when introducing a complex, highly technical capability into a system, namely, performance-based and management-based regulation. The LEAPP is drafted to anticipate that a hybrid of these regulatory approaches will be adopted (following similar emerging and disruptive technology approaches, such as that of cyber). In this regard, there will be some risks that will require meeting to an articulated performance metric, while others will require descriptive risk management and process-driven responses. In each case where a specific performance-based standard is required, this will require mandating from the acquisition agency/adopting user.  This standard is likely to be identified during later design/development stages, when this document is used as a risk mitigation prompt in consultation with Defence.

[19] EY, A Survey of AI Risk Assessment Methodologies, Aug 21

Some example methodologies that can be applied to the elements for Responsible AI for Defence are listed below. These are not exhaustive, and some methodologies may not be appropriate for particular AI capability. As the RAID is a broad LEA Toolkit, it is intended to facilitate the LEA of all AI capabilities. AI risk assessment methodology appropriate for an HR tool, for example, will differ widely from the requirements of a tool designed for target identification in the use of lethal force.

## 1. Responsible

The key internal practice available to States to ensure responsibility is part of an AI capability is 'baking' human intent into the code. This practice is measurable in a number of ranging from code review (premised on design specifications to statistical analysis of Testing, Evaluation, Verification and Validation (TEVV) applied to code function and capability effects. Other examples of methods to measure this element include: training, education, orders regarding activation; use of blockchain to record which commander/use is temporally responsible for the AI capability at any given time, assessment, testing and certification processes applied to operators and commanders prior to being permitted to operate or command operation of an AI. Clear articulation of responsibility for the AI and attendant risks at each stage of the AI capability life cycle – including where appropriate apportioning 'levels of responsibility' across its design and use phases, to individuals including operational commanders.[20]

Human Factors and Ergonomics (HFE) methods relevant to this element include: 'task analysis, cognitive task analysis, process charting, situation awareness assessment, trust assessment, mental workload assessment, teamwork assessment, interface analysis, usability evaluation, design, systems analysis, risk assessment, and accident analysis'.[21]

## 2. Accountable

The primary measures of accountability will be State prescriptions and military processes. State prescriptions (including legal frameworks, orders, doctrine, and policies) detail where accountability for using a capability rests. Military processes (often a component or result of State prescriptions) provide accountability through prescription or consequence of practice. For instance, an example methodology to measure accountability for the use of AI, may be formulated in and recorded by a blockchain enabled HOTO of command and control.

Specific frameworks include:

- Defence Risk Management Framework.

[20] MEAID

[21] USC HF&E Report

### 3. Understandable

This element can be measured by testing of technical knowledge of AI operators and commanders on the AI, physical and virtual testing and certification based upon anticipated use and use limitations, test transparency of the AI capability from activation to operation to post activity review, HFE methods and assessment of the accuracy of user mental models of the AI capability.

### 4. Explainable

Design measures to measure this element are those that ensure the AI is ‘intrinsically interpretable’. For example, undertaking TEVV testing to assess design measures. Other methods would include applying purpose-built AI to ‘interpret’ the AI system being validated; as well as standard HFE methodologies.

### 5. Reviewable

Discoverability, recordability and auditability of the algorithmic processes in the exercise of the algorithmic functionality will require integration with a military’s extant acquisition and design processes; and have regard to the standards of information and evidence required for potential future investigations and interrogation in the event of misuse of accident. This element is strongly linked to understandability and explainability.

Specific review frameworks include:

- DSTG Technical Risk Management Assessment;
- Article 36 Weapons Review.

### 6. Reliable

This element deals with biases; including considering biases in the data set itself, and the structures and systems in which they are being modelled. It can be measured using physical and virtual testing of to identify statistical failure rates; and through TEVV that addresses inputs and context for its specific use case. It can also be measured using adversarial testing and assessment against standards of performance. HFE methods to measure this element include, ‘cognitive task analysis, process charting, human error identification, situation awareness assessment, trust assessment, mental workload assessment, teamwork assessment, interface analysis, usability evaluation, performance time prediction, design, systems analysis, risk assessment, and accident analysis’.[22]

### 7. Predictable

Predictability can be broken into three parts:

- the degree to which system’s technical performance is or is not consistent with past performance;

[22] USC Report

- the degree to which any AI or autonomous system's specific actions can (and cannot) be anticipated; and
- the degree to which the effects of employing an AI system can be anticipated.[23]

Measuring this element could be achieved through TEVV, testing of data sets for hygiene (and in particular for bias mitigation), the conduct of adversarial testing, virtual/simulation and live field training. The focus of any such measures will be on the effects from such activities and will have a strong statistical component. The concept of predictability is often referred to as the 'black box' problem with AI; and thus the development and application of 'standardized metrics to grade predictability and understandability' to test against will likely to be a requirement to appropriately measure this element.

## 8. Compliant

Methods to measure this element include:

- Thorough TEVV, with a focus on identifying and assessing effects for predictability and controllability;
- Assessing the extent to which appropriate values and standards have been designed into the AI capability;
- Assessing the extent to which inputs – such as data – have been appropriately validated. For instance, ensuring good data hygiene can reduce output bias; algorithm TEVV can avoid algorithmic bias; and value sensitive design can offset automation bias.
- Use of compliance toolkits, such as the AI Fairness 360 toolkit, to identify biases and discrimination.[24]

HFE methodologies to measure this element could include: human error identification, situation awareness assessment, trust assessment, mental workload assessment, design, systems analysis, risk assessment, and accident analysis.[25]

## 9. Controllable

In the Australian context, this element refers to the systems that are placed over the system. Measuring this element requires demonstration of compliance with the control mechanism, whether that be physical, software or hardware related, or derived through a system of control for the AI capability when deployed within the specific use-case environment.

[23] Holland Michel, Arthur. 2020. 'The Black Box, Unlocked: Predictability and Understandability in Military AI.' Geneva, Switzerland: United Nations Institute for Disarmament, at p.5. Research. doi: 10.37559/SecTec/20/AI1

[24] See A Method for Ethical AI in Defence ('MEAID')

[25] University of Sunshine Coast Human Factors Report ('USC Report')

HFE methodologies that could be utilises include: physical HFE methods, task analysis, cognitive task analysis, process charting, human error identification, situation awareness assessment, trust assessment, teamwork assessment, interface analysis, usability evaluation, design, systems analysis, risk assessment, and accident analysis.[26]

## 10. Integrated

The integration of the AI into the various systems will come from measures derived to assess other elements but would benefit from TEVV to identify effects of AI functionality when operated with the various systems it is expected to interact with. As such the continuation of such testing once the AI has been accepted into service through virtual or simulated training, as well as live training in various levels of controlled to semi-uncontrolled environments would be necessary to measure this element.

## 11. Safe

There exist a wide variety of safety risk measurement and management methodologies, but the safety of the AI system can generally be measured using recursive testing, evaluation, verification and validation process, where systems (both active and non-active learning). Continual and iterative testing and certification will support the measurement of the system's safety.[27] Specific safety frameworks include:

- Defence Safety Management Framework.

## 12. Secure

Physical, cyber and EMS security measures should be considered in this element. Specific security frameworks include:

- Cyber Security Risk Assessment.

[26] USC Report

[27] Black box

# Annex A - Article 36 Review of Weapons, Means and Methods of Warfare Requirements

The purpose of this Annex is to assist those developing AI capabilities for military use that may relate to the use of force in armed conflict.

## Article 36 Review of Weapons, Means and Methods of Warfare Requirements

In Australia, an AI capability[28] that is determined to be a new weapon, means or method of warfare will be subject to an Article 36 review if it is going to be acquired by Defence. An Article 36 Review is a process used by Defence to conduct a legal review of new weapons, means and methods of warfare, proposed to be used by Defence. The purpose of the Review is to ensure that new capabilities intended for use during armed conflict can be used in a manner that complies with Australia's international legal obligations.

The Australian Government has determined that the use of AI-enabled capabilities that support or enable a new weapon, means or method of warfare must be in accordance with government direction,[i] compliant with Australian law and consistent with Australia's international legal obligations.[ii]
To assist Defence industry developing AI capabilities that may fall within Australia's definition of a new weapon, means or method of warfare, the following content outlines the basic information likely to be required by Defence should your AI capability undergo an Article 36 weapons review.

This process consists of eight steps:

Step 1 – Is your AI capability a weapon, means or method of warfare?
Step 2 – Design, technical, and performance characteristics of AI capability.
Step 3 – What is the normal anticipated use?
Step 4 – Specific law
Step 5 – General law
Step 6 – Other relevant international law
Step 7 – Public interest and the Martens Clause
Step 8 – Domestic law (if necessary)

These steps and other considerations relevant to designers and developers of AI capabilities are explained further below. The law relevant to Article 36 Reviews is a specialised field of law. Designers and developers of AI capabilities are encouraged to obtain expert legal advice on the issues detailed in this Annex to ensure their AI capability will meet legal compliance requirements.

[28] In this Annex, AI capability also includes an AI-enabled capability.

## Step 1 – Is your AI capability a weapon, means or method of warfare?

### Requirement

The requirement to conduct an Article 36 Review is based upon a determination that the AI, or the system that it enables, is assessed to be a new weapon, means or method of warfare that the Australian Department of Defence intends to use in armed conflict.

While these terms have no officially recognised international definition, Australia has provided guidance on what it considers each of these terms to mean. This information will assist in determining whether Defence would require an Article 36 Review to be conducted on your AI capability prior to final acquisition. Defence has defined the phrase 'weapon, means or method of warfare' to include:

- any 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 [objects].[29]

New weapons are:

- weapons that are being designed or created;
- those that are being acquired from another State; or
- those that are being materially altered.[30]

Defence will likely on require an AI capability to undergo an Article 36 Review where the AI capability is integral to the exercise of combat functionality, and then only with respect to that part of an AI capability's function that undertakes or enables that combat functionality. For example, a part of an AI capability that provides an AI function to fly an aircraft would not usually require review unless it also directly enabled targeting effects while flying.

'Combat functionality' combines the elements of:

- combat (the use of violence by armed forces); and
- functionality (the purpose/task that the instrument is designed or expected to undertake to represent the range of actions or functions that a weapon is capable of undertaking to apply violence).

Violence in this sense ranges from actions preparatory to a use of force (such as authorising, searching, detecting, tracking, identifying, selecting, cueing, prioritising, determining fire control), through to those using the harming mechanism during use of force actions (applying kinetic or non-kinetic force where violence is intended or expected to neutralise, damage, destroy, detain, injure, or kill). The complexity comes with determining which aspect of AI functionality is relevant to combat functionality. Identifying those components that are part of the AI capability and integral to (commonly referred to as 'critical to')

[29] Presentation by CMDR SJ White, Directorate of Operations and International Law, *Weapons review of New Weapons, Means and Methods of Warfare* (September 2020) <*https://documents.unoda.org/wp-content/uploads/2020/10/Weapons-reviews-2020-Australia.pdf*>.
[30] Defence FOI 187/20/21 Document 1, Defence Administrative Policy 1, 23 November 2020

causing damage or harm is essential to identifying what aspects of your capability will need to be capable of passing an Article 36 Review.

**For designers or developers.**

To assist designers and developers of AI capabilities anticipate whether there is a need to prepare for Defence to conduct an Article 36 Review of their capability as part of its acquisition process, the following questions (which form part of the RAID Checklist) should be considered during the design and development phases of AI capabilities:

1. Is the AI capability a ‘weapon, means or method of warfare?
    a. Is it a weapon?
    b. Is it a means?
    c. Is it a method?

If the answer is yes, to any of the above, it a ‘new’ weapon, means or method of warfare – complete this Annex.

2. If the answer is no to 1. a-c – determine if your AI capability is critical to the combat functionality of the system the AI is integrated into or otherwise controls (or aspects of).

If the answer is yes – complete this Annex.

If the answer is no – you may choose to complete this Annex but there is no requirement to do so as your AI capability is unlikely to be subject to an Article 36 Review by Defence during its acquisition process.

**Step 2 – Design, technical, and performance characteristics of AI capability.**

**Requirement**

The second step of an Article 36 Review requires a detailed description of the design, technical specifications, and performance of a new weapon, means or method of warfare. This information assists Defence determine the scope of the weapon review and includes:

- **Design.** The design characteristics and specifications of the capability.

- **Technical specifications**. This could include technical guidance (design, manufacturing process, material composition, fusing system, guidance system, integrated safety procedures and safeguards), ballistics information (speed, accuracy, damage mechanism, delivery mechanism, effects et al), analysis and assessments of weapons effects, and appropriate subject matter expert advice on the technical characteristics of the weapon or means.

- **Performance.** The performance characteristics, including a description of how it ‘operates’, and any relevant health and environmental considerations. Guidance on the conduct of Article 36 Reviews by the International Committee of the Red Cross (ICRC) suggests that 'relevant factors would include: the accuracy and reliability of the targeting mechanism (including e.g. failure rates, sensitivity of unexploded ordnance, etc); the area covered by the weapon; whether the weapons' foreseeable effects are capable of being limited to the target or of being controlled in time or space (including the degree to which a weapon will present a risk to the civilian population after its military purpose is served).'[31]

### For designers or developers

It is recommended specialist legal advice is obtained on potential military legal considerations relevant to AI capabilities during design and development phases. However, indicative specification pre-requisites that could assist in meeting the above requirements are listed below:

- **Design.** Design specifications include AI functionality, AI models, AI metrics and standards, development and environmental data, HMI, AI capability output. It must include a description of the data sets used to train any neural networks and the criteria used to inform the AI operation. For example, if the AI is intended to classify persons and objects, the data set description should include what information has been used to classify each category of person or object.

- **Technical specifications**. Technical characteristics including how the AI models work, inbuilt control mechanisms, sensors etc.

- **Performance.** Performance characteristics include how the AI functionality operates to enable AI capability including how effective any AI functionality is in areas such as reliability and predictability.

## Step 3 – What is the normal anticipated use?

### Requirement

Step 3 in an Article 36 Review identifies the 'normal anticipated use' of the weapon or means, and the manner in which that use occurs ('method').[32] This description includes the designed purpose or intended effect of the AI capability. Specific reference should be made to factors such as how it works (how it harms or damages), and what it would be used to target or engage. It is possible that an AI capability will provide the functionality to undertake, or make a recommendation to undertake, a method of warfare i.e., if it aids human decision-making. The 'normal anticipated' use of an AI capability must therefore articulate how it makes or implements decisions and the criteria the AI relies upon.

[31] ICRC Guide to the review of new weapons, means and methods of warfare (The ICRC Guide).
[32] Article 36 requires weapons reviews to encompass the use of a weapon or means 'in some or all circumstances'. This has been practically interpreted by States as a 'normal or anticipated' or 'normal or expected' use interpretation, and not all possible uses (including unlawful misuses) or effects of a weapon, means or method.

The use of AI can complicate the description of the 'normal anticipated use' in several significant ways. The 'normal anticipated use' is typically assessed in terms of both the intended or anticipated effect and how it achieves the intended or anticipated effect; but it is also a direct consequence of the AI capability being a capability to undertake methods of combat functionality. Complications can arise in the following situations:

- **Machine learning.** Where the AI utilises machine learning (i.e., the algorithms have adaptive capacity for learning, modification, or optimisation) there is the potential for the AI capability to adapt in response to training input, thus resulting in a change in either the intended effect of the capability or the method by which it achieves that intended effect.

- **Novel or emergent behaviours.** If the AI capability does not employ an algorithm that can change, the novel and dynamic contextual environment it is used in may result in novel or emergent behaviour (i.e., variation in its method of achieving its intended or anticipated effect).

- **AI selected engagement options.** Where the AI is responsible for selecting the means of undertaking the harm and can choose from a range of means and methods of employment then the assessment of 'normal and anticipated use' will become significantly more complex.

An inability to identify, and therefore assess, an AI capability's 'normal anticipated use' will mean the capability is either unlikely to pass an Article 36 Review or that the Review process will result in recommendations that limit the use of the capability to only those aspects of it that can be assessed as meeting Australia's legal obligations.

AI capabilities that enable the normal anticipated use of a means or method of warfare to be adaptive will require ongoing monitoring to determine whether the threshold requirement to conduct a new Article 36 Review is met.   That is, the requirement to undertake an Article 36 Review would be triggered because the 'normal anticipated use' is new.

### For designers or developers

An inability to identify, and therefore assess, an AI capability's 'normal anticipated use' will mean the capability is either unlikely to pass an Article 36 Review or that the Review process will result in recommendations that limit the use of the capability to only those aspects of it that can be assessed as meeting Australia's legal obligations.

AI capabilities that enable the normal anticipated use of a means or method of warfare to be adaptive will require ongoing monitoring to determine whether the threshold requirement to conduct a new Article 36 Review is met.   That is, the requirement to undertake an Article 36 Review would be triggered because the 'normal anticipated use' is new.

Consideration should be given to variables such as:

- the 'nature' of the algorithms, such as the level and complexity of code, permitted code adaptability, and decision-implementation parameters;
- the level and use of reliability, performance and legal standards applied to a decision-implementation capability; and
- the 'nurture' of the algorithms from data diet, training regimes, through to human interaction and application of internal and external controls.

Consideration should be given to when the capability can:

- change the intended effect of the capability or the method by which it achieves that intended effect;
- demonstrate novel or emergent behaviour when in a new environment; or
- independently select its method or means of employment.

### Step 4 – Specific law

#### Requirement

Step 4 in an Article 36 Review is the determination of whether a means or method is specifically prohibited or restricted by any treaties or customary international law that binds Australia.[33]

At the time of writing, there are no specific prohibitions on AI, or AI enabled, capabilities;[34] however, the development of an AI capability that relies on a prohibited or restricted means or method of warfare would be identified in this stage. For example, in Australia, an AI capability that directly or indirectly enables the use of cluster munitions would be identified as prohibited due to Australia being a party to the Cluster Munitions Convention.

[33] *Legality of the Use or Threat of Nuclear Weapons* (Advisory Opinion) [1996] ICJ Rep 226 (*Nuclear Weapons Advisory Opinion*). The Court adopted a two-fold test: (1) Is there any customary or treaty law that contains a *specific prohibition* against the threat or use of a weapon in general or in certain circumstances? (2) In the absence of a specific prohibition, is there a *general prohibition* against the threat or use of a weapon in general or in certain circumstances?

[34] For example, Chief of the Australian Defence Force, *Australian Defence Doctrine Publication (ADDP) 06.4: Law of Armed Conflict* (11 May 2006) Ch 4 provides a list and description of specifically prohibited or restricted weapons.

### For designers or developers

Consideration should be given to whether the AI capability is attached to, or enables, another capability that is novel to Australia (that is, is attached to a platform or munition that is not already in the Australian inventory).

## Step 5 – General law

### Requirement

Step 5 in an Article 36 Review requires consideration of whether there is a general prohibition against the effects caused by a means or method of warfare in general, or in certain circumstances (i.e. a focus on the effect of weapons or the means of warfare).

This analysis includes reviewing the technology against established principles relevant to unnecessary suffering, indiscriminancy, and environmental harm to determine whether the proposed means is unlawful *per se* or in certain circumstances:[35]

### The Traditional Instrument Review – Unnecessary suffering[36]

A means of warfare in its 'normal anticipated use' must be employable without causing superfluous injury or unnecessary suffering to combatants.[37] Article 36 Review analysis on this issue typically requires a balancing of the injury or suffering caused to combatants by the capability against the military utility of using that capability (i.e., its military necessity).

In practice, this aspect of an Article 36 Review will trigger consideration and comparison of:  the injury or suffering caused by the subject means or warfare against an existing capabilities that are designed to provide the same military utility, have a similar effect, or are of a similar type; any incidental or secondary effects; substantive differences in the harming mechanism or the consequent injuries or illnesses that result; and health/medical information (foreseeable harms, permanence of impairment, mechanism of injury, diagnosability,

[35] For AP I States this would involve the articulation of the AP I and customary rules; for non-AP I States only the customary international law positions. In particular for AP I, see art 35(2), covering 'weapons, projectiles and materials and methods of warfare that cause superfluous injury or unnecessary suffering'; art 51(4), covering indiscriminate attacks; and art 35(3), covering environmental modification. See Jean-Marie Henckaerts and Louise Doswald-Beck, *Customary International Humanitarian Law* (Cambridge University Press 2005) vol 2, rules 43–45, 70 and 71 for the status of these requirements.

[36] The term 'Traditional Instrument Review' is used in this Annex to refer to the type of weapons review in use by States to undertake review of weapons systems that do not contain any AI and refers to the review of the item that is used to direct or control the part of the weapon system that delivers the lethal effect.

[37] Treaty obligation of Art 35(2) to AP I and customary international law. Unnecessary suffering means 'a harm greater than that unavoidable to achieve legitimate military objectives'. While it is noted that the art 35(2) language ('weapons, projectiles and material and methods of warfare') is arguably narrower in application then that of art 36 ('new weapon, means or method of warfare'), most AP I States who undertake weapons reviews apply the broader requirements of art 36 when determining compliance with the prohibition against superfluous injury and unnecessary suffering.

treatability). This comparison enables a comparison between an existing, lawful, capability and the capability undergoing the Article 36 Review.

### For designers or developers

The inclusion of AI to replace or reduce human control over the use of a weapon requires considerations beyond the traditional review of an instrument's capability to cause unnecessary suffering. This consideration must focus upon whether the weapon design causes unnecessary suffering. This is specifically the case if the AI creates a capability to undertake a method that includes application of the primary rule of unnecessary suffering – as outlined above – as a method (on behalf of a human or operator). This is particularly important where the AI can determine when an attack should end (as this changes the requirement of military necessity to prosecute the attack). It becomes necessary to determine whether the AI can avoid the continuation of the attack causing (further) injury that is in effect superfluous or unnecessary. Using AI to apply unnecessary suffering would require assessment of:

- the AI capability's ability to avoid disproportionate suffering (by stopping an attack on a combatant once the military necessity to continue the attack ends – such as through rendering the combatant unable to fight as a result of the attack); or

- the AI capability's ability to avoid 'rendering death or permanent impairment inevitable' (through failing to identify correctly humans whose status as a combatant has ceased or failing to stop an attack on a combatant when the circumstances of the attack have changed).

A Review of an AI capability for the purpose of assessing the requirement of the principle of unnecessary suffering would require a determination regarding whether the AI capability can either apply the principle correctly, or otherwise avoid undertaking action that is contrary to this principle. This requires assessment of the AI capability's ability to correctly apply (in context) some, or all, of the concepts of:

- **military necessity** - identify if it is militarily necessary to attack the adversary and with what level of harm (see for instance Article 51(5)(b) and Article 57(2)(a)(ii) of AP1);

- **attack** (Article 49(1)) - know what actions are 'acts of violence against the adversary, whether in offence or in defence';

- **distinction** (i.e. Article 57(2)(a)(i)) - be able to differentiate combatants from civilians (Article 50(1)) or hors de combat (i.e. Article 33, 41 and 44) and thus ensure only combatants are attacked):

    - **combatants** (categories of persons referred to in Article 4 A (1), (2), (3) and (6) of the Third Convention or under Article 43 of AP1) or civilians directly participating in hostilities (Article 51(3)) – are capable of being identified;

- **civilian status** (Article 50(1)) – the capability is able to apply, where required, the presumption of civilian status for persons; and

- know what levels of effect/harm/injury are required to defeat the combatant.

Given the specialised nature of this field of law, (International Humanitarian Law (IHL)), It is recommended that designers and developers obtain expert advice to assist them in addressing these considerations.

### The Traditional Instrument Review – Environmental damage

A means of warfare in its ‘normal anticipated use’ must not be designed or expected to cause widespread, long-term and severe damage to the natural environment.[38] In a weapon review this will traditionally focus on the ‘normal anticipated’ use of the means of warfare and the instrument’s harming mechanism to assess whether it can avoid causing widespread, long-term and severe damage to the natural environment.
The threshold for long-term damage is generally understood to be measured in decades rather than year.[39]

### For designers or developers

The inclusion of AI to permit decision-implementation requires considerations beyond the traditional review of an instrument’s capability to cause environmental damage. That is, the AI creates a capability to undertake a method that includes application of the primary rule prohibiting causing widespread, long-term and severe damage to the natural environment – as outlined above – as a method (on behalf of a human or operator). This is of relevance where the AI, or AI-enabled, capability has control of the harming mechanism of a capability or platform. Where the designed or expected effects engage this rule then its specific requirements may need specific programming to ensure the AI capability will adhere to this rule. Because it is a specific prohibition it must be coded into the AI capability as a red line, to ensure compliance.

This would also require the AI capability to potentially consider:

- what is the natural environment;
- what is widespread, long-term and severe damage to the natural environment;

[38] AP I arts 35(3) and 55(1). *Convention on the Prohibition of Military and any other Hostile Use of Environmental Modification Techniques* (adopted 18 May 1977, entered into force 5 October 1978) 1108 UNTS 151 art 1 prohibits parties from the hostile use of environmental modification techniques having widespread, long-lasting or severe effects as the means of destruction, damage or injury.

[39] The ICRC Guide (n 31).

- if the instrument causes that widespread, long-term and severe damage to the natural environment;
- how to take ‘care’ to protect the natural environment (Art 55(1)); and
- the prohibition on reprisals using attacks on the natural environment.

### The Traditional Instrument Review – Indiscriminancy

A means of warfare in its ‘normal anticipated use’ must be capable of being used discriminately.[40] For an Article 36 Review this entails an assessment of whether a capability employs a method or means of combat:

- which cannot be directed at a specific military objective; and
- the effects of which cannot be limited as required by AP I;

‘and consequently, in each such case, are of a nature to strike military objectives and civilians or civilian objects without distinction’.[41]

### For designers or developers

The use of AI to permit decision-implementation requires considerations of a capability’s ability to cause indiscriminate effects. That is, the AI could undertake a method of warfare that includes application of the primary rule of indiscriminancy – as outlined above – on behalf of a human operator. This is relevant where the AI capability has control of the harming mechanism of the instrument.

The factors considered in reviewing the indiscriminancy principle will include accuracy, type, duration, and extent of effects.
The operation of the AI capability will need to comply with AP1 Article 51(4), as well as the various distinction requirements contained within AP1 and/or customary international law. This might require an assessment of the AI capability to:

- direct attacks only at specific military objectives (Article 51(4)(a));

- determine whether it is possible to direct an AI capability at a specific military objective (Article 51(4)(b)); and

- determine whether the harmful effects, in space or in time, of an AI capability can be limited to legitimate military targets (Article 51(4)(c)).

To apply the primary rule will potentially require the AI, or AI-enabled, capability to correctly apply (in context) some, or all, of the following concepts:

- **attack** (described above);

- **military objectives** – be able to identify military objectives either under defined parameters or in accordance with the two-pronged test in AP I

[40] Under weapons law – art 51(4) of AP I and customary international law – weapons, means and methods that are indiscriminate are prohibited.

[41] AP I art 51(4)(a)

Article 52(2) – whether an object 'by their nature, location, purpose or use' makes 'an effective contribution to military action'; and whose 'total or partial destruction, capture or neutralization, in the circumstances ruling at the time, offers a definite military advantage.'

- **distinction** (be able to differentiate between legitimate military objects and civilian objects, or be able to differentiate combatants from civilians or hors de combat – for instance distinction under Article 57(2)(a)(i));
    - combatants and civilians directly participating in hostilities;
    - civilian status (be able to apply where required the presumption of civilian status for objects (Article 52(3)) or persons);
- an understanding of effects (at the time of, and for a period after, the attack – Article 51(c)); and
- precautions of attack generally (Article 57(2)(a)(i)-(ii)) (explained further below).

An Article 36 Review would pay particular attention to how the AI capability satisfies the subjective requirements of targeting law (potentially by programming higher objective standards into the code); and predictability of effects. This is a complex area noting the technological problems – bias, brittleness etc. – in being able apply targeting law accurately in the circumstances of extreme variability found in combat.

### Traditional Instrument Review – Proportionality

A means of warfare must be capable of being used proportionately (see Article 51(5)(b) and Article 57(2)(a)(ii) of AP1). That is, it must not cause incidental harm to civilians and/or civilian objects which is **excessive** in relation to the concrete and direct military advantage anticipated.

This requirement has two components. First, the harming mechanism of the capability must be capable of being used proportionately (i.e. is it indiscriminate because it is expected to cause excessive incidental harm when compared to the military advantage). Second, and more specifically for AI capabilities, the AI functionality must be capable of executing human intent in accordance with the proportionality standard. AI, capabilities could be utilised to undertake components of the proportionality assessment or execute the entire assessment. In other words the AI capability needs to be assessed on its ability to apply (rather than just comply with) the principle of proportionality. The proportionality assessment requires assessment of:

- 'concrete and direct' military advantage (Article 51(5)(b) and Article 57(2)(a)(ii) of AP1 - there must be a 'relevant and proportional contribution to the objective of the military attack involved');

- collateral damage (Article 51(5)(b) and Article 57(2)(a)(ii) – 'incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof');

- the abstract and subjective concepts of 'excessive' (in relation to collateral damage weighed against military advantage) and circumstantial change (regarding status) requiring cancellation.

To apply these requirements the AI capability may also be required to apply the rules relating to:

- attack;
- military objectives;
- combatants and civilians directly participating in hostilities);
- presumption of civilian status;
- distinction;
- an understanding of effects (at the time of, and for a period after, the attack); and
- choice between military objectives (Article 57(3)).

An Article 36 Review will pay particular attention to how the AI, capability identifies actual collateral effects (if not determined by a human operator), and then computes the varying subjective and qualitative standards under proportionality. If an AI capability were acquired from another State consideration would need to be given to ensuring Australian legal standards are met by the AI (not the legal standards of the selling State).

- **Other LOAC/IHL considerations relevant to an AI enabled weapon, means or method (as required)**

If the AI enables the weapon to cause a harmful effect (for example identifying, selecting and/or recommending targets, deciding when or how to attack, assessing collateral damage) there are additional IHL rules governing the use of force that must be considered.

**Conducting attacks.** An Article 36 Review does not typically assess whether a weapon, means or method of warfare could be used in compliance with the precautions in attack. This is because the precautions in attack, outlined in Art. 57, are fundamentally directed at the human decision-makers. Article 36 Reviews, however, traditionally do not consider the lawfulness of human behaviour, focussing consideration on the lawfulness of the weapon, means, or method warfare.

This approach will likely change due to the inclusion of AI to assist or undertake decision-making and/or decision-implementation. This is because the AI in the combat functionality operates to replace direct human decision-making. That is, the algorithms of the AI will implement decisions on conducting attacks ranging from determining the target is a legitimate and appropriate military objective, to ensuring any incidental damage or injury is minimised and proportionate. The AI therefore functions as a capability to undertake a method of combat (in effect AI in the combat functionality will give effect to human intent in complying with some or all of the art 57 obligations). As such, to the extent that the AI capability purports to enable the exercise of combat functionality on behalf of human users, it will have to be assessed for compliance with the Art. 57 obligations.

An Article 36 Review will need to determine whether the AI capability in its 'normal anticipated use' implements decisions that are covered by the rules on attack; and if so, analyse the ability of the AI to participate in the conduct of such attacks within the requirements of Article 57. The relevant sub-articles of Article 57 are:

- Article 57(1) – constant care obligation.
- Article 57(2) – feasible precautions. Because AI is a decision-implementation tool in accordance with Article 57(2)(a) it will need to ensure that it permits those who plan or decide upon an attack to comply with:
- everything feasible in applying the principle of distinction or identifying that there is no special protection (Article 57(2)(a)(i));
- taking all feasible precautions in the choice of means and methods of attack with a view to avoiding, and in any event to minimizing, incidental loss of civilian life, injury to civilians and damage to civilian objects (Article 57(2)(a)(iii)); and
- refraining from deciding to launch any attack which is not expected to be proportionate (Article 57(2)(a)(iii)).
- Article 57(2)(b) – cancel or suspend an attack if it becomes apparent that the principles of distinction or proportionality are not met, or if there is an issue of special protection.
- Article 57(2)(c) – where circumstance permit, provide effective advance warning if the attacks may affect the civilian population.
- Article 57(3) – 'When a choice is possible between several military objectives for obtaining a similar military advantage, the objective to be selected shall be that the attack on which may be expected to cause the least danger to civilian lives and to civilian objects'.
- Article 57(4) – 'In the conduct of military operations at sea or in the air, each Party to the conflict shall, in conformity with its rights and duties under the rules of international law applicable in armed conflict, take all reasonable precautions to avoid losses of civilian lives and damage to civilian objects'.

**Specific prohibitions.** Included within the precautions in attack is reference to 'special protection'. AP I and II (as well as other LOAC treaties and customary international law) provide rules that give certain objects and people special protection that restrict them from being attacked or require certain processes to be undertaken prior to an attack (or military operation) being conducted against them.

Where these specific protections are not otherwise covered by prohibitions or restrictions under the general principles of LOAC the specific prohibition or restriction may need to be coded into the AI capability as a 'red line' prohibiting certain actions.

Similarly, rules which certain objects and people special protection, that restrict them from being attacked or require certain processes to be undertaken prior to an attack (or military operation) being conducted against that object or person may need to be programmed into an AI capability. Coding these specific protections into an AI capability will assist in ensuring the AI capability does not breach the special protection and enhance its compliance with Art, 57. Example rules related to special protection include:

- Protection of wounded, sick and shipwrecked – Article 10 and Article 7 of AP II;
- Protection of (including identification of as detailed in Article 18);
- Medical Units – Articles 12 & 13 (rules for discontinuance) and Article 11 of AP II;
- Civilian medical personnel – Article 15 and Article 9 of AP II;
- Medical vehicles, vessels and aircraft and coastal rescue vessels – Articles 21-24 and Article 11 of AP II;
- Prohibition on Perfidious Acts (distinguished from ruses) – Article 37;
- Prohibition on use of specified emblems (and restrictions on use of adversary's emblem) – Articles 38-39 and Article 12 of APII;
- Prohibition on 'no quarter' - Article 40;
- Safeguard of enemy *hors de combat* – Article 41;
- Restrictions on attacking persons parachuting from an aircraft – Article 42;
- Restrictions on attacking medical personnel and chaplains of the combatants – Article 43(2) and Article 9 of APII;
- Protection of persons who have taken part in hostilities – Articles 44-45;
- Protection of cultural objects and of places of worship– Article 53 and Article 16 of APII;
- Protection of objects indispensable to the survival of the civilian population – Article 54 and Article 14 of APII;
- Protection of works and installations containing dangerous forces – Article 56 and Article 15 of APII; and
- Protection of members of the armed forces and military units assigned to civil defence organizations – Article 67.

### For designers or developers

Where relevant, it is expected these rules would need to be converted into code, and then be capable of being applied in compliance with Article 57. The safeguard responsibilities under Art 57(2)(a)(i) or Article 57(2)(b) with respect to special protections would also need to be applied.

As noted above, the extent to which each rule needs to be addressed is dependent upon the 'normal anticipated use' of the A capability and linked to the environment in which is intended to be deployed.

## Additional Article 36 Reviews requirements under an AI, or AI enabled, capability review

An Article 36 Review of new weapon, means or method of warfare operates on the presumption of lawful use (i.e. that the responsibility for the lawful use of the capability belongs to the weapon user and those responsible for its use) and therefore focusses on whether a capability can be used lawfully rather than whether its use is lawful in a particular circumstance.

### For designers or developers

The introduction of AI to a weapon capability has the potential to negate the reliance on this presumption of lawful use and require careful consideration of the IHL rules governing the use of force in offence and defence. That is, depending upon the normal anticipated use of the AI capability, the review may need to consider the ability of the AI to comply with that part of IHL/LOAC which was traditionally addressed with human decision-making and action. The requirements for each AI capability will need to be assessed on a case-by-case basis. Furthermore, the aspects of the AI capability that address IHL requirements, or performance standards that permit compliance with IHL will need to be built into the AI capability.

## Step 6 – Other relevant international law

An Article 36 Review will evaluate new weapons, means or methods of warfare against any other applicable international law (relevant to conduct in armed conflict) binding upon a State, or that a State may choose to apply for policy reasons.

This requirement is assessed on a case-by-case basis in the context of the capability itself and any relevant Defence policy directives.

## Step 7 – Public interest and the Martens Clause

### Requirement

An Article 36 Review will evaluate new weapons, means or methods of warfare to determine whether it is in the public interest, the public conscience, or principles of humanity to study, develop, acquire or adopt a weapon, means or method of warfare.

The public interest entails considering any relevant concern (i.e., moral, economic, and policy) that effects legitimacy of weapon, means or method of warfare. From a legal perspective, public interest is usually restricted to developing but unsettled law or future legal trends in the law, that will potentially result in the weapon, means or method of warfare becoming illegal or being rendered militarily unnecessary.  A further consideration is provided for in Article 1(2) of AP I, commonly known as the Martens Clause which states that:

In cases not covered by this Protocol or by other international agreements, civilians and combatants remain under the protection and authority of the principles of international law derived from established custom, from principles of humanity and from the dictates of public conscience.

Australia applies the Martens Clause as preserving the operation of customary international law, including the customary principles of distinction, proportionality, prohibition against unnecessary suffering, that the means and methods of warfare are not unlimited, and extend to the protections contained under Common Article 3 to the Geneva Conventions.42 In addition to this narrow application, there are two other major competing interpretations of the Martens Clause; that it can be used as an interpretive provision, or in its widest sense that it permits the consideration of the principles of humanity or the dictates of public conscience in the application of the law. While Australia does not publicly acknowledge adhering to either of these interpretations of the Martens Clause, this does not mean it will not in the future. It is possible that other States may apply the Martens Clause in different ways to Australia, including using one of these other major applications.

### For designers or developers.

It is anticipated that in applying best practice during an Article 36 Review, Defence, as a matter of policy, will seek to identify any information suggesting the:

- capability is abhorrent to the public conscience, or it offends the principles of humanity;[43] and
- use of the capability is not in the public interest.

If such information is identified, the Review will likely consider both criteria to determine if either will affect acceptance of the capability.[44] [45][46]There is yet to be any Australian-government position released on the use of fully autonomous lethal weapons system, however, designers and developers of these types of AI capabilities should remain informed on policy developments in Australia related to the limits of use of AI capabilities to ensure their designs will meet Defence (Australian) requirements for acquisition.

[42] Rupert Ticehurst, ‘The Martens Clause and the Laws of Armed Conflict’ (1997) 317 International Review of the Red Cross 125.

[43] This is the position taken by Human Rights Watch and International Human Rights Clinic, ‘Losing Humanity: The Case Against Killer Robots’ (19 November 2012) 36 and Christof Heyns, ‘Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions’ (Human Rights Council, 9 April 2013) UN Doc A/HRC/23/47, in calling for a ban on LAWS. The argument can be flipped. For instance, Galliott and Scholz claim there is a moral imperative to develop weapons that are more compliant with LOAC. For example, weapons that are capable of averting attacks on protected symbols, protected sites and signals to surrender. See Jai Galliott and Jason Scholz, ‘AI in Weapons: The Moral Imperative for Minimally-Just Autonomy’ (2018) 1(2) Journal of Indo-Pacific Affairs 57.

[44] Many States adopt a policy position that is closer to the broad position in practice, albeit this is not to be taken as acceptance of this as being its legal position on their interpretation of the Martens Clause.

[45] See Parkes regarding weapons review being ‘cognizant of trends in the law of war or arms control law’.

[46] This could also include creating a public forum to allow the public and other interested parties an opportunity to comment on the issue.

There is yet to be any Australian-government position released on the use of fully autonomous lethal weapons system, however, designers and developers of such capability should ensure that they stay up to date on policy developments related to the limits of use of AI, and AI-enabled, capabilities to ensure their designs are likely to be accepted into service when they are reviewed.

**Step 8 – Domestic law (if necessary)**

### Requirement

There is no requirement to review domestic legal obligations during an Article 36 Review. However, during a review, issues may be raised in relation to the application of domestic law to the AI capability. An example of an exception is Australian laws that incorporate sanction control requirements.
Domestic law prohibitions or limitations on the ability of Defence (the ADF) to have the AI capability in its inventory, or on its subsequent use, must be addressed within the weapons review.[47]

### For designers or developers

Defence policy may require additional domestic legal requirements to be incorporated into its Article 36 Review process as its policies and processes relating to AI evolve. For example, Workplace Health and Safety (WHS) is an area that may be considered in the future. While WHS laws can be suspended to varying extents in armed conflict it will become increasingly more appropriate to assess the 'normal anticipated use' WHS requirements of an AI capability during acquisition, acceptance, training and pre-deployment certification. This will require consideration of elements of understandability, explainability, safety, security, controllability, reliability and predictability.

[47] Where the capability is likely to be used domestically by the military (as part of an armed conflict), then the relevant domestic law should be considered.

# ANNEX B – DEFENCE WARFIGHTING FUNCITON

The information below provides further details and examples of Defence warfighting functions and categories of activity that summarise how Defence conducts its functions.

The enterprise-wide capabilities address normal day-to-day functioning of Defence, as well as those required to field ADF force elements in combat. The purpose of listing the warfighting functions relevant to the AI capability is to assist in identifying where the AI functionality will be used and how it will contribute to Defence's functions.

Use the below functions to describe the use case of your AI capability if complete Component G – Use Case Environment.

## Annex B. 1 Combat or Warfighting Functions

| Tag | Command (Cmd) |
|---|---|
| Description | The process and means for the exercise of authority over, and lawful direction of, assigned forces.[48] |
| AI examples | AI used to support strategic, operational and/or tactical planning, including optimisation and deployment of major systems<br>AI used in modelling and simulation used for planning and mission rehearsal<br>AI that supports management of policy and procedures<br>AI used to optimise business and administrative processes, including modelling and simulation tools<br>AI used for enterprise business planning at the strategic, operational and/or tactical level |

| Tag | Force Application (FA) |
|---|---|
| Description | The conduct of military missions to achieve decisive effects through kinetic and non-kinetic offensive means. |
| AI examples | Autonomous weapons (AWs) and autonomous/semi-autonomous combat vehicles and subsystems<br>AI used to support strategic, operational and/or tactical planning, including optimisation and deployment of major systems<br>AI used in modelling and simulation used for planning and mission rehearsal<br>AI used in support of the targeting cycle including for collateral damage estimation<br>AI used for Information Warfare such as a Generative Adversarial Network (GAN-) generated announcement or strategic communication |

[48] ADF Concept for Command and Control of the Future Force, citing ADDP 01.1 Command and Control Ed 2 AL1.

| | |
|---|---|
| | AI used to identify potential vulnerabilities in an adversary force to attack<br>AI used for discrimination between combatants and non-combatants |

| | |
|---|---|
| Tag | Force Protection (FP) |
| Description | All measures to counter threats and hazards to, and to minimise vulnerabilities of, the joint force in order to preserve freedom of action and operational effectiveness |
| AI examples | Autonomous defensive systems (i.e. Close in Weapons Systems)<br>AI used for Cyber Network Defence<br>AI used to develop and employ camouflage and defensive deception systems and techniques<br>Autonomous decoys and physical, electro-optic or radio frequency countermeasures<br>AI to identify potential vulnerabilities in a friendly force that requires protection<br>AI used to simulate potential threats for modelling and simulation or rehearsal activities<br>Autonomous Medical Evacuation/Joint Personnel Recovery systems |

| | |
|---|---|
| Tag | Force Sustainment (FS) |
| Description | Activities conducted to sustain fielded forces, and to establish and maintain expeditionary bases. Force sustainment includes the provision of personnel, logistic and any other form of support required to maintain and prolong operations until accomplishment of the mission. |
| AI examples | Autonomous combat logistics and resupply vehicles<br>Automated combat inventory management<br>Predictive algorithms for the expenditure of resources such as fuel, spares and munitions<br>Medical AI systems used in combat environments and expeditionary bases<br>Predictive algorithms for casualty rates for personnel and equipment<br>Algorithms to optimise supply chains and the recovery, repair and maintenance of equipment<br>Algorithms to support the provision of information on climate, environment and topography<br>AI used for battle damage repair and front-line maintenance |

| | |
|---|---|
| Tag | Situational Understanding (SU) |
| Description | The accurate interpretation of a situation and the likely actions of groups and individuals within it. Situational Understanding enables timely and accurate decision making. |

| | |
|---|---|
| AI examples | AI that enables or supports Intelligence, Surveillance and Reconnaissance (ISR) activities including:<br>object recognition and categorisation of still and full motion video<br>removal of unwanted sensor data<br>identification of enemy deception activities<br>anomaly detection and alerts<br>monitoring of social media and other open-source media channels<br>optimisation of collection assets<br>AI that fuses data and disseminates intelligence to strategic, operational and tactical decision makers<br>Decision support tools<br>Battle Management Systems<br>AI that supports Command and Control functions<br>Algorithms used to predict likely actions of groups and individuals<br>AI used to assess individual and collective behaviour and attitudes |

## Annex B.2 Enterprise-level and Rear Echelon Functions

| Tag | Personnel (PR) |
|---|---|
| Description | All activities that support the Raising, Training and Sustaining (RTS) of personnel. |
| AI examples | AI used for Human Resource Management including:<br>record keeping<br>posting and promotion<br>disciplinary and performance management<br>recruitment and retention<br>modelling of future personnel requirements<br>prediction of HR supply and demand events and anomalies<br>AI used in individual and collective training and education including modelling and simulation<br>AI used for testing and certification of personnel<br>AI used to model the capability and preparedness of permanent and reserve personnel |

| Tag | Enterprise Logistics (EL) |
|---|---|
| Description | Activities that support rear-echelon enterprise-level logistics functions including support of permanent military facilities |
| AI examples | Autonomous rear-echelon supply vehicles and warehouses<br>AI used for optimisation of rear-echelon supply chains and inventory management<br>AI used in depot-level and intermediate maintenance, including:<br>Digital twinning<br>Predictive maintenance<br>Global supply chain analysis, prediction and optimisation<br>Enterprise-level analysis and prediction for resource demand and supply (i.e. national/strategic fuel requirements)<br>AI used in the day-to-day operation of permanent military facilities |

| Tag | Business Process Improvement (BP) |
|---|---|
| Description | Activities that support rear-echelon administrative business processes that are not related to personnel or logistics. |
| AI examples | AI used for Information Management and record-keeping<br>Informational assistants such as policy chatbots<br>AI that supports management of policy and procedures<br>AI used to optimise business and administrative processes, including modelling and simulation tools<br>AI used for enterprise business planning at the strategic, operational and tactical level |

**ANNEX C – Legal and Ethical Assurance Working Group**

This Annex is designed to outline the LEAWG structure, operation and terms of reference.

## LEAWG Terms of Reference

[See for example: https://www.apsc.gov.au/sites/default/files/2021-02/examples_of_well_and_poorly_drafted_tor.pdf]

## LEAWG Objectives

[Insert LEAWG Objectives here]

## LEAWG Membership and Points of Contact

The LEAWG consists for the following members:

Chair: [insert name, email contact]

Secretary: [insert name, email contact]

Legal lead: [insert name, email contact]

Ethics lead: [insert name, email contact]

Governance lead: [insert name, email contact]

Members: [insert names, email contacts]

## LEAWG Meeting Details

[Insert meeting details here:

E.g.: The LEWAG will meet quarterly on the first Monday of the quarter, at 1400h.

The meetings will be held in hybrid format, with in-person meeting locations to be coordinated

The LEAWG Secretary will distribute the meeting agenda two (2) weeks prior to the meeting; and minutes will be provided no later than three (3) weeks after the LEAWG concludes.]

[i] Including overarching political aims, legal basis, policy objectives, and the command guidance derived from military end states and objectives. Australia (System of Control, 25-29 March 2019).
[ii] Australian Intervention on Ethics and Sociology - CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems (LAWS) – 14 May 2014. Australia (Item 6b, 29 August 2018). Australia (26 March 2019).