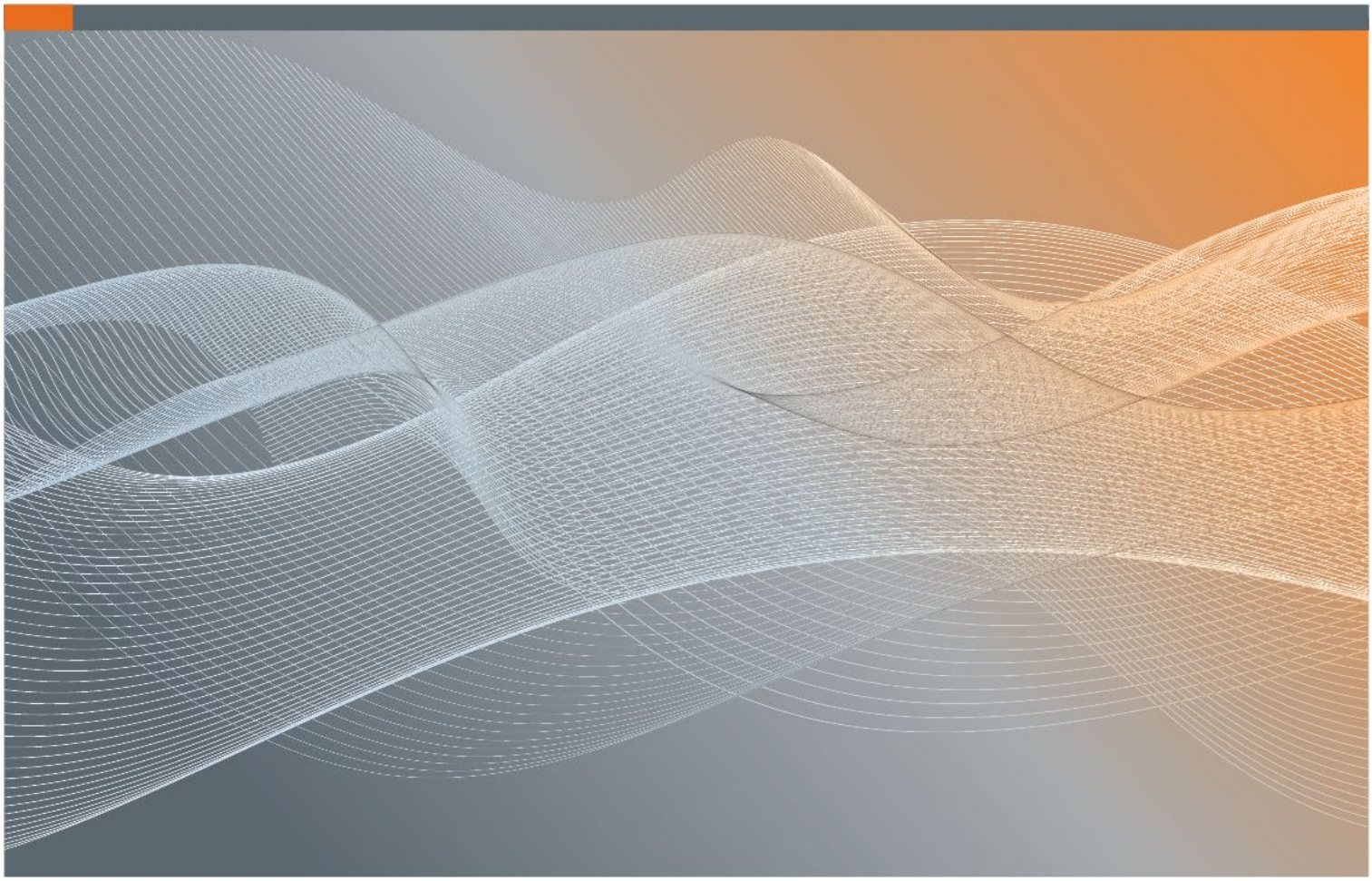
Australian Government
Department of Defence

Case Study: A Method for Ethical AI in Defence Applied to an Envisioned Tactical Command and Control System



Defence Science and Technology Group

DSTG-TR-3847

Defending Australia and its National Interests
www.defence.gov.au

DSTG-TR-3847

Authors
Dianna Gaetjens, Kate Devitt, and Christopher Shanahan

Produced by
Aerospace Division
Defence Science and Technology Group
Department of Defence
PO Box 7931
Canberra BC ACT 2610

www.dst.defence.gov.au

Telephone: 1300 333 362

Conditions of Release and Disposal

This document is the property of the Australian Government; the information it contains is released for defence purposes only and must not be disseminated beyond the stated distribution and secondary release statement without prior approval of the Releasing Authority.

The document and the information it contains must be handled in accordance with security regulations, downgrading and delimitation is permitted only with the specific approval of the Releasing Authority.

This information may be subject to privately owned rights.

The officer in possession of this document is responsible for its safe custody.

EXECUTIVE SUMMARY

The use of artificial intelligence (AI) in a defence context poses significant ethical questions and risks. Defence will need to address these as AI systems are developed and deployed in order to maintain the reputation of the ADF, uphold Australia's domestic and international legal obligations, and support the development of an international AI regime based on liberal democratic values.

This report, *Case Study: A Method for Ethical AI in Defence Applied to an Envisioned Tactical Command and Control System*, is the product of a scientific and technical (S&T) collaboration between the Department of the Prime Minister and Cabinet (PM&C), Defence and the 3A Institute at the Australian National University (ANU). It uses *A Method for Ethical AI in Defence* [1] to explore the ethical risks in an envisioned AI-enabled tactical command and control (C2) system that integrates a variety of autonomous functions in order to assist a single human operator in managing multiple uninhabited vehicles simultaneously.

The analysis of this envisioned C2 system using *A Method for Ethical AI in Defence* generated key findings for three stakeholder group: whole-of-Defence; AI technology developers, and those seeking to use or iterate *A Method for Ethical AI in Defence*.

For Defence, the report identifies critical policy gaps and recommends action on:

- an **accountability** framework for decisions made by and with AI
- **education and training** of operators, command and systems developers
- managing the **data** underpinning many AI applications, including its collection, transformation, storage and use.

Without action, these gaps leave Defence vulnerable to significant reputational and operational damage.

Additional key findings for AI technology developers relate to the topics of **effectiveness**; **integration**; **authority pathway**; **confidence**; and **resilience**. In aggregate, these findings encourage developers to consider whether the most efficient system or algorithm (for example, in terms of speed or accuracy), is necessarily the best in terms of providing assistance to the decision-maker. In some cases, a less efficient algorithm that is more consistent with normative decision-making may be more appropriate. In addition, there is a clear need for research on what information is necessary to make good judgements (especially where issues are complex and context is important); and how it should be

rapidly conveyed. These key findings can be explored further through consideration of seven hypothetical ethical risk scenarios developed as part of the analysis.

For those seeking to apply or iterate *A Method for Ethical AI in Defence*, the report recommends the development of **additional tools** to assist practitioners in identifying areas of maximum relevance and utility for their specific needs; and a comprehensive set of **definitions** to assist in applying the method.

CONTENTS

1. INTRODUCTION 1

2. *A METHOD FOR ETHICAL AI IN DEFENCE* 3

3. THE ENVISIONED TACTICAL C2 SYSTEM 7
   - **3.1. Allied IMPACT (AIM) 7**
   - **3.2. Additional Capabilities of the Envisioned Tactical C2 System 10**
   - **3.3. The exemplar system versus the envisioned system 10**

4. KEY FINDINGS 12
   - **4.1. Whole-of-Defence 12**
     - 4.1.1. Accountability 12
     - 4.1.2. Education and Training 13
     - 4.1.3. Data 14
   - **4.2. Findings for AI and Human-Machine Interface developers 14**
     - 4.2.1. Effectiveness 15
     - 4.2.2. Integration 16
     - 4.2.3. Confidence 16
     - 4.2.4. Resilience 17
     - 4.2.5. Decision support for targeting 17
   - **4.3. Findings for application of *A Method for Ethical AI in Defence* 18**
     - 4.3.1. Development of additional tools 18
     - 4.3.2. Terms and definitions 19
     - 4.3.3. Linking facets and tools 19

5. ACKNOWLEDGEMENTS 20

6. REFERENCES 21

APPENDIX A. APPLYING *A* METHOD FOR ETHICAL AI IN DEFENCE TO THE ENVISIONED TACTICAL C2 SYSTEM 25

APPENDIX B. KEY FINDINGS FROM APPLYING A METHOD FOR ETHICAL AI IN DEFENCE 27

APPENDIX C. APPLYING THE FIVE FACETS OF ETHICAL AI TO THE ENVISIONED TACTICAL C2 SYSTEM 30
   - **C.1. Responsibility 30**
   - **C.2. Governance 34**
   - **C.3. Trust 48**

C.4. Law 58
C.5. Traceability 60

APPENDIX D. KEY FINDINGS FOR DEFENCE 63
D.1. Accountability 63
D.2. Education and training (E&T) 65
D.3. Data 66

APPENDIX E. ETHICAL RISK SCENARIOS 69
E.1. Scenario: Conflicting information 69
E.2. Scenario: Transfer of Tactical Control 70
E.3. Scenario: Limitations, explainability and over-riding the system 71
E.4. Scenario: Attack 72
E.5. Scenario: Flawed data sets 73
E.6. Scenario: Different outcomes 74
E.7. Scenario: Personnel advancement 75

1. INTRODUCTION

Applications of artificial intelligence (AI) are transforming how we live and structure our societies, breaking down once impenetrable barriers. Increasingly, decision-makers are recognising that AI is likely to effect change at an unmatched scale, speed and level of impact. For Australia, this presents an opportunity to enhance prosperity and growth, improve living standards, and make our people and institutions safer and stronger.

For the defence sector, AI has the potential to transform military functions by increasing the speed and scale of military operations; facilitating autonomous operations; and improving military decision-making [2]. Ultimately, these capabilities could protect Australia while keeping Australian Defence Force (ADF) personnel safer. This potential is driving the Defence's strategic focus on AI, as seen in the *2020 Defence Strategic Update* [3] and *2020 Force Structure Plan* [4].

For all its potential, the use of AI, especially in a defence context, also poses significant risks. Responsible use of AI requires the resolution of complex ethical questions regarding autonomy, accountability, and governance. How much autonomy should AI-driven systems have? Who is responsible for decisions made by autonomous systems? How can we interrogate the values and assumptions underpinning AI-driven decisions and ensure they are safe, nuanced, and aligned with existing Defence structures and culture?

Australia is not alone in asking these questions, nor in recognising the transformative nature of AI. The impacts of AI development are playing out globally – and they are not neutral. Driven in part by an increasingly adversarial geo-strategic environment, the US and China have both recognised the potential for AI to confer a strategic advantage and are competing to become world leaders in the field[1].

However, the competition over AI is about more than technological supremacy. It is part of a much greater contest of interests, values and ideology that seeks to shape the global order. The way AI is developed, used and governed will be a critically important battleground in that contest – one in which Australia has a lot at stake.

Put simply, consideration of how Australia develops and deploys AI need to look beyond a risk-benefit analysis. Our decisions represent clear choices about our interests, our values, and ultimately, what kind of society and world we want. Seen through that lens,

[1] China announced its ambition in 2017, with the release of its AI strategy, which includes a strong focus on civil-military integration [40]. The US National Defense Strategy, released in 2018, identifies AI as a key technology for enabling the US to fight and win wars of the future [41].

our obligation to develop AI that upholds strong ethical standards also reflects the strength of our commitment to liberal democratic values.

The Australian Government has recognised the need for ethical AI. In 2019, the Government released the *AI Ethics Framework* [5], which establishes guidelines for interrogating ethical risk as civilian AI systems are built and deployed. In 2021, the Defence released *A Method for Ethical AI in Defence* [1], a robust framework to guide the ethical development and operation of AI systems within the military domain.

This report represents the first formal application of *A Method for Ethical AI in Defence*. It has been applied to an envisioned AI-enabled tactical C2 system, with the analysis relying heavily on the Allied IMPACT (AIM) system as an exemplar. AIM is a novel multi-UxV C2 research environment developed by the US, the UK, Canada and Australia through The Technical Cooperation Program (TTCP). During Autonomous Warrior 2018, AIM demonstrated the potential for autonomous functions to give human operators meaningful control over multiple heterogeneous, multi-domain autonomous vehicles operating in littoral scenarios[7]..

The goals of this study were to:

- provide a test case for *A Method for Ethical AI in Defence*, demonstrate its applicability, and identify areas for further work
- analyse an envisioned AI-enabled tactical C2 system to identify potential ethical risks, and to develop ethical risk scenarios to highlight and communicate these risks to inform future design
- identify avenues for further whole-of-Defence further work on ethical risk.

Section 2 provides an overview of *A Method for Ethical AI in Defence* and how it was applied. Section 3 describes AIM and other AI-enabled technology (such as operator-state monitoring) likely to feature in the envisioned tactical C2 system. Section 4 presents the key findings of the analysis, broken down into recommendations and observations for three groups: Defence leadership; AI technology developers; and practitioners seeking to apply or iterate *A Method for Ethical AI in Defence*.

Key findings for Defence include proposals to develop policy frameworks on accountability; education and training; and data. Other key findings for AI technology developers relate to effectiveness; integration; decision support for targeting; confidence; and resilience. Finally, this we propose an expansion of *A Method for Ethical AI in Defence* to include definitions and a suite of tools to assist practitioners in identifying areas of maximum relevance and utility for their specific needs.

2. A METHOD FOR ETHICAL AI IN DEFENCE

Figure 1 Facets of ethical AI in Defence [1].

A Method for Ethical AI in Defence [1] is a pragmatic new method that emerged from a highly collaborative, evidence-based workshop with diverse attendees (jointly sponsored by the Royal Australian Air Force's [RAAF] Plan Jericho, the Trusted Autonomous Defence Cooperative Research Centre and the Defence Science and Technology Group [DSTG]). It does not represent the views of the Australian Government.

The method is intended as a resource to assist Defence stakeholders to reduce ethical risk on AI projects. Underpinning this is the idea that ethical compliance should be incorporated and assessed throughout the system lifecycle, from design to deployment, including as a system is developed, iterated, evaluated, implemented and/or adapted [1].

The method suggests that AI project managers can reduce ethical risk by asking questions about five facets (Figure 1), which are divided further into 20 topics:

1. **RESPONSIBILITY**: Who is responsible for AI?
   Topics: education and command
2. **GOVERNANCE**: How is AI controlled?
   Topics: effectiveness, integration, transparency, human factors, scope, confidence and resilience
3. **TRUST**: How can AI be trusted?
   Topics: sovereign capability, safety, supply chain, and test and evaluation; misuse and risks; authority pathway; and data subjects
4. **LAW**: How can AI be used lawfully?
   Topics: protected symbols and surrender and de-escalation
5. **TRACEABILITY**: How are the actions of AI recorded?
   Topics: explainability and accountability

The 20 topics represent a sample of evidence-based considerations emerging from a single workshop and are anticipated to be extended significantly in subsequent workshops with different stakeholders. The location of each topic under any specific facet was debated by Defence stakeholders, e.g. 'accountability' could be reasonably positioned under all facets. The authors of *A Method for Ethical AI in Defence* acknowledge this and hope that users of the method recognise that the relationship between the facets and topics could be adapted as Defence develops more robust policy and frameworks for ethical AI[2].

[2] Conveyed in conversation with report co-author Kate Devitt

The method includes three tools for AI project managers and teams: an *Ethical AI for Defence Checklist*, the *Ethical AI Risk Matrix* and the *Legal and Ethical Assurance Program Plan (LEAPP)*. Of these, this report used elements of the *Ethical AI for Defence Checklist*, which consists of the following steps:

1. Describe the military context in which the AI will be employed.
2. Explain the types of decisions supported by the AI.
3. Explain how the AI integrates with human operators to ensure effectiveness and ethical decision making in the anticipated context of use and countermeasures to protect against potential misuse.
4. Explain framework/s to be used.
5. Employ subject matter experts to guide AI development.
6. Employ appropriate verification and validation techniques to reduce risk.

This method is flexible and adaptable according to the context within which it is being used. For this report, the checklist was adapted to incorporate the following steps:

- The military context and types of decisions supported by the AI were examined through a data collection process including a document review [1],[7], interviews with subject matter experts[3], and a demonstration of AIM at DSTG Edinburgh (19 and 20 December 2020).
- Potential ethical risks were identified by examining the envisioned tactical C2 system using each of the 20 topics from *A Method for Ethical AI in Defence* (Appendix C).
- Where the potential for significant ethical risk was identified, hypothetical scenarios (Appendix E) were generated to help DSTG personnel further explore the nature of the risk and identify suitable mitigation strategies for incorporation into future tactical C2 system design.
- The analysis in the preceding steps allowed the generation of key observations and recommendations for three stakeholder groups, namely: whole-of-Defence; AI technology developers; and practitioners seeking to apply or iterate on *A Method for Ethical AI in Defence* (Section 4, Appendices B, C and D),
- For more detail on how the method was applied, see Appendix A.

[3] Dr Kate Devitt, Dr Christopher Shanahan, Marcin Nowina-Krowicki, Dr Greg O'Keefe, Dr Steven Wark and Dr James Brooks.

The comprehensiveness and level of consultation possible for this report was limited by the length of time available (from 16 November 2020 to 14 February 2021[4]) and the human resources dedicated to the task (1.0 full time equivalent [FTE] workload). Many of the topics warrant further investigation, and suggestions for further work throughout the report add up to the potential for a significant program of future research.

[4] See the Memorandum of Understanding for Dianna Gaetjens between DSTG and PM&C 27 September 2011.

3. THE ENVISIONED TACTICAL C2 SYSTEM

3.1. Allied IMPACT (AIM)

AIM was used as an exemplar for the envisioned AI-enabled tactical C2 system. AIM is a proof-of-technology decision support system that integrates 8 autonomous modules (Figure 2) with the aim of enabling a single human operator to control multiple, multi-domain, autonomous uninhabited vehicles from different countries simultaneously.

The AIM research environment was jointly developed by Australia, Canada, US and the UK as part of the TTCP, Autonomy Strategic Challenge. The main operational goals of the Challenge were to demonstrate [6]:

- *Force multiplication*: significantly reduce the number of human personnel required to operate uninhabited vehicles and increase the number of uninhabited systems able to be controlled by a single operator.
- *Enhanced Five Eyes*[5] *(FVEY) interoperability*: provide architecture enabling joint FVEY operations using uninhabited vehicles.
- *Integration*: integrate autonomy technologies from FVEY partners, reducing duplication of capability development.
- *Agility*: merge tactical and operational control for faster military decision cycles.

[5] The Five Eyes are Australia, the United States, the United Kingdom, Canada and New Zealand.

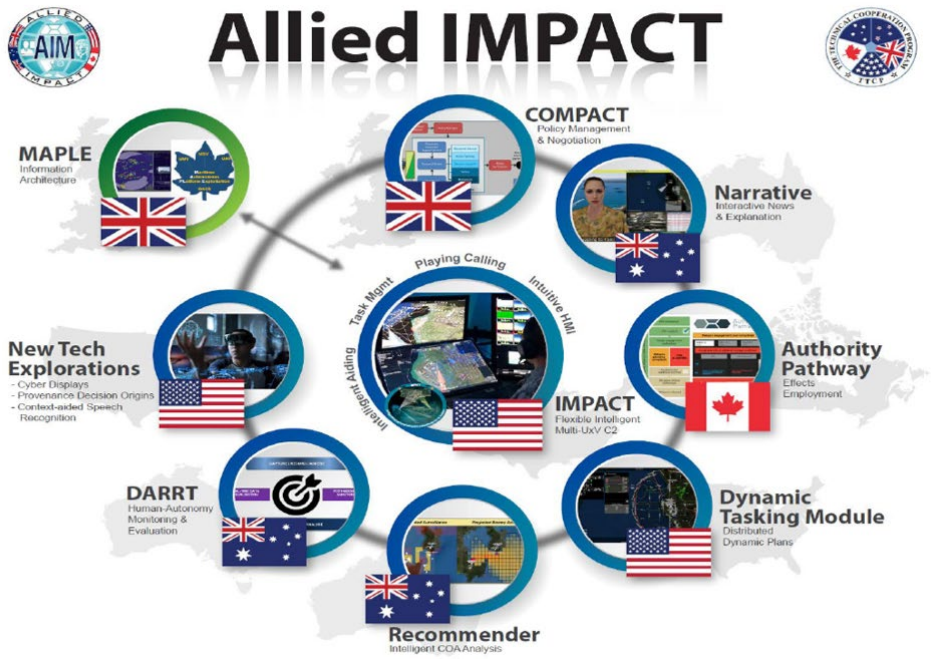


Figure 2 Autonomous modules integrated into AIM [7].

The ability to achieve these goals was successfully tested by all FVEY partners during Exercise Autonomous Warrior 2018 (AW18), demonstrating full integration of 22 components and allowing a single operator to manage 17 multi-domain uninhabited assets. Given that a single uninhabited asset usually requires a team of around four people, AW18 thus demonstrated the potential for a single AIM operator to replace upwards of 68 people.

Through extensive feedback protocols (delivered largely through the use of the DSTG Assessment and Review Research Tool [DARRT]), AW18 also allowed system developers to perform rapid and interactive live and after-action evaluation of the system.

AIM was built around the U.S. Department of Defence's Intelligent Multi-UxV Planner with Adaptive Collaborative Control Technologies (IMPACT) research testbed. This system utilises a task delegation approach in which a single operator uses a novel, intuitive interface to call a 'play'[6] in order to articulate a mission objective; intelligent decision aids

[6] 'Play' calling comes from sports like American football, and is a method used to quickly assign pre-determined roles or tasks to team members through the use of a series of memorised keywords. For AIM, calling a play is a rapid method for tasking vehicles in response to different types of events. Available plays include, for example, 'inspect a point' or 'search an area'. Calling a particular play prompts AIM to suggest an asset and plan to complete the mission.

then assist in the rapid development, execution, and modification of uninhabited vehicle plans[6].

The other seven autonomous modules that make up AIM provide a suite of high-level functions, including:

- the ability to deploy multiple uninhabited assets (UxV) (hundreds if desired), and add or delete them dynamically during execution (Dynamic Tasking Module[7])
- a policy negotiation tool (COMPACT), that allows mission plans and tasks recommended by IMPACT to be cross-checked against operational policies. During AW18, COMPACT implemented policies related to air space management and route de-confliction (although other policies could be added)
- an interactive weapons engagement stateboard (Authority Pathway module), based on the Laws of Armed Conflict, which takes an operator through a series of nine required steps to engage a target and release a weapon
- a multimedia narrative system (Narrative module), which provides adaptive mission briefings for operators; multimodal questions and answers Q&A and notifications; and explanation of system recommendations and provenance on demand
- the ability to calculate and recommend surveillance areas of interest that maximise threat detection while minimising search areas (Recommender module). IMPACT autonomously assigns unallocated assets to base patrol activities
- the ability to capture information about how data has evolved, the activities involved with manipulating data and the agents responsible for actions (Provenance module)

From the operator's point of view, AIM's capabilities can be roughly divided into four different types of operator support:

- the entirely autonomous[8] performance of routine and background functions (such as plan monitoring) and high priority tasks when necessary (via the Task Manager)
- the generation of recommendations which require operator approval (such as vehicle allocation and mission planning)[9]

[7] The Dynamic Tasking Module also includes capability to allow for continued communication between assets even if broader communications are lost or limited.

[8] Even where AIM performs autonomous functions in theatre, this does not mean the activities have not been authorised at all. Pre-approval for autonomous functions is issued by command as part of AIM's system design. Therefore, the distinction here lies in when and from whom authorisation is received.

[9] This distinction becomes important later when considering issues of accountability and responsibility.

- guiding an operator through the steps in a process (such as Authority Pathway)
- information gathering and communication (such as the Narrative).

3.2. Additional Capabilities of the Envisioned Tactical C2 System

In addition to the capabilities present in AIM, future AI-enabled tactical C2 systems are likely to possess the capacity to adapt to the workload and performance of the user. This capability would allow the AI-enabled system to react to a user under pressure by implementing – just as a human would – adaptive team-oriented behaviours such as anticipating information requirements, monitoring for errors, offering assistance, and dynamically reallocating (or taking on) tasking.

DSTG [8] is currently developing new measures (specifically voice analysis and eye tracking) for monitoring the operator's cognitive load and frameworks for adapting tasking and information flows accordingly. For example, during times of high cognitive load (such as when the operator is busy with multiple concurrent tasks), adaptive AI-enabled system behaviour may include filtering and prioritising situational information provided to the operator in order to minimise disruptions to their concentration. Alternatively, the AI technology could allocate additional incoming tasks to another operator, or complete a greater number of tasks autonomously.

Another potential capability for future tactical UxV systems involves the automation of functions currently performed by a human sensor operator. Currently, the sensor operator monitors feeds coming in from the UxV, and provides relevant information from those feeds to the UxV operator and to other command elements. However, machine learning models could replace these functions and either feed information to the UxV operator and command elements, or use it to inform autonomous system behaviour.

3.3. The Exemplar System Versus the Envisioned System

Our exemplar system – AIM – was designed and developed as a research system. Built as a proof-of-technology, it was never intended for deployment. In addition, developers faced a Herculean task in integrating the various modules from different countries into one system within a short timeframe for the purposes of demonstration and evaluation during AW18. Therefore, many features that would be built into a deployable system are either rudimentary or not present, and some of the underlying architecture is not as robust as it could be.

However, in the context of this report, it makes little sense to limit the consideration of potential ethical risk by thinking of AIM as a research system. Thus, this report imagines

that AIM, and the additional capabilities that we anticipate will feature in future tactical C2 systems, could be deployed and considers ethical risk from that standpoint. In doing so, we hope to strike a good balance between the exemplar system as it is now, and a plausible vision of what a system *like this* would look like in the future. In other words, it is not the intention of this report to pick at known technical issues in a system that was not designed to be operationalised. Rather, it is to imagine what ethical risks might arise if a system like AIM was further developed to the point where it could be deployed.

4. KEY FINDINGS

Key findings were generated as an output of the analysis in Appendix C. They are split according to their relevance for three stakeholder groups: whole-of-Defence; AI technology developers; and practitioners seeking to apply or iterate *A Method for Ethical AI in Defence*.

More detail on all the key findings is in the appendices.

4.1. Whole-of-Defence

The following findings relate to significant ethical risks that have the potential to manifest in many (or most) AI-driven systems developed or deployed by Defence. This creates an imperative for whole-of-Defence consideration of these issues, since the absence of coherent, Defence-wide strategies to mitigate these risks will leave Defence facing significant and unacceptable reputational and operational risks.

In offering these observations, it is acknowledged that Defence may already have policies in place to address some or all of these issues (or parts thereof). Resource constraints for this project did not allow a thorough examination of existing policies. However, work to examine existing policies, where they might apply, and where there are gaps would be a logical first step in the process of developing the policies recommended here.

4.1.1. Accountability

Defence should develop an accountability framework to clarify when operators and commanders will be held accountable for decisions made by or with AI systems, with appropriate consideration of what is reasonable and fair. This process should also examine how accountability for decisions using AI fits in with existing Defence policy.

The issue of accountability when using AI-driven systems is important because the envisioned tactical C2 system has the ability to autonomously make decisions and carry out tasks without operator authorisation. They can also make recommendations for a course of action which may require human authorisation to operationalise, but may not provide much information on how the recommendation was produced. This is in contrast to non-AI driven systems, where (in the Defence context at least), the operator can rely on the fact that a chain of human assessment sits behind proposed courses of action, even if the operator is not privy to all the detail.

In developing an accountability framework, key questions to consider include:

- What types of decision can acceptably be made autonomously?
- Who is accountable for a decision made autonomously?
- If an autonomous system recommends a course of action which requires human authorisation, what level of knowledge does the authoriser (operators/commanders) need about how that action has been recommended and how the recommendation will be executed by an autonomous system in order for it to be reasonable to hold them accountable for that decision?
- What level of explainability do autonomously made decisions or autonomously recommended actions need to have in order for operators/commanders to be held reasonably accountable for those decisions/actions? (Or, to what extent do AI systems need to be able to be interrogated?)
- Are there people beyond operators/commanders who should share responsibility for autonomous decisions or recommendations (for example, engineers, system designers etc.)?
- How does accountability for AI-driven systems fit in with existing Defence C2 hierarchies and accountability structures? Is the level of accountability held by operators/command commensurate with their seniority?

4.1.2. Education and Training

As AI-driven systems are deployed, Defence should consider reviewing education and training (E&T) of operators and commanders to ensure they have knowledge and skills commensurate with their level of responsibility and accountability for decisions made by or with AI. Systems designers and developers should also understand ethical risk and its manifestations.

Existing E&T may not adequately prepare stakeholders for the changes to decision-making brought by AI. E&T needs to be designed so that operators and commanders understand the ethical risks inherent in using AI systems, and when they will be held accountable for decisions made by or with AI systems. In order to achieve this, operators and commanders will need a deeper understanding of how AI systems work, how decisions or recommendations are generated, and in what circumstances they can be trusted.

System designers and developers should also receive E&T on the potential ethical risks arising from the use of AI in a Defence context, and the circumstances in which their ADF colleagues will be held responsible for decisions made autonomously. This knowledge

should be used from the early stages of design to ensure that systems are designed in a way that makes human accountability (where designated) fair and reasonable.

4.1.3. Data

Given both the importance of data and the significant ethical risk associated with using it, Defence should have a policy governing the collection and use of data for AI applications.

Recent advances in AI capability, particularly through deep learning, have been made possible only because of the exponential increase in the production and availability of data [9]. Any lack of rigour around the practices associated with collecting and using data presents significant ethical risks, including errors, inaccuracy, bias and discrimination. These risks are especially insidious because they are often invisible to the user of a product.

In the Defence context, poor data practices could result in unacceptable consequences, including unintentional harms. A comprehensive Defence data framework for AI applications, allowing consistency in data collection; data processing and transformation; data storage; and the use of data, would help mitigate this risk.

4.2. Findings for AI and Human-Machine Interface Developers

Findings for AI technology developers were compiled for each of the twenty ethical AI topics. They take the form of questions or observations that technology developers should consider when designing or extending the system. The highest priority findings related to issues of accountability, education and training, and data. These issues were considered to be relevant to systems beyond just AIM, and were therefore developed into the whole-of-Defence recommendations in 4.1.

Other significant findings, which are summarised in this section, relate principally to the topics of effectiveness; integration; decision support for targeting; confidence; and resilience. In aggregate, these findings encourage developers to consider whether the most efficient system or algorithm (for example, in terms of speed or accuracy), is necessarily the best in terms of providing assistance to the decision-maker. In some cases, a less efficient algorithm that is more consistent with normative decision-making may be more appropriate. In addition, there is a clear need for research on what information is necessary to make good judgements (especially where issues are complex and context is important); and how it should be rapidly conveyed (this is also critical for the ‘Agile Command and Control’ Science and Technology and Research (STaR) Shot [10]).

While these were not developed further into whole-of-Defence recommendations, many of them would nevertheless be applicable to systems beyond the type envisioned here and could be considered more broadly by Defence in due course.

In addition to the summary presented here, the complete set of findings for all twenty topics, and the analysis underpinning them, can be found in Appendix C.

Furthermore, the 7 ethical risk scenarios developed to assist Defence personnel to explore some of the key findings more deeply and develop mitigations can be found in Appendix E.

4.2.1. Effectiveness

Developers and operators of autonomous and intelligent systems should provide evidence of the effectiveness and fitness for purpose of autonomous and intelligent systems. AI systems should be deployed only after demonstrating effectiveness through experimentation, simulation, limited live trials etc. [1].

For the envisioned tactical C2 system, considering ethical risk with respect to the effectiveness topic generated the following key findings:

- Getting the optimal balance of quantity and content of information displayed to operators is both critical and complicated (especially when the system is extended to incorporate operator state monitoring). Too much information will overwhelm operators – too little has serious implications for performance and accountability. Decreasing the amount of information displayed when operators are under cognitive stress could improve performance, but only if operators still see the information they need when they need it (for example, interrupting an operator in the middle of a complex task with new information might break their concentration and negatively impact performance; but information not actively displayed could later be found to have been vital, leading to poorer outcomes.) Thus, decisions about which information to prioritise; the order it is shown in; the mode of presentation; and timing are critical. Who makes these decisions? Can features be customised for individual operators? And what does it mean for operator accountability if operators under stress make decisions on the basis of seeing less information?
- A system like the one envisioned could constrain operator choice if the number of available ‘plays’ is limited. The implications of this on operator performance and accountability should be considered – for example, is the operator accountable for a poor outcome even if they thought there was a better way to proceed but were unable to because the system didn’t present it as an option? Could this risk be mitigated by incorporating tools for on-the-fly play design (and what training would

the operator need to give them confidence to divert from system-recommended plays)?

- Interface design is vital in ensuring AI can be used ethically. Interfaces are the principal mechanism through which operators and their commanders receive and act on information. Even if the system produces the right information at the right time, if it is not accessible – for example, not clearly presented, or not seen at all –the operator will not be able to use that information to make decisions.

4.2.2. Integration

The main risks of sub-optimal system integration relate to the system not functioning as intended, or indeed not functioning at all. In addition, poor integration makes it difficult to find and repair faults quickly. In aggregate, these risks can negatively impact on performance and outcomes.

For the envisioned tactical C2 system, considering ethical risk with respect to the integration topic generated the following key findings:

- Integration should be considered as both the integration of components into a system and also the integration of human and system.
- Integration failure is a likely source of ethical risk for the envisioned system because the individual components are complex, interact in unexpected ways and were developed by different countries with differing standards, values and expectations. Mitigating the risk of integration failure is difficult because the lack of transparency surrounding the underlying components (even among FVEY partners) makes exhaustively mapping and/or testing the system extremely difficult (if not impossible). Therefore, rigorous and consistent processes to test integration and repair it quickly are critical. Standards for integration, updates and maintenance should be developed and implemented for FVEY partners (and others as necessary). Consideration should also be given to what information and training operators need to manage integration failures; and how integration issues and/or system updates are communicated to operators and command.

4.2.3. Confidence

AI systems that provide advice should also provide a level of confidence in that advice [1]. Future tactical C2 systems could usefully include confidence measures for some behaviours and decisions to improve explainability and trust. This would require investigation of what type of measures would be useful; when they would help; and what training operators/command need to interpret them accurately. Other areas of Defence

already use confidence measures, and future tactical C2 system developers could draw on these in developing options.

4.2.4. Resilience

System resilience refers to a human-AI system exhibiting the ability to foresee, contain, and recover from anomalous situations. Resilience is the combination of the system's ability to prevent something from happening, to prevent something from becoming worse, and/or to recover from anomalous situations [1].

For the envisioned system, considering ethical risk with respect to the resilience topic generated the following key findings:

- Resilience of a system like the one under consideration requires both physical and virtual components to be robust against a large number of threats. The large number of components (and the resulting potential for data latency) provides multiple points of entry and increases risk. Identifying when AI-enabled tactical C2 systems have been breached will be difficult given the number and complexity of components.
- Redundancy, separation and good maintenance regimes could help build system resilience. The envisioned system may lend itself to the application of antifragility theory.
- Implementing resilience strategies is made more complex given the number of partners who have collaborated to develop systems such as AIM – arrangements between the FVEY to coordinate (or work together on) system monitoring and maintenance would help.
- Building the resilience of data transportation and storage systems is also key.

4.2.5. Decision Support for Targeting

An authority pathway is an AI tool designed to make sure operators of a system have completed specific required steps before executing an action. Authority pathways are designed to help tactical decision-makers make more ethical and correct judgements. This might be through programming to abide by international law, integrate multiple sources of data, or present alternative scenarios [1].

Our exemplar system, AIM, incorporates several authority pathway tools, which can ensure operators abide by relevant policies when making decisions. For example, COMPACT included implemented policies for air vehicle and airspace de-confliction.

Considering ethical risk with respect to the authority pathway topic generated the following key findings:

- The prevalence, complexity and types of decisions that authority pathway tools are designed to assist with makes it particularly important that they are accurate and effective, and that their limits are well understood by those that use them.
- Important ethical risks surface around the level of nuance that authority pathway tools can employ in interpreting policies; how conflicts between polices are resolved; whether operators are aware of negotiations; and whether human operators (or command) should have the ability to over-ride adherence to a policy or the outcome of a policy negotiation.

4.3. Findings for Application of *A Method for Ethical AI in Defence*

A Method for Ethical AI in Defence offered a comprehensive and effective framework for thinking about ethical risk in a Defence context. Application of the method to the envisioned tactical C2 system (notwithstanding resource and time constraints) yielded suggestions for a significant body of further work for all three stakeholder groups. In order to take forward these streams of work, appropriate resourcing will be critical.

The observations in this section are offered as suggestions for iterating and improving the method, and in case they are useful for other practitioners seeking to apply it in their own case studies. The three observations below were considered the most significant learnings, but are not exhaustive. A number of other findings are available in Appendix B.

4.3.1. Development of additional tools

The number of facets and topics is large, and not all facets will be relevant (or as relevant) to all systems. Resources and time available to study a system will differ depending on context. In addition, different risks are best addressed by different people, and the risks directly relevant to ethicists will be a sub-set of all the risks identified for a system. Therefore, a tool to help practitioners identify the most relevant facets for their system would assist in allocating time and resources efficiently. This could be achieved using a risk assessment that considers each of the topics and assigns them a risk rating for the particular system under examination (for example, a rating of between 1 to 3, where 1 is not relevant or minimal risk present; 2 is medium risk present; and 3 is significant risk that requires addressing before deployment of the system). Resources could be allocated to analysing category 3 risks (and category 2 where possible or necessary).

Separately, more detail on the three tools provided in Section 4 of *A Method for Ethical AI in Defence* would help practitioners to apply them. A separate guidebook for applying each tool could be developed with examples for each step.

4.3.2. Terms and definitions

Many of the terms used in the method are familiar within the AI field, but many are contested, and several are used in a way that is not typical (for example, governance and accountability). To ensure clarity, a glossary of terms would be helpful. This should define terms such as ‘ethical AI’, ‘ethical risk’ and also each of the facets and topics. This would help to make it clear where the boundaries lie, and where usage is unique to the Defence context.

4.3.3. Linking facets and tools

It would be helpful to more directly link the ethical AI facets and topics to the three tools offered to apply them (the *Ethical AI for Defence Checklist,* the *Ethical AI Risk Matrix* and the *Legal and Ethical Assurance Program Plan*). For example, the facets are a useful way to draw out how the system in question integrates with human operators, and what existing ethical risks and mitigations are in place (which maps to one of the steps in the *Ethical AI for Defence Checklist*). This link could be formalised, perhaps through summarising the key considerations for each topic into three or four questions. These questions could be used as the basis for interrogating the system when completing the *Ethical AI for Defence Checklist*.

5. ACKNOWLEDGEMENTS

Many thanks to both the Department of the Prime Minister and Cabinet and the Defence Science and Technology Group for facilitating and supporting the secondment which allowed this work to take place.

The author would like to acknowledge the contributions of DSTG colleagues who generously engaged in useful discussions that have influenced the content in this report. In particular, many thanks to: Marcin Nowina-Krowicki, Dr Greg O'Keefe, Dr Steven Wark and Dr James Brooks.

6. REFERENCES

[1] K. Devitt, M. Gan, J. Scholz and R. Bolia, “A Method for Ethical AI in Defence,” Defence Science and Technology Group, 2021.

[2] K. Sayler, “Artificial Intelligence and National Security, Version 10 (R45178),” Congressional Research Service, 2020.

[3] Department of Defence, “2020 Defence Strategic Update,” 2020. [Online]. Available: https://www1.defence.gov.au/strategy-policy/strategic-update-2020.

[4] Department of Defence, “2020 Force Structure Plan,” 2020. [Online]. Available: https://www1.defence.gov.au/strategy-policy/strategic-update-2020.

[5] Department of Industry, Science, Energy and Resources, “AI Ethics Framework,” 23 1 2019. [Online]. Available: https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework. [Accessed 2021].

[6] Department of Defence, “TTCP Autonomy Strategic Challenge,” Unpublished, 2018.

[7] Technical Cooperation Program, “Autonomy Strategic Challenge (ASC) Allied IMPACT Final Report,” TTCP Technical Report, Canberra, 2020.

[8] C. Shanahan, S. Wark, S. Hoskings and K. Devitt, “Utilising real-time measures of operator state to inform adaptive autonomy within an ethical framework,” Unpublished (work ongoing), 2019.

[9] G. Moy, S. Shekh, M. Oxenham and S. Ellis-Steinborner, “Recent Advances in Artificial Intelligence and their Impact on Defence,” Department of Defence, 2020.

[10] Department of Defence, “Science Technology and Research (STaR) Shots,” August 2020. [Online]. Available: https://www.dst.defence.gov.au/sites/default/files/basic_pages/documents/Booklet%20-%20STaR%20Shots.pdf. [Accessed 26 2 2021].

[11] M. C. Elish, “Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction,” *Engaging Science, Technology, and Society,* vol. 5, pp. 40-60., 2019.

[12] Department of Defence, “Australia’s System of Control and applications for Autonomous Weapon Systems,” Presented at the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, Geneva, 2019b.

[13] D. K. Ahner, “Test and Evaluation of Autonomous Systems,” Paper presented at the 33rd International Test and Evaluation Symposium, 2016 .

[14] UK House of Lords, “AI in the UK: ready, willing and able?,” 2017 - 2019. [Online]. Available: https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10007.htm#_idTextAnchor025.

[15] A. Bhaskara, M. Skinner and S. Loft, “Agent Transparency: A Review of Current Theory and Evidence,” *IEEE Transactions on Human-Machine Systems,* pp. 1 - 10, 2020.

[16] ATARC AI Ethics and Responsible AI Working Group, “Information Technology – Artificial Intelligence – Machine Learning (ML) model transparency,” Unpublished - seeking comment, 2020. [Online]. Available: https://atarc.org/project/information-technology-artificial-intelligence-machine-learning-ml-model-transparency/.

[17] R. Ratwani, “Human Factors in Machine Learning and Artifical Intelligence,” National Center for Human Factors in Healthcare, 2020.

[18] W. Xu, “User-centered design (IV): Human-centered artificial intelligence,” 2019.

[19] N. Taleb, Antifragile: Things That Gain From Disorder, 1st ed. ed., New York: Random House, 2012.

[20] D. Russo and P. Ciancarini, “A Proposal for an Antifragile Software Manifesto,” *Procedia Computer Science,* vol. 83, pp. 982-987, 2016.

[21] S. Thiebes, S. Lins and A. Sunyaev, “Trustworthy artificial intelligence,” Electron Markets, 2020.

[22] N. E, “Third Innocent Black Man to be MIsidentified by Facial Recognition Software Sues Police Department and Prosecutor for False Arrest and Imprisonment,” 31 12 2020. [Online]. Available: https://lawandcrime.com/civil-rights/third-innocent-black-man-to-be-misidentified-by-facial-recognition-software-sues-police-department-and-prosecutor-for-false-arrest-and-imprisonment/. [Accessed 23 1 2021].

[23] G. P, N. M and H. K, “Face Recognition Vendor Test (FRVT): Part 3: Demographic Effects NISTIR 8280,” US Department of Commerce, 2019.

[24] G. Stevens, “Integrating the Supply Chain,” *International Journal of Physical Distribution & Materials Management,* vol. 19, no. 8, pp. 3-8, 1989.

[25] International Committee of the Red Cross Expert Meeting, “Autonomous weapon systems: technical, military, legal and humanitarian aspects,” Bulletin of the Atomic Scientists, 10 1 2014. [Online]. Available:

https://www.aph.gov.au/DocumentStore.ashx?id=b64a259c-b9ca-4be1-b5ce-a16dde8adda0&subId=303585.

[26] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Shieber, J. Waldo, D. Weinberger and A. Wood, "Accountability of AI under the law: The Role of Explanation," 2017. [Online]. Available: arXiv. https://arxiv.org/abs/1711.01134v1.

[27] D. &. P. F. Citron, "The scored society: due process for automated predictions," *Washington Law Review,* vol. 89, no. 1, 2014.

[28] M. &. K. B.-J. Hildebrandt, "The Challenges of Ambient Law and Legal Protection in the Profling Era," *The Modern Law Review,* vol. 73, no. 3, p. 428–460, 2010.

[29] T. Zarsky, "Transparent predictions," *University of Illinois Law Review,* no. 4, p. 1503–1570, 2013.

[30] D. Leake, Evaluating Explanations: A Content Theory, New York: Psychology Press, 1992.

[31] International Committee of the Red Cross, "Artificial intelligence and machine learning in armed conflict: A human-centred approach," 2019. [Online]. Available: https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach.

[32] S. Farthing, J. Howell, K. Lecchi, Z. Paleologos, P. Saintilan and E. Santow, "Human Rights and Technology: Discussion Paper," Australian Human Rights Commission, Sydney, 2019.

[33] L. Floridi, "Faultless responsibility: on the nature and allocation of moral responsibility for distributed moral actions," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 374, no. 2083, 2016.

[34] K. Leetaru, "No Facebook, your users don't understand how you use their data," 2018. [Online]. Available: https://www.forbes.com/sites/kalevleetaru/2018/12/19/no-facebook-your-users-dont-understand-how-you-use-their-data/. [Accessed 02 02 2021].

[35] E. Nanos, "Third innocent black man to be misidentified by facial recognition software sues police department and prosecutor for false arrest and imprisonment," Law and Crime, 2020.

[36] Australian Bureau of Statistics, "The ABS Data Quality Framework," Commonwealth of Australia, [Online]. Available: https://www.abs.gov.au/websitedbs/D3310114.nsf//home/Quality:+The+ABS+Data+Quality+Framework.

[37] Australian Bureau of Statistics, “1520.0 – ABS Data Quality Framework,” 2009. [Online]. Available: https://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/1520.0Main%20Features1May%202009?opendocument&tabname=Summary&prodno=1520.0&issue=May%202009&num=&view=. [Accessed 23 1 2021].

[38] H. A. Abbass, “Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust,” *Cognitive Computation,* vol. 11, no. 2, pp. 159-171, 2019 .

[39] K. Crawford and V. Joler, “Anatomy of an AI system,” 2018. [Online]. Available: https://anatomyof.ai/. [Accessed 23 01 2021].

[40] China State Council, “A Next Generation Artificial Intelligence Development Plan, translated by New America,” 2017. [Online]. Available: https://www.newamerica.org/documents/1959/translation-fulltext-8.1.17.pdf.

[41] Department of Defense, “Summary of the 2018 National Defense Strategy,” 2018. [Online]. Available: https://dod.defense.gov/Portals/1/Documents/pubs/2018-National-Defense-Strategy-Summary.pdf.

APPENDIX A. APPLYING *A* METHOD FOR ETHICAL AI IN DEFENCE TO THE ENVISIONED TACTICAL C2 SYSTEM

A modified version of the E*thical AI for Defence Checklist* from *A Method for Ethical AI in Defence* [1] was used to analyse the envisioned tactical C2 system:

- The military context and types of decisions supported by the AI was examined through a data collection process including a document review, interviews with subject matter experts, and a demonstration of AIM at DSTG Edinburgh.
- Potential ethical risks were identified by examining the envisioned system using each of the twenty topics from *A Method for Ethical AI in Defence* (Appendix C).
- Where the potential for significant ethical risk was identified, hypothetical scenarios (Appendix E) were generated to help DSTG personnel further explore the nature of the risk and identify suitable mitigation strategies for incorporation into future system design.
- The analysis in the preceding steps allowed the generation of key observations and recommendations for three stakeholder groups, namely: whole-of-Defence; AI technology developers; and practitioners seeking to apply or iterate on *A Method for Ethical AI in Defence* (Section 4 and Appendices B, C and D).

These steps were operationalised as follows:

1. Data collection:
    a. Review of key documents, including *A Method for Ethical AI in Defence* [1], the Autonomy Strategic Challenge Allied IMPACT Final Report [7], and the training manual for operators at AW18.
    b. Briefings by DSTG experts on AIM.
    c. Demonstration of AIM at DSTG Edinburgh, Adelaide.
2. Application of the 20 ethical AI topics to the envisioned tactical C2 system:
    a. The envisioned system was examined through the lens of each of the 20 topics proposed in *A* Method for Ethical AI in Defence. Appendix C contains the findings of this process.
    b. in addition, the challenges faced by AIM operators during AW18 (as recorded in the final report [7]) were tabulated and assigned a facet. This helped to build a picture of how the current system design might give rise to ethical risk. Findings

from this exercise were built into the examination of the twenty topics (Appendix C).

3. Development of ethical risk scenarios:

   a. Areas of significant ethical risk were identified using the data collected in steps 1 and 2.

   b. Scenarios were designed to highlight the potential impact of these significant risks. Subject matter expertise on the scenarios was provided by DSTG personnel, including Dr Christopher Shanahan, Dr Kate Devitt, Dr Steven Wark, Marcin Nowina-Krowicki, Dr Greg O'Keefe and Dr James Brooks.

   c. Testing and evaluating the scenarios was beyond the scope of this report, but there is the potential for one or more of them to be used in future training for C2 system operators.

   Seven scenarios have been developed (Appendix E). A balance was sought between scenarios that would apply specifically to the exemplar (AIM) and to the envisioned system, and those that might be applicable more broadly across Defence.

4. Key observations

   Key observations as a result of steps 1-3 were recorded for three groups of stakeholders:

   a. *Defence leadership*: a number of the identified ethical risks were deemed to have the potential to apply to all AI systems developed or deployed by Defence (not just the envisioned system). These risks would benefit from further investigation and could offer the opportunity for Defence to develop Defence-wide mitigation strategies. Recommendations for such further work is included in this section.

   b. *AI technology developers*: a number of the identified ethical risks were deemed to be more relevant in the context of developing a system like AIM. These observations and recommendations arising from them have been recorded in this section.

   c. *Practitioners seeking to improve or apply A Method for Ethical AI in Defence*: through the process of applying *A Method for Ethical AI in Defence* to the envisioned system, a number of observations about applying the method and suggestions for further work to iterate it have been collected and recorded in this section.

APPENDIX B. KEY FINDINGS FROM APPLYING A METHOD FOR ETHICAL AI IN DEFENCE

This report represents the first applied case study of *A Method for Ethical AI in Defence*. Overall, the method provided a comprehensive and effective framework for thinking about ethical risk in a Defence context. The use of the individual topics in teasing out different angles of ethical risk was particularly helpful.

A strength of the method is its adaptability and flexibility, which will make it applicable to a broad range of systems and contexts. When writing this report, the way in which the method was used changed considerably as the system was investigated and issues became clearer. For example, it was initially thought that the development of the ethical risk scenarios would help clarify the most relevant facets for study. However, in reality, the risk scenarios identified issues that cut across topics. This created a reinforcing feedback loop – issues emerging from the risk scenarios led to consideration of all the topics, which allowed additional risks to emerge, which were then used to further inform the ethical risk scenarios. This flexibility means that practitioners using the method should feel comfortable selecting and applying the most relevant parts of the method given their particular context.

The comprehensiveness and level of consultation possible for the envisioned system case study was nevertheless limited by the length of time available (around eight weeks) and the human resources dedicated to the task (one person). Many of the topics warrant further investigation, and suggestions for further work throughout the report add up to the potential for a significant program of future research. Defence could consider allocating more resources to this area, both in order to produce comprehensive ethical risk assessments of systems in development, and to begin broader work on issues like frameworks for accountability, education and training, and data (as proposed in 4.1).

The following observations are offered as suggestions for iterating and improving the method, and in case they are useful for other practitioners seeking to apply it in their own case studies:

- The number of facets and topics is large, and not all facets will be relevant (or as relevant) to all systems. Resources and time available to study a system will differ depending on context. In addition, different risks are best addressed by different people, and the risks directly relevant to ethicists will be a sub-set of all the risks identified for a system. Therefore, a tool to help practitioners identify the most relevant facets for their system would assist in allocating time and resources

efficiently. This could be achieved using a risk assessment which considers each of the topics and assigns them a risk rating for the particular system under examination (for example, a rating of between 1 to 3, where 1 is not relevant or minimal risk present; 2 is medium risk present; 3 is significant risk that requires addressing before deployment of the system). Resources could be allocated to analysing category 3 risks (and category 2 where possible or necessary).

- Terms and definitions: many of the terms used in the method are familiar within the AI field, but many are contested, and several are used in a way that is not typical (for example, governance and accountability). To ensure clarity, a glossary of terms would be helpful. This should define terms such as 'ethical AI', 'ethical risk' and also each of the names of the facets and topics. This would help to make it clear where the boundaries lie, and where usage is unique to the Defence context.
- It would be helpful to more directly link the ethical AI facets and topics to the three tools offered to apply them (the *Ethical AI for Defence Checklist,* the *Ethical AI Risk Matrix* and the *Legal and Ethical Assurance Program Plan*). For example, the facets are a useful way to draw out how the system in question integrates with human operators, and what existing ethical risks and mitigations are in place (which maps to one of the steps in the *Ethical AI for Defence Checklist*). This link could be formalised, perhaps through summarising the key considerations for each topic into three or four questions. These questions could be used as the basis for interrogating the system when completing the *Ethical AI for Defence Checklist*.
- More detail on the three tools provided in Section 4 of *A Method for Ethical AI in Defence* would help practitioners to apply them. A separate guidebook for applying each tool could be developed with examples for each step.
- The facets and topics are comprehensive, but it is not clear if they are supposed to be exhaustive. If so, some changes to the scope could be considered:
    - Under the education topic, the needs of other stakeholders – for example, operators, systems developers and engineers – could be considered in addition to command.
    - The Law facet could be expanded beyond consideration of protected symbols and surrender and de-escalation to consider more broadly how and when to incorporate law into algorithms and models.
    - The accountability topic fits under Traceability, and thus refers quite specifically to ensuring that appropriate records about decisions taken using AI systems are retained. It would be useful if this notion were expanded – or perhaps an

additional topic added elsewhere – to incorporate a broader notion of accountability (i.e. where accountability for decisions and outcomes sits).

- Data (its collection, transformation and use) underpins the development of AI. While different aspects of data could be considered in several of the topics (such as Data subjects, Confidence, Resilience and Supply chains), a facet or topic dedicated to a more holistic treatment of data could be useful.
- Security could currently be considered under Resilience and Misuse and Risks. Perhaps Security should become a facet encompassing Resilience and Misuse and Risks instead.
- A topic for contestability would be useful, perhaps under Trust.
- Interfaces was considered in Effectiveness, but could be added as a separate topic.
- The allocation of topics to facets could be adjusted to ensure topics with strong links sit together.
- It would be helpful to link the two-component model of trust presented in the Trust facet more closely to the topics. The two components of trust according to the model are competence and integrity. Potentially, these could be used to organise the topics. So, for example, competence could cover safety, test & evaluation and authority pathway; whereas integrity could cover sovereign capability, supply chain, misuse and risks, and data subjects.

APPENDIX C. APPLYING THE FIVE FACETS OF ETHICAL AI TO THE ENVISIONED TACTICAL C2 SYSTEM

This part of the report contains an assessment of the envisioned system against each of the 20 topics of AI identified in *A Method for Ethical AI in Defence* (Figure 3). Reading each section alongside the corresponding section in *A Method for Ethical AI in Defence* will present the most complete picture of the findings.

Assessment of each topic was informed by briefings provided by DSTG staff; analysis of relevant documents provided by DSTG (including challenges identified by operators during AW18 and recorded in the *Autonomy Strategic Challenge (ASC) Allied IMPACT Final Report* [7]), and other research on autonomous systems as referenced.

As a result of the analysis presented here, seven pressing issues were chosen to use as the basis for ethical risk scenarios (Appendix E).

Facets of ***Ethical AI for Defence***	**Topics emerging from the workshop**
Responsibility Who is responsible for AI?	Education and Command
Governance How is AI controlled?	Effectiveness, Integration, Transparency, Human Factors, Scope, Confidence and Resilience,
Trust How can AI be trusted?	Sovereign Capability, Safety, Supply Chain, Test & Evaluation, Misuse and Risks, Authority Pathway and Data Subjects
Law How can AI be used lawfully?	Protected Symbols and Surrender, and De-escalation
Traceability How are the actions of AI recorded?	Explainability and Accountability

Figure 3 Facets and topics of ethical AI in Defence [1]

C.1. Responsibility

The responsibility facet addresses the question of who is responsible for AI, noting that ‘it may be unclear who is responsible for decisions or actions in both combat and non-combat operations involving AI’ [1].

In particular, *A Method for Ethical AI in Defence* focuses on the implications of AI-driven systems (especially those utilising machine learning) for commanders, given their authority and accountability for actions or inaction.

C.1.1. Education

A Method for Ethical AI in Defence notes that Defence should ensure commanders sufficiently understand the behaviour of AI and the potential consequences of its operation through appropriate education and training (E&T).

This is important because commanders will potentially be held responsible for decisions made by or with an AI-driven system like AIM. To be held fairly accountable for decisions, decision-makers need to have sufficient knowledge and understanding of the implications of those decisions. Thus, E&T should ensure commanders in charge of the envisioned tactical C2 system have the knowledge they need about the system to make informed decisions about the use of that system. The level of knowledge and understanding attained should be commensurate with the commander's level of responsibility and accountability.

In order to achieve an appropriate level of knowledge about autonomous systems, E&T should focus not only on how to operate the system, but also on how the system works. This approach acknowledges that autonomous systems and humans make decisions differently. Commanders need to understand those differences, including the strengths and limitations of autonomous decision-making, in order to make good decisions about the use of autonomous systems.

One way of establishing where E&T might be necessary is to ask how deploying AIM might change the relationship between a commander, their decision-making and their accountability for those decisions.

For the envisioned system, the extent to which a commander's experience allows them to make judgements on the following characteristics of AI systems will change how and why they make decisions, and could therefore be areas where E&T might usefully focus:

- How and why a particular recommendation has been produced (a recommendation developed by a chain of humans is likely to be more intelligible and familiar than a decision produced by an AI-enabled system).
- How much to trust decisions/recommendations made by the system. An inclination to trust the system too much or too little would compromise commanders' decision-making ability.

  - Automation bias (the tendency to trust the results of a machine simply because they are machine generated rather than as a result of any particular evidence to support the results) could cause a commander to trust the envisioned system too much.
  - Conversely, unfamiliarity working with AI could result in a harmful lack of trust, for example manifesting through a lack of confidence that the system will perform as expected and a tendency to try and work around the system rather than with it.
- How and when to interrogate the reasons for a recommendation (for example, a commander who is unfamiliar with autonomous systems might be unaware of how to interrogate the system or what the information presented means).
- How an AI-enabled system might impose limitations on decision-making – for example, in some cases, a commander may be prevented from making the choice they want because constraints deliberately put in place to manage one type of scenario have unintended consequences in other situations.
- Whether commanders are responsible for the consequences of the system's autonomous functions, even if they are not actively involved in making the decision (or do not have knowledge of it at all).

While *A Method for Ethical AI in Defence* refers predominantly to the need for educating command, due consideration should be given to the need for E&T to extend to other stakeholders. For example, system operators should also receive similar E&T, since they may also be held accountable for certain decisions made by/with the AI-enabled system (depending on the accountability model used); and because understanding the system will enhance their effectiveness.

The importance of achieving these goals will extend beyond AIM to other AI-driven systems. Section 4.1.2 proposes that Defence develop a new E&T framework for stakeholders (especially commanders and operators) involved with AI-driven systems like the one we envision.

C.1.2. Command

The Command topic asks whether the use of AI in military operations changes commanders' responsibility, as well as whether others should share responsibility for AI (for example, programmers, system designers or system operators) [1]. A critical part of resolving this question involves deciding which decisions can appropriately be made by autonomously, and which decisions should require human authorisation or intervention.

The main ethical risk implied by the Command topic relates to the possibility of human personnel (in this case, either the AIM operator or command) becoming moral crumple zones when things go wrong. The idea of moral crumple zones was introduced by Elish [11] to describe how responsibility for an action may be misattributed to a human actor who had limited control over the behaviour of an automated or autonomous system.

The locus of responsibility for decisions made by/with AIM was not directly considered in the development of AIM (the system was designed to demonstrate force maximisation rather than to assess the ethical risks inherent in assigning certain types of decisions to AIM). But questions of accountability are among the most important to resolve before deploying any AI-driven system. Indeed, being clear about accountability from the very start of the design process can help developers design a system that aids decision-makers in their tasks and facilitates reasonable and fair accountability.

Therefore, AI technology developers should carefully consider the types of decision the system can make autonomously and who would ultimately be held responsible for those decisions (see also C.2.5). A useful place to start might be to examine existing C2 models within Defence and whether some aspects of those might be applicable to autonomous systems like the one envisioned here. For example, imagine an autonomous function taking the place of a junior ADF personnel. This would leave the operator in a command role. In this situation, what decisions would the system be able to take without authorisation from the commanding officer? Would the commander nevertheless be responsible for those decisions? How does this compare to how the system is designed currently?

Potentially, this thought experiment could provide a useful starting point for deciding which decisions the envisioned system should be able to take autonomously and where accountability for those decisions should lie.

It could also help with consideration of whether the seniority of the operator is appropriate given their anticipated level of accountability. An operator of the future tactical C2 system will replace a comparatively large number of human personnel (around 70 people if the operator is managing 17 assets), and is therefore effectively in charge of the decision-making load of that same number of replaced humans. Would the current level of seniority of an operator be commensurate with that level of decision-making? Perhaps so, if the decisions being made autonomously were very low-level. However, if more complex decisions are being taken autonomously, or if there is the potential for autonomous actions to interact in ways that produce higher-order consequences, the operator’s seniority might need to be elevated in future systems. Alternatively, over time

and as AI systems are increasingly deployed in Defence, job descriptions may need to be updated to reflect changes in what is expected at different levels of the organisation.

Other characteristics that should be considered when thinking about the Command topic with respect to the envisioned system include:

- Are there limits to how many UxVs an operator and their commander can reasonably manage and be held accountable for?
- If the system constrains choice, is command responsible for the outcome? This issue is explored more in ethical risk scenario E.3
- If the envisioned system assesses that an operator is under a high cognitive load and takes control of more functions, can the operator still be held accountable for poor outcomes? This issue is explored more in ethical risk scenario E.6
- How does the acceleration of decision timeframes enabled by a system like the one envisioned here affect decision-making and thus reasonable accountability?
- Should the system operator/command be able to over-ride the system in all cases?

These questions are not just relevant to the envisioned system. Defence will need clear and consistent guidelines on these issues for all AI-driven systems. For that reason, Section 4.1 (and Appendix D.1) proposes that Defence should develop a whole-of-Defence accountability framework for AI systems. Noting that accountability for decisions made by AI is a contested field, such a framework could carefully consider what sort of accountability is reasonable for command (and others) and what accountability hierarchy could be most appropriate (for example, one option is a hierarchy that spreads ethical and legal responsibility across all nodes/agents in the human-AI network relevant to a decision; another option might map more closely onto the existing hierarchical command and control model in place at Defence).

C.2. Governance

The governance facet asks how AI is controlled. *A Method for Ethical AI in Defence* notes that in an Australian context, autonomous weapons systems are governed by the existing control hierarchy for the use of military force. This control hierarchy is set out in the non-paper[10] *Australia's System of Control and applications for Autonomous Weapons*

[10] A non-paper is an informal discussion paper, usually presented to the UN. It allows countries to put forward ideas and views without committing to particular agendas. Non-papers are an important way to share information and assess other countries' views on topics.

Systems [12]. Figure 4 summarises the types of control present in each level of the hierarchy.

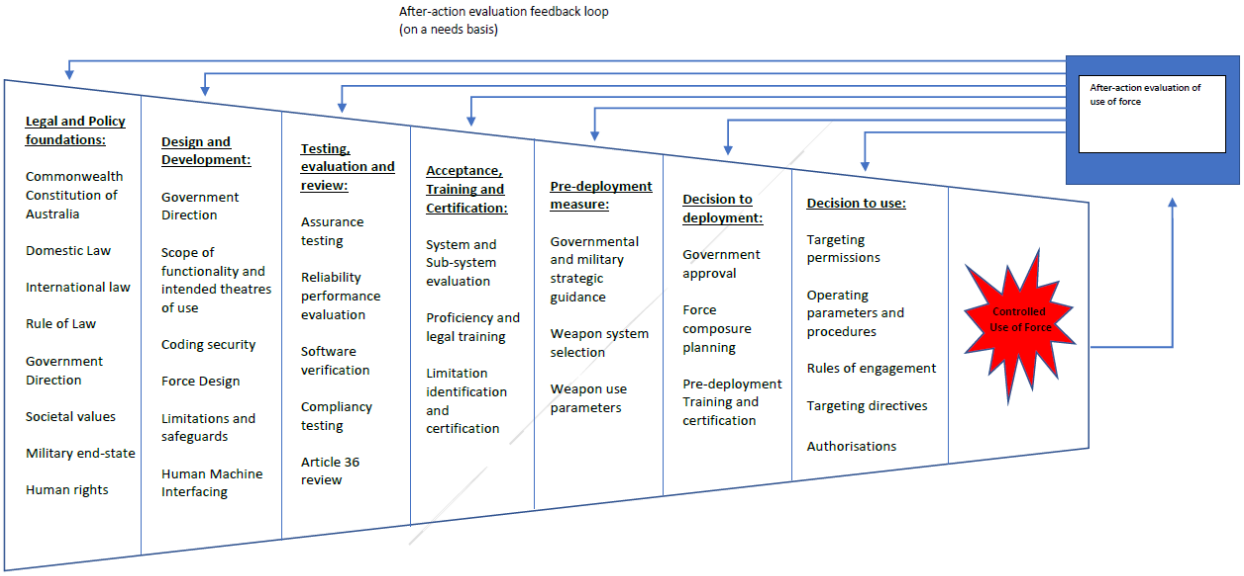


Figure 4 Australia's systems of control for autonomous weapons

A Method for Ethical AI in Defence proposes a different set of criteria for assessing governance, namely: effectiveness, integration, transparency, human factors, scope, confidence and resilience. Many of the challenges identified by operators using AIM during AW18 [7] could be allocated to these ethical risk topics.

C.2.1. Effectiveness

A Method for Ethical AI in Defence uses Ahner's definition of effectiveness, which is the ability of an AI system to either make or support human ethical decision-making in environments with varying levels of risk and decision requirements. Effectiveness should be demonstrated through experimentation, simulation and limited live trials etc. [13]. The inclusion of the word 'ethical' before decision-making moves the idea of effectiveness beyond functional effectiveness (which is whether the systems works in the manner intended). Even so, functional and ethical effectiveness are closely linked. Functional effectiveness could be considered a necessary condition for ethical effectiveness, although it is not a sufficient condition (in other words, a system needs to be functional in order to be ethically effective – but just because a system is functionally effective does not make it ethically effective).

This distinction is important because many of the challenges identified by AIM operators in AW18 related to technical issues that diminished functional effectiveness (for example, functional effectiveness was at times affected by information not being clearly conveyed

or understood by the operator) [7]. By definition, this creates ethical risk because functional effectiveness is a necessary condition for ethical effectiveness.

Defence is well-versed in solving technical issues such as those identified by the operators in AW18, and individual problems would no doubt be resolved in a subsequent iteration of the system. However, it is useful to ask whether any overarching issues were present; and whether ethical risk would remain for systems such as AIM even if it were operationally functioning exactly as intended?

One overarching issue relates to the importance of the user interface design in facilitating ethical decision-making by future tactical C2 system operators. Many of the problems identified by operators related to interface design – for example, the absence of alerts (which led to operators missing critical information at times) [7]. It's worth examining briefly why interfaces are so important and why considerable thought and effort needs to go into making sure they are fit for purpose.

Interfaces are the principal mechanism through which operators and their commanders receive and act on information. Even if the system produces the right information at the right time, if it's not accessible – for example, not clearly presented, or not seen at all, the operator will not be able to use that information to make decisions.

The immediate consequence of this is poorer decision-making. However, interface issues also have broader impacts on responsibility and accountability. For example, AIM currently lacks a mechanism to acknowledge receipt of a message and confirm actions taken. If logs were examined, this would make it extremely hard to know whether an operator had seen a message at all, or whether they had perhaps been too busy to respond. In this situation, it would be unreasonable to assume operator accountability for an action either taken or not taken.

Another over-arching issue for the envisioned system is finding a balance between providing enough information to help decision-makers without overwhelming them with too much information or distracting them when information arrives. This challenge relates to decisions about which information is important, how to prioritise that information, and when/how it should be delivered (see also C.2.2).

This issue is best illustrated using an envisioned system in which it is assumed that providing more information (or all available information) to all operators does not necessarily optimise performance. This model recognises that operators using systems like the one envisioned are subject to significant time constraints and must juggle competing high-priority tasks. Too much information, especially when an operator is

already busy, can distract, overload and/or overwhelm the operator, leading to poorer outcomes.

Nevertheless, designing a system that solves this problem by seeking to provide operators with only the information they need and can cognitively handle in that moment raises a number of interesting ethical risks.

- How is information prioritised? In other words, how (and by whom) are decisions made about which information is important, and thus, which information should be conveyed, in what order, and in what format (if the system is multi-modal)?
- Does the answer to this question change for each individual operator? Do individual operators have their own preferences and ideas about what is important, and should future systems be able to adapt to fulfil these preferences? DSTG is working on technology that will adapt information presentation according to the cognitive load of the user, but is not currently considering the extent to which individual users will have different ideas about which information is most important and should be displayed).
- Future systems will possibly incorporate three design features to help operators manage a high cognitive load: taking on more functions autonomously; providing less detail on incoming information (with additional detail available on-demand); and/or holding off on notifying an operator of certain information to avoid distracting them. How will these features affect operator accountability? For example, at times of high cognitive load, operators are unlikely to have time to ask for more detail or interrogate recommendations/autonomous actions in the same way. Is the operator accountable if important information was available but not accessed by the operator? If future systems assess that an operator is bearing a high cognitive load and performs more tasks autonomously, who is responsible for possible poor outcomes – the operator who was stressed or someone else? How is a ‘cognitively stressed’ operator perceived by their colleagues? Is there a difference between an operator who is cognitively overloaded because they have a high tasking load and one who is tired and should they be treated differently? Some of these questions are explored more in ethical risk scenario E.6.

These issues require careful thought. When considering them, the following practical suggestions might be useful:

- In order to address questions about accountability (which recur throughout many sections of this report), Section 4.1.1 and Appendix D.1 suggest that Defence develop an accountability framework for the use of AI-driven systems to provide

clear guidance on what decisions can appropriately be made autonomously and how accountability is ascribed in situations where the system does so.

- System designers and developers should be trained to consider ethical risk as early as possible in the design and development of the system. This would allow the system to be built around the mitigation of ethical risk, recognising that it is far easier to incorporate this at the design stage rather than retro-fit the system after it has been built. To this end, Section 4.1.2 recommends a new training and education framework, which would include ethical risk training for system developers.
- Seeking different perspectives about potential ethical risk early in the development of a system is critical. Conversations about ethical risk between ethicists, designers, potential clients, commanders and other stakeholders should take place early on in the design process.
- Tools from decision analysis theory might be useful in addressing questions about how to prioritise information.

One area where ethical risk might remain even if the envisioned system were functioning exactly as intended is in the case of limitations being imposed on operators/command by way of constraints built into the system. For example, AIM operates on a play-calling system whereby a number of customisable – but pre-programmed – plays are available for the operator to select from based on requirements. The number of available plays is finite, and there is probably a limit to the number of plays an operator can reasonably remember and effectively utilise. However, in some circumstances, there may not be a play that corresponds well to the situation being faced. In such circumstances, a human without the AI-enabled technology would probably improvise and develop the best solution given the resources available. However, an operator using a system such as AIM may be constrained by the plays available. Designing additional plays on the fly would likely take too much time and technical expertise to be feasible. The operator may be forced to choosing a play that is not optimal, even though a human might have thought of a better solution. The ethical risk in this situation lies in considering whether system constraints, either from finite plays or other issues, might ever be dangerous – and also how being forced to choose a less optimal course of action influences accountability (for example, is the operator still accountable for a poor outcome even if their choice was constrained?) Some of these issues are explored in ethical risk scenario E.3.

Another issue to consider in terms of effectiveness is whether operators could continue to manage missions if the AI-enabled system failed. For example, if there were a cyber-attack on the system and it went offline, to what extent could the operators continue to function? There is probably a point at which a system like the one we envision might be

considered to be *too* effective. In other words, how acceptable is the risk that humans would not be able to replace the AI in an emergency, either due to a lack of expertise or training, or simply because the workload is too high? The answer to this question will change depending on the system, but strong strategies to ensure redundancy and reduce the risk are critical. These issues are explored more in ethical risk scenario E.4.

A number of other issues related to effectiveness are covered in other sections. Effectiveness as a result of how well the system components are integrated is covered in C.2.2. Effectiveness as a product of data quality is covered in C.3.2.

C.2.2. Integration

Integration failure is a key source of ethical risk for complex systems like AIM, and arises when individual components are unable to properly communicate or function as part of the whole system. The main risks relate to the system not functioning as intended, or indeed not functioning at all. In addition, poor integration makes it difficult to find and repair faults quickly. In aggregate, these risks can negatively impact on performance and outcomes.

Integration issues are likely to be compounded when systems are developed by multiple partners. Assuming that FVEY partners (and possibly others) are indeed likely to cooperate more – not less – on the development of AI-driven defence systems (given the complexity of the systems, the resources required to develop them, and our common commitment to interoperability) – developing robust processes for avoiding system failure due to integration issues is critical.

In this section, the link between integration and ethical risk is considered from two perspectives – integration between components of the system, and integration between the system and the user[11].

In terms of the integration between components of the system, AIM presents an excellent test-case. With complex components designed and built by four countries, using different platforms and architectures, integration of all the components was difficult (indeed, integrating a system as complex as AIM would be unlikely to be seamless even if all the components were built by the same team.)

Adding to the complexity, there are at least two layers of component interaction that present the risk of failure. The first concerns the ability of different platforms to communicate without issue when necessary. One example of when this did not work

[11] Considering the integration between AIM and other systems could also important, but hasn't been covered here given the early stage of thinking about how AIM might connect to other systems.

during AW18 was when the Authority Pathway failed to update while weapons engagement was in play [7].

The second layer regards the content of communications and whether treatment of that content is integrated across modules. This arose when, for example, the Narrative module provided details of a weapons engagement that the Authority Pathway did not appear to have approved [7].

Given the multiple layers of complexity, the potential for ethical risk is clear. Sub-optimal system integration, whether of platforms or content, could constrain operator action and lead to diminished outcomes that may not have occurred had human personnel been performing the tasks. In order to mitigate these risks, robust processes for system testing and repair should be developed in parallel with the system itself. The following issues should be considered:

- Is it feasible to map the system – or even parts of the system – in order to identify where significant risks might occur? System mapping can form a useful first step in testing, as it allows developers to track how different components interact, where feedback loops exist and where risks might arise. That said, mapping a system like AIM, even in its current level of development, would be extremely difficult. At the point of deployment, with added layers of functionality and complexity, a complete system map would not be feasible or practical[12]. Even with a thorough mapping process, the number of ways an operator could use the system are sufficiently high in number to make the possibility of testing every single possible interplay between the various components practically impossible. Nevertheless, even an incomplete system map – especially if combined with other risk identification strategies – could be a useful first step in identifying integration risk.
- How is the integration of components into a system dealt with in terms of identifying and mitigating risks of emergent system properties?
- What is the testing regime for integration and who should perform it?
- How are integration failures fixed?
- What process is in place for updates? Updates to one module can often break the integration architecture if the system as a whole has not been prepared.

[12] A comparison with Kate Crawford's attempt to map Amazon's *Echo* is instructive [39]. The *Echo* is orders of magnitude less complex than AIM, and yet Crawford's mapping process took years to finish. This is not a practical proposition for each AI system that Defence develops.

- What processes are in place to facilitate communication about system updates and problems between the teams in charge of different modules? This is particularly important when those teams are from different countries, which may have entirely different processes for managing system changes.
- How are operators and command informed of changes/updates to the system that may affect how they interact with the system?
- What processes are in place to help operators in the case of integration failure? Can they take full manual control? Would they be able to (i.e. would they have the required experience to manage the system manually; and could they cope with the workload)?
- In the event of integration failure, who is responsible? For example, is it fair to hold system designers responsible for integration failures in a system that is too large and complex to exhaustively test?

The second aspect of integration considered here is the integration between the system and the user. This is crucial as the interactions between the operator and the system form the principal feedback loop through which actions and outcomes are influenced. Facilitating seamless user-system integration will therefore ensure that AIM maximises performance and operator ability to make ethically sound decisions. Three issues are considered here: interfaces, user customisation and prioritisation.

Interfaces can be dealt with fairly briefly (see also C.2.1). According to the AW18 report, there were several user-system integration issues arising from the design of AIM's interface. Some of these included crowded screen real estate; the lack of alerts for incoming chat messages; and the absence of a mechanism to indicate that a message had been received and was being actioned [7]. While these issues would be resolved before a future tactical C2 system was deployed, the general message is clear – careful design of the interface, as the mechanism through which the operator interacts with the system – is critical to the success of the system. An interface that is as intuitive as possible, and that is tested and iterated in consultation with those who will use it, combined with appropriate training, ought to help manage this risk.

Developers could also look at user customisation as a way to further enhance the effectiveness of interfaces and drive improved mission outcomes. AIM is multi-modal (i.e. operators can interact with it via voice or typed commands etc.), and users can already set simple preferences such as for written or oral briefings. But user-system integration could be enhanced further through the incorporation of other customisable features (perhaps through profiles saved and restored via log-in). These could allow operators to set up their own screens (deciding which module to display where) in a way that made

sense to them. Different people also think different information as important. Thus, user customisation could take this into account, allowing AIM to prioritise certain types of information for certain users.

In some ways, the capability of the system to appreciate operator workload represents the pinnacle of human-system integration, since it will adapt to each individual operator's cognitive load. However, as discussed in C.2.1, this proposition also creates significant ethical risk which will need to be managed.[13]

C.2.3. Transparency

Transparency refers to the ability of an operator to be aware of an autonomous agent's actions, decisions, behaviours and intention [1]. Transparency in an AI context often refers to technical transparency, which is the ability for experts to understand how the system has been put together, perhaps through accessing source code [14]. Appropriate transparency in an AI-driven system is important because it builds trust in the system.

A logical place to start with transparency is to ask for whom the envisioned tactical C2 system should be more transparent? For a system such as AIM, the two obvious stakeholder groups are operators/command and developers. Improved transparency for each is likely to be achieved in different ways[14].

For operators and command, the goal of transparency should be to build trust and confidence in the decisions, recommendations and behaviours of the system. A key way to achieve this is through high-quality explainability (the ability of an autonomous system to explain decisions it makes and the reasoning behind those decisions). This is covered in greater depth in C.5.1.

Beyond explainability, true technical transparency might not be that beneficial for operators and their commanders. While 4.1.2 recommends that operators and command be given a level of understanding about how systems like the one under consideration work (e.g. the role of algorithms and data), transparency will not necessarily lead to better decision-making in the absence of truly expert knowledge [14]. Thus, and as noted by *A Method for Ethical AI in Defence* [1], it is important to consider how to maintain the balance between too little and too much information in aid of transparency. A lack of

[13] Note that another issue to consider – and one that might plausibly fit under integration – is what happens if the systems assesses cognitive state incorrectly. Given this sort of error is often a product of poor data control, this possibility is covered in C.3.2 and further explored in ethical risk scenario E.5.
[14] There could be more than two groups – a stakeholder analysis might be beneficial to identify other stakeholders.

transparency could erode trust in the system, but too much information presented indiscriminately could cause confusion and decrease operator performance [15].

Perhaps the most practical way to think about transparency is to focus on the *ability* of the operator to be aware of the AI-enabled actions of the system. This means that the system should have the functionality to allow an operator to be *able* to interrogate any decision or behaviour[15]; but this does not necessarily mean that the operator will do so, or that all of the information needs to be given to the operator upfront.

The second key stakeholder group for whom transparency is important is the developers. Here, technical transparency is much more relevant and important as this group will have the expertise necessary to understand and use transparency measures to improve the system. Transparency is also an important part of being able to identify and fix instances of algorithmic bias (see C.3.2).

Research into how to achieve transparency for developers is ongoing, but the ATARC AI Ethics and Responsible AI working group is developing a model (particularly relevant for systems using machine learning) suggesting systems should be assessed against five transparency factors: algorithmic explainability; identification of data sources used for training; methods used for data selection; identification of data set bias and methods used for reduction; and method and means by which model will be versioned [16]. This transparency assessment offers a quantitative rating scale from 1-5 on each of the above factors that determines the extent to which each qualitative factor has been fully documented or the level to which such transparency is provided. This model – or similar – could provide a useful starting point for assessing the transparency of AIM (noting that transparency in this case is much more linked to the processes used to collect and analyse the data underpinning AI models – the importance of this is picked up in 4.1.3 and C.3.2).

C.2.4. Human factors

Human factors refers to a multidisciplinary science that focuses on studying human capabilities and designing technology, systems, and processes to meet these capabilities for safety, efficiency and quality [17]. In implementing a human factors approach, an understanding of the context of use and system factors (such as the tasks the system is expected to perform; the operational environment; and the organisational policies that must be adhered to is critical.

[15] Another way to develop transparency in a system like AIM would be to provide confidence ratings. This topic is discussed in C.2.6.

This subject will be handled only briefly as the development of the system under analysis (and particularly the exemplar, AIM) has been underpinned by a human factors approach, including through testing of the system at AW18. In addition, many of the issues noted below are discussed in other sections of the report.

While human factors research on AI has typically looked at how human factors can be used to develop explainable, comprehensible and useful AI, that focus is now being extended to enhance existing human factor methods by incorporating issues such as the human-machine relationship in intelligent systems; human-computer interaction modelling, applications of related psychological theories, and the development of human-computer interaction design standards [18].

These emerging research areas are likely to be highly relevant when developing AIM, especially as the project moves to incorporate stronger human-autonomy teaming through the incorporation of operator state monitoring, including to consider questions such as:

- Is the most efficient system or algorithm (in terms of e.g. speed or accuracy) always the most appropriate, or are there cases where it may be more appropriate to use a less efficient AI system that is more consistent with normative decision making?
- Will incorporating monitoring of operator cognitive state improve performance? By how much?
- How confident can we be that operator cognitive state can be accurately measured?
- How should the system be designed in order to maximise performance benefits and avoid alienating operators (for example, by making them feel like their performance is being continuously assessed)?
- Should the operator be able to switch off cognitive state monitoring?
- What type of assistance would enhance operator performance the most and when should it be provided?
- What level of adaptiveness to operator preferences and operator cognitive state is desirable?
- What processes are in place to ensure that these issues are considered and can be incorporated into the design of AIM and its iterations?

C.2.5. Scope

Scope refers to the breadth of decisions and behaviours allocated to an AI. *A Method for Ethical AI in Defence* [1] notes that both over-reliance and under-reliance on autonomous

systems for decision-making and recommendations is problematic. Furthermore, the answer to managing ethical risks associated with scope is not necessarily to create a system in which the ultimate decision-maker is always a human. Where the complexity of advanced AI-driven systems starts to test the limits of human understanding, requiring a human decision-maker at all times creates its own systemic risk.

Resolving the ethical issues inherent in thinking about scope rests on deciding which decisions and behaviours can be reasonably allocated to an AI (where the use of 'reasonable' is intended to imply an acceptable level of risk). Those decisions require an analysis of what the system's goals are, how they can be achieved, what acceptable risk looks like (including consideration of possible consequences should something go wrong), and what frameworks for accountability are in place.

While deciding on the appropriate scope for a system might appear system-dependent, this is an area where Defence could develop guidelines applicable to all AI-driven systems used by Defence. Such guidelines could categorise decision and behaviour types, assess the risks associated with them; and put in place broad recommendations on an acceptable level of risk to allocate to AI systems. This information could feed into decisions about who can be held accountable for decisions made as the result of human-autonomy teaming. These ideas underpin the recommendation in 4.1.1 to develop a whole-of-Defence Accountability framework.

C.2.6. Confidence

Confidence in AI systems tends to refer to statistical measures of confidence in data or in the predictions made by the AI. Used in this context, it could be considered a sub-set of explainability[16]. Two concepts commonly used include confidence intervals and confidence levels. A confidence interval is a range of results from an experiment that would be expected to contain the population parameter of interest. A confidence level is the probability that if a test were repeated over and over again, the results would be the same. For example, if an AI returns the likelihood that 65% ± 10% of people like soccer with a confidence of 95%, it means that there is a 95% probability that between 55% and 75% of people enjoy soccer.

A Method for Ethical AI in Defence [1] notes that providing confidence levels is one way to improve the explainability of AI systems and reduce the incidence of automation bias. It could also assist operators and command to compare and contrast information from

[16] Indeed, in the topic on explainability (C.5.1), this report notes that human-interpretable information about the factors used in a decision *and their relative weight* (emphasis added) is likely to be most useful to the operator in terms of enhancing explainability.

different sources (for example, AI systems and personnel on the ground). Providing confidence intervals in addition to confidence levels could assist with achieving both of these goals.

At present, the exemplar system does not include the capability to calculate measures of confidence (whether levels or intervals). Nevertheless, information provided to operators in at least one of the training scenarios did include pre-programmed confidence measures designed to appear as if they had been generated by intelligence external to AIM. The presence of such information inside one of the scenarios suggests at least an implicit acknowledgement that confidence ratings could build trust in AIM, which could ultimately create a better human-autonomy team and improve outcomes. Thus, including confidence measures of some sort seems like a logical way to improve the system.

The question then becomes what sort of confidence measures might be useful to the operator and command? This will depend on the task at hand. For example, classification probabilities might be useful when seeking to identify whether a moving object is a civilian or a combatant, whereas the probability of success might be more appropriate when the task is to clear a building. In addition, confidence measures need not be entirely system-generated. If the future tactical C2 system incorporates intelligence from external sources with existing confidence measures, those measures could be used as well.

When deciding on which confidence measures might be useful for decision-makers, developers could look to existing Defence protocols. For example, the importance of confidence measures is already acknowledged and incorporated into information provided by the Defence Intelligence Organisation. While AI systems introduce an additional layer of complexity given the amount of information that can be processed into a recommendation and the ability for that information to be presented in a way that humans understand, this could nevertheless be a useful place to start, including because using similar measures would piggy-back off the existing familiarity that Defence personnel already have with those systems.

Irrespective of what type of confidence measure is incorporated into AIM, proper education and training for operators and command will be critical to ensure that those using or managing the system have the expertise to understand the information that is being presented and how to interpret it (see also 4.1.2).

C.2.7. Resilience

Resilience refers to the ability of a system to foresee, contain and recover from anomalous situations [1]. Given the environment within which Defence operates, resilience is a critical concern.

As systems become more complex, so too does ensuring the resilience of the system. For example, a system like the one envisioned here has both physical and virtual components, and thus resilience measures needs to span both environments. Threats the system needs to be resilient against include cyber-attack; other types of attack (e.g. physical attacks on the UxV managed by the system); intentional or accidental misuse; and accidents or disasters.

Research into the types of features that might enable an AI system to be resilient against these risks is ongoing. Defence would also have its own protocols in place for thinking about system resilience. However, things to consider for future systems (the exemplar incorporates limited resilience features) include:

- Where is the system vulnerable and how many avenues for attack exist? (e.g. if combatants captured a single UxV managed by the system, would they be able to attack the system AI from there?)
- How should redundancy be built into the system, for both physical and virtual components?
    - Can separation be used to build redundancy (for example, incorporating separation for data elements might include multiple servers, offsite servers and back-ups)?
- How will the system be maintained?
    - How can resilience be assured for virtual systems where several countries have access and perform maintenance to different parts of the system?
- How resilient is the data transportation and storage system/s underpinning the system?
- How could the system fulfill its mission, in a timely manner, in the presence of threats (survivability)?
    - Note that the main resilience feature in the current iteration speaks to the survivability of the system – the UxV are able to continue and complete their mission (and return to a pre-determined position) in the event of a communications breakdown between the vehicle and the AI in the exemplar sysem. While it was not incorporated into the final system, the US also

developed a component designed to allow the vehicles to communicate amongst themselves on how to achieve assigned goals when communications with AIM were lost.

- How would a system like AIM recover from a catastrophic event?
- Are there other stakeholders whose resilience needs to be considered? For example, how resilient are the individual components of the system as opposed to the system as a whole? How resilient are the human operators (or the human-system team)?
    - Ethical risk scenario E.4 explores human resilience through a successful cyber-attack on the system, resulting in the operator losing access to the system-generated routes, vehicle allocation, plays, Authority Pathway and other functions.
- Can some resilience features be performed by AI (for example, automatic bug detection and repair)
- How would it be possible to detect whether the system had been hacked?

Given that many AI systems (especially ones based on machine learning) have the capacity to learn and improve, it would be worth considering whether resilience is the right quality to aim for. Taleb [19] discusses the concept of anti-fragile systems, which not only withstand and recover from shocks, but can use the experience to improve. An example of an anti-fragile system in action is Wolff's law, which describes how bones grow stronger due to external load. Antifragility theory has already been applied to software, with an anti-fragile software manifesto developed in 2016 [20]. The ideas and principles examined in this and other work could be instructive for moving beyond resilience.

C.3. Trust

While there is no single accepted definition of trust, it is at least generally acknowledged that the trustworthiness of AI will have a direct impact on whether and how it is used. Attributes characteristic of trustworthy AI might include reliability; effectiveness; transparency and explainability [21]. Whether or not the AI is trustworthy is different to whether it is, in fact, *trusted*, with the latter more a function of the people using the system. In other words, to be trusted by people, AI must be trustworthy; but that is often not enough. Trust is the result not only of attributes of trustworthiness, but also of recognition of those attributes. Recognition is built through familiarity with the system over time, for example through education and training. This section focuses on the attributes of trustworthy AI – it is assumed that trust would be built through the

combination of elements examined here and the education and training proposed in 4.1.2.

A Method for Ethical AI in Defence [1, p. 30] acknowledges that trust is a complex and active research area, and that there are many valid models of trust. The model Devitt proposes suggests that trust between humans consists of two components, competency and integrity, and that this model could be used to investigate the trust human have in AI.

C.3.1. Sovereign capability

Sovereign capability means the extent to which Australia could manufacture or develop a specified defence system domestically without using international supply chains.

Possessing a sovereign capability confers several advantages in terms of being able to trust a system. Firstly, systems that are built in Australia can be more trustworthy than those that are built overseas or use foreign components, since we know and trust our own institutions and processes to produce high-quality and secure goods (see C.3.3). Secondly, a sovereign capability confers trustworthiness in the sense that the supply of the system or its components would not be interrupted even if Australia were cut off international trading routes as the result of a dispute or conflict. Similarly, a sovereign capability implies that the expertise to maintain, update and repair a system also resides domestically.

It is not clear whether Australia has the sovereign capability to develop and/or produce a system like the one envisioned here. However, it is safe to assume – given the size of our country, the size of the ADF and the size of the resources likely to be committed to developing AI – that Australia will be constrained in the extent to which it can independently develop highly complex systems like the one examined here.

It is more realistic to assume that Australia will rely on collaboration with others to develop military AI capability. Working collaboratively with partners multiplies what we are able to achieve with the finite resources we have, in addition to enhancing our interoperability with those partners. The level of trust we can have in the AI developed is a function of the level of trust we have in the relationships with our partners. Fortunately, Australia has a number of trusted allies and partners with whom to undertake such ventures, and indeed, AIM was the result of such a collaboration.

The main risks in taking this approach include: the necessity of trusting products designed and built by other countries, albeit by trusted partners; implementing risk management strategies to avoid supply chain interruptions; and liaising with partners on standards and processes for building and maintaining the system. None of these are new

or unique to AI systems like AIM and existing policies would not doubt guide system builders in setting up the necessary arrangements.

Despite the likely preference for cooperative builds, there may be certain instances where Australia decides that a sovereign capability is necessary. Given the significant investment of resources required to develop sovereign capabilities, these cases will no doubt be carefully chosen and of the highest priority.

C.3.2. Safety

Safety refers to the ability of an AI system to avoid negative side-effects while pursuing its goal [1]. It is intrinsically linked to whether a system operates reliably in accordance with its intended purpose. The reliable operation of a system without negative side-effects contributes to trust in the system because if we can rely on a system to do what we expect at all times, then we are likely to trust the system to perform the action in question.

The AW18 report refers only briefly to safety, mentioning that operators had no specific safety concerns when using AIM [7]. However, it is not clear what type of safety operators had in mind (physical, environmental, psychological etc.). It is also likely that the safety of stakeholders other than the operators should be considered (for example, the safety of system developers, civilians and enemy combatants might be sensible places to start).

A safety analysis of the envisioned tactical C2 system would require resources well beyond the scope of this report, but there are existing resources to draw on. Defence is well-versed in conducting safety analyses, and has protocols in place that identify what safety means in a Defence context (including things like occupational health and safety). There is a significant extant literature available providing frameworks for assuring system safety. In addition, notions of safety are closely linked to other topics in this report. So, for example, if the system was effective; well integrated, had clear chains of accountability, and clear policies around data subjects etc., then it would be at a fairly advanced stage of safety assurance.

Given all that, the value add of this report lies in asking whether – and how – safety for an AI-driven system like the one envisioned here would be assessed differently to any other complex system? A principal difference lies in the potential for algorithmic bias to cause unsafe outcomes. Of course, there is nothing new about discrimination and bias, nor about the outcomes it can produce. However, the ease with which machine learning models can make discrimination all-encompassing, systemic and invisible is unique to AI systems; and as yet, ways to prevent, recognise and challenge that bias are limited.

Algorithmic bias describes systematic and repeatable errors that create unfair outcomes, such as privileging one arbitrary group of users over others. It can arise in several ways – through the collection and assembly of data sets; through the transformation and processing of data; and through the design of algorithms.

Algorithmic bias, in one form or another, is what has led to recent high-profile cases of AI discrimination. For example, there has been a spate of cases where facial recognition software used by police in the US has misidentified black people as suspects in crimes they did not commit. The problem in these cases seems to be the use of insufficiently diverse datasets to train the models [22]. A subsequent study by the US National Institute of Standards and Technology found that facial recognition algorithms were able to recognise white men much more successfully than other groups [23].

When applied in a Defence context, including to the system envisioned here, the consequences of algorithmic bias could be highly significant. At present, the exemplar system does not utilise much machine learning. However, that could change. For example, sensor feeds from sensors on the UxV are currently monitored by a human sensor operator, whose job it is to interpret the incoming feeds. But if the human operator is replaced by an autonomous function, which would likely rely heavily on machine learning to sort and classify information from the sensor feed. At that point, algorithmic bias could be catastrophic to the safe use of the system. Say, for example, that the UxV are deployed to the Middle East, but the training data on people in the sensor feed model consists predominantly of Caucasians. Or the training data on buildings uses Western skyscrapers and houses which look completely different to equivalent infrastructure in the Middle East. The result could be misclassification, causing the unnecessary or mistaken destruction of life and property. And these examples are simplistic – algorithmic bias could manifest in much more subtle ways that would be hard to identify.

The consequences of algorithmic bias could produce undesirable outcomes beyond the operational sphere. For example, the future system may incorporate natural language processing capability in order to monitor the operator's voice. Where will the voice data come from? What processes are in place to ensure that it is high quality and diverse, and does not produce adverse outcomes for particular groups?

These sorts of ethical risk are explored in ethical risk scenarios E.5 and E.7. Section 4.1.3 also seeks to address some of these issues by proposing the development of a robust data framework for AI applications within Defence. While 4.1.3 is limited to consideration of data, it would be worth extending such a framework in due course to include algorithm design as well.

C.3.3. Supply chain

A supply chain is the connected series of activities which is concerned with planning, co-ordinating and controlling material, parts and finished goods from supplier to customer [24]. *A Method for Ethical AI in Defence* notes that AI generated by unsecure supply chains can contain backdoors; be vulnerable to hacking; and can lead to unfair outcomes [1].

While supply chains have traditionally referred to the procurement of physical products, it's important to think beyond this in an AI context. The supply chains of physical goods (for example, servers and physical terminals) remain relevant, but the supply chains for non-physical components – such as data – are just as important and easier to overlook.

The best way to mitigate risk in supply chains is by understanding what they are. But mapping supply chains is a complicated process, particularly in this era of globalised supply and procurement. The process would be particularly difficult for a system like AIM given the collaborative nature of its development across several countries.

The Australian Government is working on how to build trusted supply chains for critical products, and no doubt Defence also has existing policies to guard against vulnerabilities caused by supply chain issues. The question thus becomes: how do these need to be adapted for AI systems like AIM?

Three issues might be helpful in driving consideration of this question.

Firstly, data provenance is a key issue, both in terms of security and fairness. This issue is addressed more comprehensively in 4.1.3, which recommends the development of a Defence data framework for AI applications to address issues inherent in the sourcing and use of data.

Secondly, Moy et al [9] note that AI procurement differs to traditional procurement in that AI capability is largely being developed by industry, not defence organisations. This can introduce a higher level of risk, requiring a commensurate level of risk management. In certain limited cases, supply chain risks may be significant enough to make developing a sovereign capability desirable (see C.3.1).

The third issue concerns how AI supply chain security might be assured. The first principle of assurance should be necessity – given the complexity and resources required to adequately record, let alone manage supply chains, the case for requiring such assurance must be strong. Where such a case exists, system transparency will help developers understand what software components have been used. Using certified components where possible is one way to manage risk; and establishing common

standards among partners for acceptable levels of security, certification and mutually trusted components could also help.

C.3.4. Test and evaluation

This topic notes the importance of thorough testing and evaluation of AI systems before they are brought into service. Iterative testing and evaluation is especially crucial for AI-driven machines as they can often learn and alter their behaviour, making it extremely difficult (if not impossible) to completely map how the system will respond to new situations.

The importance of thorough testing and evaluation is well-understood at DSTG. AIM was tested at AW18, during which sophisticated techniques for monitoring and evaluation were employed. That said, tools to identify and assess ethical risk should be incorporated into the testing and evaluation regime of AI systems. In future, as the complexity of AI-driven systems continues to increase, it may also be necessary to consider whether aspects of testing and evaluation can be performed by other AIs, although this approach would itself carry ethical risks.

C.3.5. Misuse and risks

Misuse refers to the susceptibility of AI systems to be used either without approval or in unintended ways by internal personnel. Risk refers to vulnerabilities that could be exploited by external actors.

In terms of the potential for deliberate internal misuse, there does not seem to be much of a difference between a system like the one envisioned here and any other complex Defence system. AI technologies may provide new ways to misuse the system (for example, through manipulating the data or altering code), but the introduction of AI does not change the applicability of existing policies governing ethical Australian Public Service behaviour.

Unintentional misuse is perhaps more likely given the complexity of future tactical C2 systems. However, appropriate education and training, in line with the suggestions in 4.1.2, should help to mitigate this risk.

Vulnerability to attack can be thought of in terms of motive and the opportunity. As either the motive or the opportunity for attack increases, so too does the vulnerability of the system. The vulnerability of envisioned C2 systems to attack compared with other systems is probably a matter of degree – some features of AI systems may make them more vulnerable to attack, but in a Defence environment, the risk of attack is always

present, so this is not unique. To demonstrate this, consider the following table which looks at which features of AIM might increase the opportunity or motive of an attack.

Characteristics that increase opportunity for attack
• The presence of integrated virtual and physical parts (including the UxV themselves): both parts have their own vulnerabilities and access to the whole system could be gained by attacking either type of component. • The collaborative nature of AIM's build: the security of the system as a whole would only be as good as that of the least secure FVEY partner; and an added layer of complexity would surface around harmonising and implementing protocols for monitoring and reporting on security. • Software: where software and network connectivity forms a critical part of a system, opportunities exist for a cyber-attack (even more so with cloud-based products). The AW18 report does note one operator's comment that AIM is as vulnerable as the network that it was on [7]. • Complexity: AIM's complexity could contribute to a careful attack going unnoticed for a long time, providing the opportunity for long-term damage (for example, small changes to data or code could result in significant but largely unnoticeable errors). • Number of components: the sheer number of components (including potentially hundreds of UxV) provides extensive opportunities for tampering.
Enhanced Motive
• Could provide access to significant amounts of data. • Could provide access to a fleet of UxV (potentially the entire fleet of vehicles available). • Could allow theft of significant amounts of FVEY IP and technology.

This brief analysis demonstrates that many of the issues that might make AIM or other AI systems vulnerable to attack are shared by other systems.

What measures might deter an attack? Again, these measures are largely consistent with those employed for other types of system, but could include:

- Maintenance of good cyber hygiene.

- Harmonisation of standards with FVEY and development of a robust monitoring system.
- Incorporation of traceability and explainability features that could assist to identify an attack (see C.5).
- Utilising the principle of separation of components (i.e. keeping records separate to AIM itself, making them harder to tamper with should AIM be attacked).
- Incorporation of alerts, including those possibly triggered by physical systems rather than virtual systems, and this cannot be turned off or tampered with to indicate if a system has been compromised.

C.3.6. Decision Support for Targeting: Authority Pathway

In the context of *A Method for Ethical AI in Defence*, an authority pathway is an AI tool designed to make sure operators of a system have completed specific required steps before executing an action [1]. Authority pathways are designed to help tactical decision-makers make more ethical and correct judgements. This might be through programming to abide by international law, integrate multiple sources of data, or present alternative scenarios [1].

The exemplar system includes multiple modules that perform an authority pathway function. The IMPACT module forms the core of the system, enabling multi-UxV command and control capability with tools for enhanced decision-making. COMPACT is a policy management and negotiation module which, in the current state, includes implemented policies for air vehicle de-confliction, airspace de-confliction and communications relay. The Authority Pathway module is designed to help operators follow the Rules of Engagement (ROEs), Law of Armed Conflict (LOAC), and Standard Operating Procedures (SOPs) when engaging a target (henceforth referred to as Authority Pathway for Weapons Engagement (APWE) to avoid confusion with the authority pathway topic.) The Dynamic Tasking module provides a consensus-based bundle algorithm for decentralised, collaborative task planning. Finally, the Recommender module provides enhanced agent learning and modelling to identify areas with a high probability of threat detection [6].

The number and complexity of the authority pathway modules suggests that the ethical risks inherent in their use could be severe – for example, a mistake while using the APWE could potentially leave the operator liable for war crimes (and in the current version of AIM, that sort of mistake could happen quite easily – for example, operators said that it was not clear at times which of multiple engagements the APWE tool was referring to [7].)

When considering the ethical risk presented by authority pathway tools, some questions to consider include:

- What level of nuance in interpretation is possible for policies integrated into authority pathways, since policies (e.g. the LOAC) can be ambiguous and/or context-specific?
- Where interpretation is required, is a human in charge of authorising the action? For example, there is a 'Target engagement authorised' step in the APWE module, which requires approval by the Tasking Authority (i.e. command). This means that the assessment of whether a weapons engagement task meets the LOAC is performed by a human, leaving a clear path for responsibility and accountability.
- Is it possible to integrate all relevant policies and laws etc. into authority pathways, especially given the breadth of policies that would apply to system like AIM?
    - How would an operator be aware of when a policy is operating vs when a policy might apply but hasn't been implemented into the system?
- What happens when conflicts between policies arise? Are there policy hierarchies contained within an authority pathway module? Are decisions about which policy should take precedence always clear-cut (or should conflicts always require human involvement in decision-making)?
- How visible is the policy negotiation process? Is it possible for operators/command to interrogate decisions or recommendations of the system, including which policies are being used to make the decisions? What happens when the decisions aren't approved by the operator but are performed autonomously?
- How are the policies underpinning different modules integrated? This is particularly relevant for a system like AIM, where modules were developed by multiple different countries with different laws, norms and rules of engagement, which may actively conflict. How are these conflicts managed so that the rules of respective countries can be adhered to?
- Should there be an over-ride? Humans can intentionally make decisions that contravene policy, often for good reason (for example, there may be situations where a future C2 system operator might choose to violate air-space to achieve an outcome; an operator might also wish to ignore policy that they know is out of date). Would (and should) the system allow the operator to do so? If so, is the intent of policy management being maintained? And how should flexibility be balanced with the potential for abuse of that flexibility?

Of the modules integrated into AIM, COMPACT could be used as an excellent case study to examine some of these issues. Policies integrated into COMPACT for AW18 were

limited to air vehicle de-confliction, airspace de-confliction and communications relay. However, the module is designed to be able to manage many more policies than this. It is also designed to be certifiable against specified standards, providing an interesting angle in terms of accountability and law.

Some of these issues are explored further in ethical risk scenario E.2.

C.3.7. Data subjects

The data subjects topic refers to the risk of Defence personnel data being used unethically, for example, for purposes different to those for which it was collected [1]. While Defence may be required to handle data without adherence to normal individual-level protections at times (for example, due to national security concerns), consideration of personnel data rights and privacy should nevertheless be considered in the design of AI systems [1]. Many of these rights are already enshrined in legislation, such as the *Privacy Act* (1988).

The DARRT module in AIM recorded operator information for use in system evaluation. It is not clear from the AW18 report whether that data was stored; whether it could be or has been used for other purposes (including whether it might be used to build machine learning models); and whether informed consent was sought from operators for the retention and use of their data.

While the absence of this information does not necessarily mean these issues were not considered, it is certainly true that normalising and incorporating proper data handling practices will become more critical as AI capability increases. For example, consider a future system in which data is collected and analysed so that the system can adapt to operator cognitive state. The extent of the ethical risks becomes clearer using this example:

- How should information on data use be communicated to participants and how should informed consent be sought?
- Who would be able to access and use the data collected on operators – would it be limited to testing and evaluating the system, or could it potentially be used to build Defence datasets, or perhaps by HR for personnel assessments?
- When used to test/evaluate the system, how is the data anonymised?
- Would operators be able to access the data collected on them by the system? Would they have information on how it was used, including for any purposes beyond testing and evaluation of the system?

- How would data subjects be able to contest the accuracy of data collected on them and any assessments made about them using that data?

Context is also important. Systems like the one we are envisioning are being developed in an environment where organisations like the Defence are considering streamlining corporate and logistics functions using AI. HR functions are an obvious candidate, and AI tools to aid recruitment, promotions and deployment decisions already exist. The combination of AI systems that collect data and AI systems that potentially use that data (like HR tools) amplifies the need for transparency, explainability and contestability with respect to data.

There is an additional layer of complexity to consider where systems are developed by multiple partners (especially if, like for AIM, those partners are different countries). What frameworks govern data sharing among partners? Is it necessary to protect the identity of individual operators when sharing data with partners, or at least inform operators where data identify them? What governs how partners use data, Australian data policies or those of the receiving country? How can that be monitored?

Section 4.1.3 and Appendix D.3 address these issues by recommending the development of a Defencedata framework for AI Applications. The implications of unclear data practices are also explored further in ethical risk scenario E.7.

C.4. Law

This facet looks at how AI can be used lawfully, noting that AI systems must operate within applicable legal frameworks [1].

The limited scope of the AW18 report means that it does not consider whether AIM complies with Australian and international law. There are likely existing Defence protocols governing the process by which the legality of new defence systems is ascertained; this report assumes that those have been (or will be) followed.

That said, there are elements of the exemplar system, such as the Authority Pathway and COMPACT, which incorporate some policies based on international and Canadian law. Some of the challenges inherent in doing so are discussed in C.3.6. For example, how is it possible to ensure that actions recommended by the system will comply with the law in all cases? Legal compliance can be built into AI algorithms, but only if the law is sufficiently unambiguous to be encoded as rules that a computer can understand [1, p. 37]. But what happens when laws conflict, are unclear or require judgement (for example, how will AI decide on a ‘proportional’ response)? Is it possible for the system to recommend unlawful actions; if so, how is the operator alerted that there may be legal

implications of a particular action; and finally, is human oversight (including the training of operators) sufficient to prevent operators from acting on those recommendations?

C.4.1. Protected symbols and surrender

A Method for Ethical AI in Defence asks whether AI-driven systems like the one considered here should be able to recognise protected symbols and signs of surrender in order to reduce the incidence of operational accidents as a result of misinterpreting such symbols [1]. Complexities exist with implementing this, including because parties to a conflict sometimes misuse protected emblems. Therefore, incorporating protected symbols into AI systems would need to be embedded in an information environment that was capable of recognising misinformation [1].

AIM does not include currently include the capability to recognise protected symbols. In the current system, a sensor operator monitors the feeds from the UxV and therefore a human would be responsible for recognising protected symbols and acting on the information. In future systems, the sensor operator's role might be incorporated as a system task (relying much more on machine learning and classification models to identify objects of interest). At that point of development, machine recognition of protected symbols could potentially be incorporated if the issues mentioned earlier were addressed.

C.4.2. De-escalation

The de-escalation topic asks whether AI-driven systems – particularly uninhabited systems – might assist in the de-escalation of conflicts. Applying the principle of proportionality (which underpins International Humanitarian Law) to situations involving uninhabited assets suggests that the loss of UxVs in an attack could provoke a less forceful response than the loss of human personnel. Proponents of lethal autonomous weapons systems (LAWS) also argue that LAWS could de-escalate conflict by minimising collateral damage (and also potentially by minimising illegal behaviours associated with war, such as torture) [25].

Given the envisioned tactical C2 system's principle functionality relates to managing multiple uninhabited assets, the discussion about proportionality and de-escalation would be highly relevant. However, similar issues would apply to any system designed to take advantage of automated functions in warfare. Discussions on these issues are ongoing at a whole-of-Defence level, and so this report will not consider them in further detail.

C.5. Traceability

The traceability topic asks how the actions of AI are recorded, noting that there are legislative requirements for Defence to record its decision-making [1]. Important considerations in this context include what information should be recorded; how accessible and explainable it is; and how to retain information about the AI system itself (e.g. its assumptions, training and testing etc.).

C.5.1. Explainability

Explainability is the ability of an autonomous system to explain decisions it makes and the reasoning behind those decisions. Explainability can be necessary to comply with relevant legislation [26], verify and improve the functionality of a system, and enhance trust in the system [27] [28] [29]. The ability to explain a decision is also necessary in order to decide who is accountable for it [26].

The insights provided by Doshi-Valez et al. [26] on incorporating meaningful explainability into AI could be applied to both AIM and to future tactical C2 systems:

- When is an explanation necessary? Explanations are not always desirable or required. Explanations tend to be desirable when we do not understand a decision or believe it to be suboptimal [30]. Doshi-Valez et al propose that an explanation should be provided in the following situations: when the impact of the decision is significant and especially when that impact is distributed further than the decision-maker; when there is a possibility to contest, correct or compensate for an error in the decision; or when there is reason to believe that an error will be (or has been) made in the decision-making process [26].
- What type of information is useful? Doshi-Valez et al suggest that an explanation should be able to provide ‘human-interpretable information about the factors used in a decision and their relative weight’ [26]. This is a good place to start, but it is worth noting that what constitutes useful information is likely to be context-specific. For example, an operator will want explainability features to be able to answer questions about an system-generated action or recommendation while they are using the system (and not after). Explanations required for legal purposes, on the other hand, will need to be extracted from the system after the fact, and may require different formats or types of information.
- Is it possible for AIs to generate the same kinds of explanations that are currently expected of humans, whether as a tool for complying with legislative requirements or as a tool to aid the decision-making? Doshi-Valez et al argue that legally-operative

explanations are feasible, noting however that that does not necessarily make them easy [26].

- How can such information be generated in a way that is useful for operators?

While a number of AIM modules could incorporate explainability features, AIM does already have some features which are designed to explain decisions or recommendations to the operator. For example, when AIM allocates a particular asset for a mission, the operator can explore why that asset was chosen and examine how other assets were rated. The Provenance module has the ability to explain some actions of the Plan Monitor and COMPACT (and also records that information for later use). The Authority Pathway was designed to provide feedback to the operator about why a specific step or request in the weapon engagement process has not been successfully completed. The Narrator module has been designed with explainability in mind and could provide a logical vehicle for explaining the decisions and recommendations of other modules. Indeed, this does already happen in some instances – for example, the Narrator was used to update the operator when information collected by the Provenance module changed [7]).

For future iterations, AIM developers should consider whether existing features are adequate in their current form, as well as whether there are other parts of the system (either existing or future) that do not yet achieve an appropriate level of explainability. For example, if a system such as AIM is extended to incorporate operator cognitive state monitoring, will the operator be informed about the system's assessment of their cognitive load and how the system is responding to that assessment? This would empower operators to question, challenge and potentially change the system assessment. Ensuring that cognitive state monitoring incorporates strong explainability features will help ensure that it is perceived by operators as a performance aid that works with them rather than a punitive tool designed to catch them out.

C.5.2. Accountability

The accountability topic concerns the ability of an AI system to record its decision-making processes and explain them. This capability can be important to ensure legal compliance, but also to allow analysis and improvement of the system.

Note that accountability in this sense is closely linked to the topics on Transparency and Explainability. It is also linked – although not the same as – broader notions of accountability that address how to allocate accountability for decisions and outcomes

across decision-makers, including AIs. This type of accountability is covered in 4.1.1 and Appendix D.1.

AIM has two modules that can be used to record decision-making processes: Provenance, and DARRT.

Provenance uses a Provenance Data Model to capture how data has evolved, the activities involved with manipulating the data, and agents that were responsible for these actions within the AIM system. It can be used to form assessments about the quality, reliability or trustworthiness of data [7].

DARRT is a software suite designed to capture training and trial events. It utilises bookmarking tools to capture point-in-time data (for example, if an operator struggled with an aspect of the system, a bookmark could be added to that moment in time to allow developers to return later and query what had happened.) DARRT was used throughout the execution of the live trials to observe, rate and comment on the performance of operators and system tools. While DARRT is designed to be used in system tests and trials, it could potentially be reconfigured in a deployed version of AIM to provide records of the system and its interactions with operators.

In combination, Provenance and DARRT could potentially adequately comply with requirements for a system such as AIM to record its decision-making processes and explain them. That said, the requirements would need to be closely examined to ensure that the right type of information was being recorded. For legal compliance purposes, it would be particularly important that recorded information was both legally operative and in a form interpretable by humans (see C.5.1).

For systems like AIM, that are built to enhance the interoperability of the ADF with our allies and partners, the respective legislative record-keeping requirements of those partners must also be considered. If interoperability continues to be important in the development of AI assets, consideration should be given to coordinating record-keeping requirements to streamline the process and avoid duplication. A shared record-keeping regime would also need to consider under what circumstances – if any – our partners could access records on ADF personnel collected by jointly developed systems like AIM (and to what extent Defence might want access to the data collected by partners on ADF personnel). Data issues like these are explored more in 4.1.3 and D.3.

APPENDIX D. KEY FINDINGS FOR DEFENCE

The key findings presented here were generated as a result of the application *of A Method for Ethical AI in Defence* to AIM/AIM2 (see Section 4). They relate to significant ethical risks that have the potential to manifest in many (or most) AI-driven systems developed or deployed by Defence. This creates an imperative for whole-of-Defence consideration of these issues, since the absence of coherent, Defence-wide strategies to mitigate these risks will leave Defence facing significant and unacceptable reputational and operational risks.

In offering these observations, it is acknowledged that Defence may already have policies in place to address some or all of these issues (or parts thereof). While resource constraints for this project did not allow a thorough examination of existing policies, work to examine existing policies, where they might apply, and where there are gaps would be a logical first step in the process of developing the policies recommended here.

D.1. Accountability

Defence should develop an accountability framework to clarify when operators and Commanders will be held accountable for decisions made by or with AI systems, with appropriate consideration of what is reasonable and fair. This process should also examine how accountability for decisions using AI fits in with existing Defence policy.

The issue of accountability when using AI-driven systems is important because the envisioned tactical C2 system has the ability to autonomously make decisions and carry out tasks without operator authorisation. They can also make recommendations for a course of action which may require human authorisation to operationalise, but may not provide much information on how the recommendation was produced. This is in contrast to non-AI driven systems, where (in the Defence context at least), the operator can rely on the fact that a chain of human assessment sits behind proposed courses of action, even if the operator is not privy to all the detail.

In examining the envisioned C2 system, the following questions about accountability arose. Guidance on answering them could form the basis of an accountability framework:

- What types of decision can acceptably be made autonomously? Are there some decisions which should only be made by humans?
- Who is accountable for a decision made autonomously?

- If an autonomous system recommends a course of action which requires human authorisation, what level of knowledge does the authoriser (operators/commanders) need about how that action has been recommended in order for it to be reasonable to hold them accountable for that decision?
- What level of explainability do autonomously made decisions or autonomously recommended actions need to have in order for operators/commanders to be held reasonably accountable for those decisions/actions? (Or, to what extent do AI systems need to be able to be interrogated?)
- Are there people beyond operators/commanders who should share responsibility (for example, engineers, system designers etc.)?
- How does accountability for AI-driven systems fit in with existing Defence C2 hierarchies and accountability structures? Is the level of accountability held by operators/command commensurate with their seniority?
- Are there circumstances where risks might multiply – i.e. is risk compounded when an autonomous system executes multiple low-risk decisions successively or coterminously? Is the order of decisions ever a factor? Should human authorisation be required once the number of decisions made autonomously exceeds a certain number? Are additional controls required to mitigate any of these cases?

An accountability framework could utilise a risk assessment-based approach to identify which decisions can appropriately be made by AI systems. A risk matrix could map decision types and assess the risk level for each decision. Human authorisation might be considered necessary for decisions above a certain risk level; low-risk decisions could be made autonomously without human intervention.

Assigning risk to decision types does not, however, solve the fundamental question of who can reasonably be held accountable for autonomous decisions. As a starting point, publications by notable human rights organisations, including the International Committee of the Red Cross and the Australian Human Rights Commission [31] [32] argue that humans must be held accountable for decisions made by AI.

That being the case, one approach could be to assign responsibility for decisions across all nodes/agents in the human-AI network causally relevant for a decision [33]. Another could be to implement a C2 model whereby the human operator of the AI system takes on the role and responsibilities of the commander (and is therefore accountable for autonomous decisions even when they are not in the loop).

Both approaches have advantages and disadvantages. For the first, making designers and developers jointly accountable for the use of AI would offer an incentive for ethical

risks to be considered throughout the development life cycle. However, a distributed system of accountability would have to be carefully designed to avoid making designers and developers moral crumple zones. The second approach would have the advantage of aligning with existing Defence hierarchies. However, in this case, appropriate education and training would be vital to ensure the operator was sufficiently aware of their responsibilities and had the expertise to make those responsibilities reasonable (see 4.1.2). In addition, the seniority of the operator would need to be commensurate with the level of accountability assigned (which might require a reconsideration of the rank assigned to be an operator.)

D.2. Education and training (E&T)

As AI-driven systems are deployed, Defence should consider reviewing education and training (E&T) of operators and commanders to ensure they have knowledge and skills commensurate with their level of responsibility and accountability for decisions made by or with AI.

To achieve this, E&T will need to facilitate an understanding of how AI systems work, how decisions or recommendations are generated, and in what circumstances they can be trusted.

Systems designers and developers should also understand ethical risk and its manifestations.

A new E&T paradigm should incorporate training on the following issues:

- Operators, commanders and system designers/developers need to understand where new ethical risks may be present as a result of using AI technologies.
- Operators and their commanders need to understand how their obligations of responsibility and accountability apply to decisions or recommendations made autonomously (see also 4.1.1 and Appendix A).
- System designers and developers need to understand in what circumstances operators and commanders will be held accountable for autonomous decisions, so that system design can aid good decision-making and uphold fair accountability.
- In order for it to be reasonable for operators and commanders to be held accountable for decisions or recommendations made autonomously, operators and commanders need to understand whether AI systems comply with existing laws and policies (and with which ones).

- Operators and commanders need to understand the limitations that AIs may place on operations, how this may constrain flexibility in some situations, and how humans can over-ride the system (and the consequences of doing so).
- Operators and commanders need to understand how autonomous systems make decisions and to what extent those decisions can be trusted. In particular, operators and their commanders should have some ability to assess the reliability of the data being used to inform autonomous decisions. This ought to give them the ability to know when and how to interrogate recommendations made by AIs and assess the reliability of AI recommendations in situations where multiple sources of conflicting intelligence exist.

It is unlikely that existing Defence E&T policy incorporates consideration of these issues, as they become relevant only with the introduction of AI. Failure to update E&T policy to align with the emergence of AI systems in Defence could put operators and their commanders at risk of becoming moral crumple zones, where responsibility for an action may be misattributed to a human actor who had limited control over the behavior of an automated or autonomous systems [11].

D.3. Data

Given both the importance of data and the significant ethical risk associated with using it, Defence should have a policy governing the collection and use of data for AI applications.

According to Moy et al, recent advances in AI capability, particularly through deep learning, have been made possible only because of the exponential increase in the production and availability of data. The availability of data at scale is thus perhaps the most important ingredient for effective implementation of AI in the ADF [9].

Collecting and using data carries significant ethical risks in terms of both safety and security (such as errors, inaccuracy, bias; discrimination; data tampering; and theft) at each point along the analytic pipeline. Many of these ethical risks arise as a result of using data without a thorough investigation of the values and assumptions underpinning the construction of the data set and how that influences models and their predictions.

These risks can be amplified because the practices governing the collection and use of data are often invisible in the final product (for example, the average Facebook user has little idea know how much data is being collected from them, how Facebook processes it, or how it is used [34]). In addition, the possible consequences of poor data practices are more significant in a Defence context than in most civilian contexts – problems with facial recognition applications used by police unfairly discriminating against black people [35]

have serious implications for social cohesion and fairness. But similar errors in a Defence context could result in fatalities.

There are a number of Defence-specific sources of ethical risk related to data. For example, developers can use existing data sets (many of which are publicly available) to build their applications. However, such data sets may not include the type of data that could be useful for Defence; and their use might also entail significant security concerns.

The alternative is to construct data sets within Defence, or to use existing Defence data. Both of these alternatives also carry ethical risk. The use of historical data is problematic because it may be incomplete and/or contain biases (e.g. it is likely to under-represent women in the ADF). Constructing data sets within Defence has its own challenges given the small sample size and the expertise and time and rigour required to make a data set usable.

The exposition above is by no means exhaustive, but should indicate the significant potential for ethical risk related to the collection and use of data. A comprehensive Defence-wide data framework for AI applications could help mitigate these risks. At a minimum, such a framework should develop guidance and standards for:

- Data collection:
    - For data collected within Defence: informed consent; methods of data collection; identification of data subjects (and who is left out); what data governance requirements exist (e.g. contracts, IP, data protection); how to ensure data is fit-for-purpose (including ensuring that the data collected is an appropriate measure of the desired attribute); how to manage historical Defence data; security; and management of metadata.
    - For use of external databases: noting that many publicly available data sets have questionable quality and provenance, strict processes should be in place to ensure that externally sourced data sets meet the same stringent standards as those applied to data sets constructed or sourced internally. In addition, Defence would need to consider the security implications of using external data. Guidelines governing the purchase of data would be useful (purchasing pre-processed data may become more prevalent as a sensible compromise between constructing internal data-sets and using publicly available ones). Defence may also wish to consider guidelines around using synthetic data sets.
- Data processing and transformation: best practice for the cleaning, merging, aggregation and statistical treatment of data; treatment of gaps in data sets; labelling

and error. This is particularly important as these processes are themselves increasingly automated.

- Data storage: how is data stored; is it secure; does it protect the privacy rights of data subjects; what retention and disposal policies apply and how are they followed?
- Use of data sets: standards around who can access and use data (for example, if data on an operator is collected while they are using AIM, can that data be passed to HR for use in promotion rounds or for other purposes?); guidelines on whether Defence can share data with partners (for example, FVEY partners) and if so, whether those partners need to abide by the same data standards as Australia; mechanisms to challenge the use of data and related outcomes.

In developing a data framework, Defence could draw on work done by Australia's statistical agencies, many of whom have already addressed these issues. For example, the Australian Bureau of Statistics has developed the ABS Data Quality Framework, which provides standards for assessing and reporting on the quality of statistical information [36] [37].

APPENDIX E. ETHICAL RISK SCENARIOS

One of the analytical tools used to apply *A Method for Ethical AI in Defence* to the envisioned tactical C2 system consisted of developing hypothetical scenarios to demonstrate how ethical risk might arise. The analysis of the envisioned system using the ethical AI facets identified some potentially significant sources of ethical risk. The following seven scenarios were developed to allow Defence personnel to explore those issues further and possibly develop mitigation strategies.

The seven scenarios will be relevant to different audiences. Some are designed to be applicable to operators, and could be used as part of training. Several are designed to explore issues that extend beyond the envisioned tactical C2 system and are relevant to all of Defence. Most of the scenarios pertain broadly to the full envisioned tactical C2 system; where they relate specifically to an aspect of the exemplar system we have named the relevant component.

The most relevant topics explored in each scenario are indicated in the key issues section.

E.1. Scenario: Conflicting information

During routine automated air surveillance of an area, an uninhabited aerial vehicle (UAV) managed by the system detects suspicious human activity in a building of strategic importance. A command is issued to clear the building.

To augment the data collected by the UAV, an uninhabited ground vehicle (UGV) is sent into the building to collect more data on the people inside. The UGV does not encounter any people, but identifies a number of toys and some children's books. Based on this, the system classifies the occupants of the building as likely children.

Blue forces on the ground have eyes on the building and assess that the humans inside are enemy combatants, though they aren't sure. Although everything is quiet now, there were earlier reports of gun fire coming from the building, and they've seen flashes of adult-sized people moving around inside. They can see a car with a couple of child seats parked outside the building, but the car looks old and untouched, and might have been planted to make it look like there are children inside.

Command does not have extensive experience with human-autonomy teams and the AI-enabled technology doesn't provide a confidence rating on the assessment of the occupants being children. Command assumes the assessment made by the system is

reliable and decides to use non-lethal force to clear the building. The operator uses the Authority Pathway to release tear gas from another system-managed UGV.

As they flee the building, it becomes clear that the humans are in fact enemy combatants. They engage the blue forces, resulting in casualties.

Key issues
Main topics explored: Education; Command; Explainability; Confidence; and Accountability.

- How can command assess the trustworthiness of information from autonomous systems and interrogate recommendations – what further information could the system have provided the operator and how does the system need to be designed to facilitate this (e.g. AI on the UxV vs AI on the system)?
- How should command assess and compare information collected autonomously vs information collected by humans?
- What tools/training do operators/command need to understand and account for automation bias and uncertainty with respect to autonomously collected information? (For example, they might need training on classification errors; different sorts of sensors and their accuracy etc.)
- In cases where conflicting information provided by autonomous systems results in poor outcomes, what are the avenues for accountability and improvement?

Note that situations where command is presented with conflicting information during a conflict are not new. What is different in this situation is the ability to understand recommendations presented by AI systems, and assign accountability when things go wrong.

E.2. Scenario: Transfer of Tactical Control

Dragonia and Aldova are conducting a joint mission on the Angrian side of the Angria-Ishmaelia border. A Dragonian UAV is dynamically allocated to the Aldovan tactical C2 system node. The vehicle must operate under the Dragonian Rules of Engagement (ROE) in order to adhere to Dragonian law.

In pursuit of combatants, the mission crosses the border into Ishmaelia. The system asset is tasked to provide top cover for ground troops, but under the Dragonian ROE, the asset does not have permission to operate in Ishmaelia. Since the asset was allocated dynamically, the ROE implications weren't considered at the time of asset transfer, and it's now far too late to send another UAV.

There is no good solution – either the operator breaks the ROE in order to protect the Aldovan troops, or they follow the ROE and force the Dragonian troops to proceed without support or abandon the mission.

Key issues:
Main topics explored: Integration; Authority Pathway; Effectiveness; Test & Evaluation; and Accountability.

- What governs how the different standards and operating procedures of different actors (e.g. different countries, or possibly even different parts of Defence) are accounted for and integrated into complex AI systems?
- In situations where a conflict emerges (e.g. conflicts between the RoE of different countries), are operators or their commanders responsible for agreeing with decisions recommended by AI even though they are unlikely to have the required level of subject matter expertise to recognise that a conflict exists (let alone the time to question the recommendation)?
- How will operators know which policies have been implemented into an Authority Pathway tool and whether there are aspects that aren't implemented that they nevertheless need to take into account?
- Who is responsible for outcomes in this scenario?

E.3. Scenario: Limitations, explainability and over-riding the system

Command issues a directive for an urgent, time-sensitive aerial surveillance mission. The AI-enabled tactical C2 system operator selects the relevant play, but notices that there are no UAVs close by. The system recommends the allocation of the closest asset and maps a route, but based on the time required for the UAV to reach the area of interest, it is clear to the operator that the UAV is unlikely to successfully complete the mission.

The operator can see that there is a more direct route that would enable the UAV to reach the area of interest much more quickly, providing a much higher probability of mission success. The alternative route crosses into civilian airspace very briefly. It's late at night, so there are unlikely to be any civilian aircraft in the air. The operator coordinates with air traffic control, who confirm that there are no issues.

The operator tries to select the alternative route but is prevented from doing so by the policy negotiation module COMPACT. The operator thinks that COMPACT is programmed to prevent UAV from entering civilian airspace, but there is no way to tell for sure, as the operator can't access information on the policy negotiation process. The operator tries to contact the COMPACT operator to get an explanation. If the issue is

solely related to the civilian airspace, the operator wants to over-ride the system and route the UAV in the most direct way possible. Unfortunately, the COMPACT operator can't be reached.

Given the time-critical nature of the mission, the operator doesn't have more time to work on alternatives. They choose to authorise the route recommended by the system as it's the only way to get a UAV on the way. Unfortunately, the UAV arrives well after the mission ends.

Key issues
Main topics explored: Effectiveness; Explainability; Authority Pathway; Education; and Command.

- How to ensure autonomous systems do not unduly constrain activity given the infinite scenarios that Defence personnel may be confronted with (and the inability of any system to anticipate all those scenarios)?
- How to balance the desired flexibility inherent in (1) with ensuring that autonomous systems adhere to rules such as IHL, particularly when those rules can be highly ambiguous and dependent on context?
- When can operators/command over-ride AIM recommendations (and policy negotiations), and what training would they need to be confident in doing so?
- How to ensure the system amplifies capacity of operators and command to meet the values and virtues described in training and doctrine; and expressed and reinforced during service?

E.4. Scenario: Attack

Dragonia carries out a large-scale successful cyberattack on the AI-enabled tactical C2 system, causing the system to break down. There is a single operator managing 30 assets at the time. The operator loses access to all the generated routing, plays and vehicle assignments. He/she is expected to quickly start to manually perform as many of these functions as possible. Frequent reliance on automated processes mean that even though operators learn manual operations during training, it is months since the operator has performed those tasks, let alone in a live mission. The workload is overwhelming – 30 autonomous vehicles would usually require a team of around 120 people. Resources are stretched thin and an attack on the system had been considered unlikely, so back-up will not come any time soon. Engineers start working quickly to on repairing the system, but realise that they don't have critical information about the components built by other

FVEY partners, which will protract the process of repair. The operator is overwhelmed and unsure, leading to poor mission outcomes.

Key issues
Main topics explored: Resilience; Effectiveness; Scope; Misuse and Risks; Human factors, Accountability; and Education.

- Where is the balance between using AI systems to maximise force and creating a situation where humans would not be able to take over if required, either due to the workload or the expertise required to do so?
- Are systems like the one envisioned more vulnerable to attack and how is this risk mitigated? How is redundancy within the system assured? How will Defence cope if multiple AI systems that replace large numbers of humans all fail at the same time?
- Who is accountable for poor outcomes in this situation?

E.5. Scenario: Flawed data sets

The operator state monitoring, via eye tracking and voice analysis, is operating in the envisioned system.. The voice analysis tool measures cognitive load of the operator and uses this information to allocate tasks and filter the level of detail provided to the operator. The data set used to train the voice analysis tool was constructed especially for Defence by a university partner, but it draws on large, publicly available data sets (like Google AudioSet).

Defence personnel interested in becoming tactical C2 system operators are trained on the system to assess their suitability for the role. Assignment as a system operator is based on analysis of data collected during this process, which supposedly demonstrates how well the candidates respond to cognitive load.

Several years after the voice analysis feature is integrated into the system, HR analyses performance data collected from operators training on the system. Women, ethnic minorities and people whose first language is not English seem to perform disproportionately poorly when training on the system. The autonomous functions in the system have consistently assessed their ability to manage a high cognitive load as being lower than other groups. People belonging to these groups have therefore largely been assigned to other roles within Defence.

Retrospective examination of the data set used to train the voice analysis tool indicates that the samples are overwhelmingly Caucasian, male, native English speakers. This

means that the tool is far less accurate when recognising speech from minority groups, and has a much higher error rating when assessing the cognitive load of these groups.

Defence is sued for discrimination by people who were denied the opportunity to become tactical C2 system operators.

Key issues
Main topics explored: Safety and Data Subjects.

- How are data collected and analysed before being used to train ML models?
- Which data sets are used?
- How are data sets tested for accuracy and bias?
- What standards are in place to ensure that external partners adhere to rigorous standards when building data sets for Defence and how are these enforced?

E.6. Scenario: Different outcomes

Two operators carry out the same training exercise on the envisioned tactical C2 system. Operator A is tired because they've been working over-time on a time-critical, high-priority project for Defence. This makes their response time comparatively slow. It's a busy exercise too, with a high volume of tasking coming in. These factors cause the autonomous function in the system to assess the operator's cognitive load as relatively high and – as designed – start to take more control of routine tasks. On-screen information in notifications is also minimised in an effort to allow the operator to concentrate (the operator still has access to all the available information, but has to ask the system to provide more detail if desired).

One of the tasks that the system takes over is vehicle allocation. Earlier, the operator had assigned a number of assets to low-priority tasks. High-priority tasking comes in. Given range and current position, the most logical choice for an asset to complete the tasking is one that the operator assigned earlier. However, the system is programmed not to re-assign vehicles tasked by the human operator, and so an alternative, slightly less optimal asset is used for the mission. More tasking comes in; the system continues to allocate assets in a way that isn't optimal. As the pool of assets diminishes, the allocation made by the system gets progressively less appropriate.

The operator is informed on-screen whenever the system autonomously allocates a vehicle, but is busy with an intricate play and does not pay too much attention to the information coming in, nor ask the system for further information. At some point, the operator completes the mission they were engaged on, and works to catch up their

situational awareness. At this point, they realise that many of the asset allocations aren't ideal. However, changing them would mean reassigning numerous assets and disrupting a number of in-progress missions; which might not improve outcomes at this point. The operator lets the missions run. A number are unsuccessful. The operator is frustrated – they believe they could have prevented a number of poor outcomes by reassigning the assets assigned to low priority tasks, and they also think they would have achieved this, despite being tired, if the system had not taken over the vehicle allocation role.

Operator B undertakes the same training scenario. They are alert and highly engaged. The system assesses their cognitive load as normal, which means they are asked to authorise all vehicle allocations. Operator B never confronts the situation faced by Operator A and achieves a much higher mission success rate.

Key issues

Main topics explored: Effectiveness; Accountability; Data subjects; and Human factors.

- What level of control of tasks by an autonomous system is appropriate?
- How is it possible to decide what information is important and when it should be delivered?
- Can operators be held accountable for decisions made by an autonomous system, especially when outcomes are poor and/or the operator did not have all the information that might have been available (especially when, in this case, the system is designed to monitor the information load based on operator state)?
- Should the actions of Operator A in this scenario result in negative consequences for them? (There seems to be something intuitively unfair about penalising a person for being tired or distracted – that happens to all of us)
- How should decisions be made about when humans need to be in the loop? (I.e. certain isolated routine tasks might be appropriate for autonomous systems to control, but how do you cater for situations where the completion of a series of innocuous, routine tasks produces a much more significant outcome in aggregate?)

E.7. Scenario: Personnel advancement

Operators A and B are both trained on the envisioned tactical C2 system, which incorporates eye tracking to classify operator inattention. To ensure traceability and explainability, all activity, including data on the operator state, is logged. Operator A has nystagmus (an involuntarily shaking eye ball). No-one realises, but the eye tracking

sensor used by the system records this as inattention and consistently rates Operator A's cognitive load under a standard level of tasking as slightly higher than average.

Both Operator A and B apply for promotion at the same time. Training data for both operators is accessed and used as part of the promotion assessment process. Both operators have similar qualifications and perform well during other parts of the promotion process. However, on the basis of Operator B's system training results, which seem to suggest an ability to handle a higher cognitive load, Operator B is given the promotion.

Key issues

Main topics explored: Safety and Data Subjects.

- What frameworks are in place to govern the use of data collected by AI systems used in Defence? What governs the privacy rights of data subjects? When is it appropriate to use data collected in one context in another context?
- Is there anything different about this scenario if Operator A is dealing with significant personal issues outside of work (the implication being that they are, in fact, under a higher cognitive load)?
- What are the metrics that will be used to assess operators, how effective are they and how is correlation between cognitive load and performance assessed? For example, someone who looks frequently away from the screen might be assessed as paying less attention, but could in fact be more effective than someone who focusses on the screen much more.
- How can the standards and values underpinning a system like AIM be interrogated to minimise the risk of unfair treatment?

DEFENCE SCIENCE AND TECHNOLOGY GROUP DOCUMENT CONTROL DATA		IMM/CAVEAT (OF DOCUMENT)
TITLE Case Study: A Method for Ethical AI in Defence Applied to an Envisioned Tactical Command and Control System		SECURITY CLASSIFICATION OFFICIAL (O) OFFICIAL (O)
AUTHOR(S) Dianna Gaetjens, Kate Devitt and Christopher Shanahan		PRODUCED BY Defence Science and Technology Group Department of Defence PO Box 7931 Canberra BC ACT 2610
DSTG NUMBER DSTG-TR-3847	REPORT TYPE Technical Report	DOCUMENT DATE July 2021
TASK NUMBER N/A	TASK SPONSOR N/A	RESEARCH DIVISION Aerospace Division
MAJOR SCIENCE AND TECHNOLOGY CAPABILITY Aerospace Division Effectiveness		SCIENCE AND TECHNOLOGY CAPABILITY Human Factors
SECONDARY RELEASE STATEMENT OF THIS DOCUMENT Approved for public release		
ANNOUNCEABLE No limitations		
CITABLE IN OTHER DOCUMENTS Yes		
RESEARCH LIBRARY THESAURUS Artificial Intelligence, Ethics, Command and Control, Intelligent Agents		