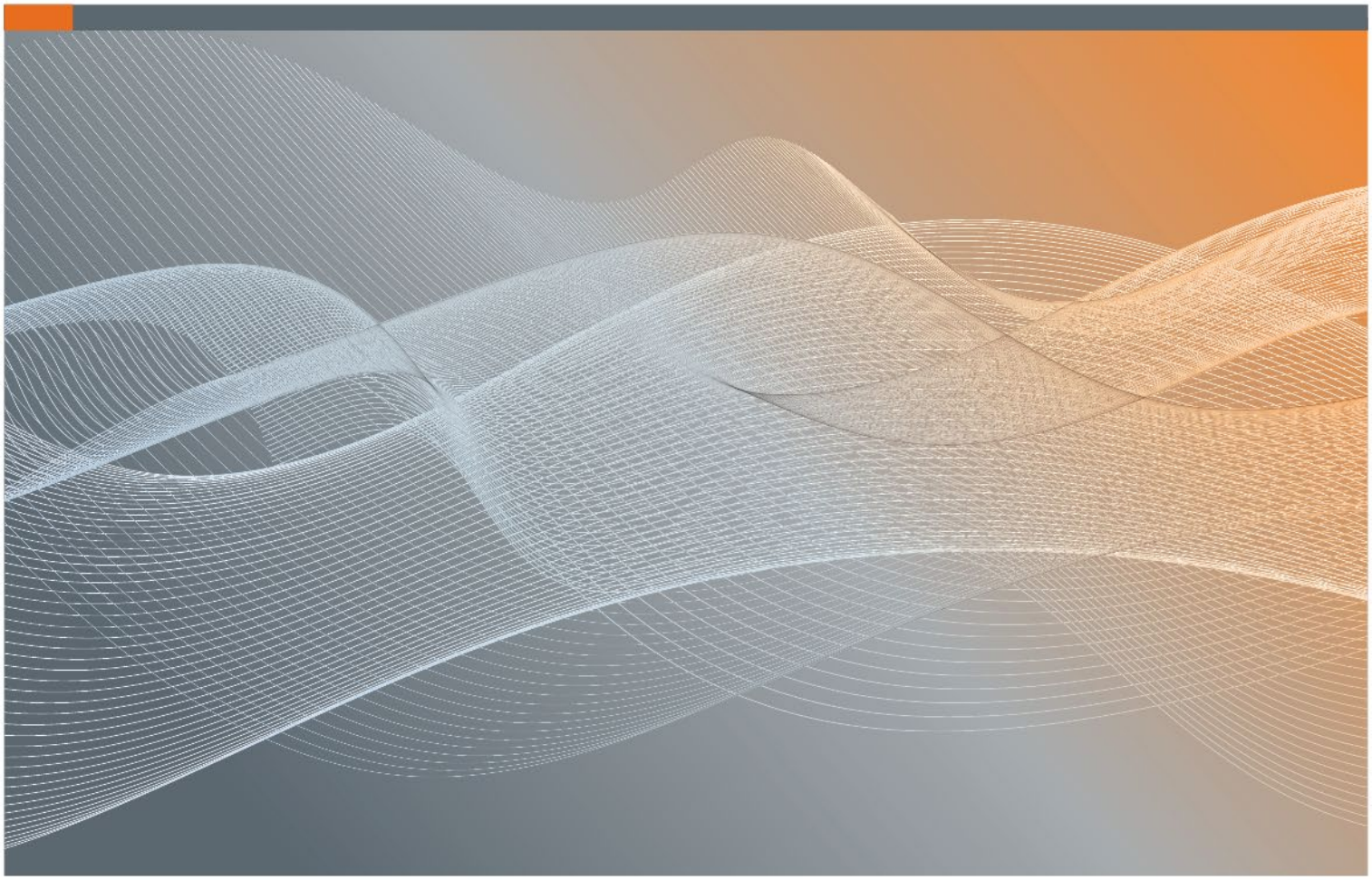
Australian Government
Department of Defence

A Method for Ethical AI in Defence



Defence Science and Technology Group

DSTG-TR-3786

This is a report on the outcomes of a workshop only and does not represent an official position of Defence. It represents views expressed by participants and stakeholders of the workshop.

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

DSTG-TR-3786

Authors
Kate Devitt[1,3], Michael Gan[2], Jason Scholz[3] and Robert Bolia[1]

[1] Aerospace Division, Defence Science and Technology Group
[2] Plan Jericho, Royal Australian Air Force
[3] Trusted Autonomous Systems Defence Cooperative Research Centre

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

EXECUTIVE SUMMARY

Recent developments in the field of artificial intelligence (AI) have highlighted the significant potential of the technology to increase Defence capability while reducing risk in military operations. However, it is clear that significant work also needs to be undertaken to ensure that introduction of the technology does not result in adverse outcomes. Defence's challenge is that failure to adopt the emerging technologies in a timely manner may result in a military disadvantage, while premature adoption without sufficient research and analysis may result in inadvertent harms.

To explore how to achieve ethical AI in Defence, a workshop was held in Canberra from 30 July to 1 August 2019. 104 people from 45 organisations attended, including representatives from Defence, other Australian government agencies, the Trusted Autonomous Systems Defence Cooperative Research Centre (TASDCRC), civil society, universities and Defence industry.

The outputs of the workshop represent a small part of a substantial and ongoing investment in appropriate methodologies, frameworks and theories to guide the development, evaluation, deployment and adaptation of ethical AI and autonomous systems across Defence and the TASDCRC.

This report articulates the views of participants and outcomes of the workshop for further consideration and does not represent the views of the Australian Government. This report will be provided to support the development of Defence policy, doctrine, research and project management.

Aim: The aim of the workshop was to develop a pragmatic and evidence-based ethical methodology for AI projects in Defence.

Objective: The objective of the workshop was to bring together some of the best national and international subject matter experts, work through complex moral issues and create a pragmatic methodology to ensure ethical AI into the future.

Method: Workshop attendees contributed evidence-based hypotheses to discussions with a view to developing methods to inform military leadership on the ethics of using AI and autonomous systems in a Defence context. Consultation with Defence stakeholders after the workshop consolidated the outputs of this report.

The workshop resulted in the identification of five facets of ethical AI in Defence (see Figure 1), twenty evidence-based topics to be explored when considering AI and a method for ensuring ethical AI in Defence.

Figure 1 Facets of ethical AI in Defence

The facets that emerged from the workshop represent broad areas of inquiry and provide a framework and resource for further investigations into ethical AI. The facets were identified by categorising evidence-based participant-driven hypotheses, taking into account applicable guidelines and principles from government, professional bodies and academia.

Workshop attendees noted that existing ethical AI principles varied by type and justification, could conflict and contradict each other and thus needed to be grounded in a clear methodology and additional governance frameworks in order to be effective. Therefore, rather than propose singular ethical AI principles for Defence, this report aims to provide those developing AI with facets of ethical AI that should be considered,

including the questions to ask, topics to consider and methods that may be relevant to Defence AI projects and their stakeholders.

The facets of ethical AI for Defence and the associated questions align with the unique concerns and regulatory regimes to which Defence is subject to. For example, in times of conflict, Defence is required to comply with international humanitarian law (IHL, *lex specialis*) and international human rights law (*lex generalis*) in armed conflict (*jus in bello*). Defence is also required to comply with international legal norms with respect to the use of force when not engaged in armed conflict (*jus ad bellum*) when applying military force. International humanitarian law, particularly the concepts of proportionality, distinction and military necessity, has no direct non-military equivalent and as such requires a specific set of requirements and responsibilities that must be considered.

Practical Methodology for Ethical AI in Defence

There was consistent agreement during and after the workshop that an effective and practical methodology would support AI projects to manage ethical risks. Three tools have been developed by the workshop organisers to assist Defence and Industry in developing AI systems for Defence. The three tools are:

- An **AI Checklist** for the development of ethical AI systems
- An **Ethical AI Risk Matrix** to describe identified risks and proposed treatment
- For larger programs, a data item descriptor (DID) for contractors to develop a formal **Legal, Ethical and Assurance Program Plan** (LEAPP) to be included in project documentation for AI programs where an ethical risk assessment is above a certain threshold.

It should be noted that the facets, questions, topics and methods identified in this report are the outcomes of a single workshop only, rather than an exhaustive review of all ethical AI considerations. Information in this report has the potential to further understanding of ethical considerations in Defence, however, subsequent ethical AI research and consultation by Defence, the TASDCRC will yield more comprehensive frameworks. To assist in facilitating this research and consultation we have developed supporting tools, including a brochure and poster, which can be downloaded along with this publication from http://www.dst.defence.gov.au/ethicalAI.

This page is intentionally blank.

CONTENTS

1. BACKGROUND

The rapid growth of artificial intelligence (AI) capabilities in the Defence sector led to the recognition that Defence requires a better understanding of the ethical issues associated with the emerging technology, as well as a robust and relevant framework to guide the development and operation of systems containing AI. The Royal Australian Air Force's (RAAF's) Plan Jericho[1] realised that experts from multiple disciplines needed to come together to address this lack of understanding and frameworks and commenced concept development for an AI ethics workshop in 2018. In early 2019, Plan Jericho, Defence Science and Technology Group (DSTG) and the Trusted Autonomous Defence Cooperative Research Centre (TASDCRC) agreed to jointly plan and run a workshop in Canberra—see Figure 2.

Figure 2 Participants listening to expert speakers and engaging in workshop activities

[1] https://www.airforce.gov.au/our-mission/plan-jericho

The intent was to develop a pragmatic ethical methodology for AI projects in Defence. The lead planners for the workshop were Dr Kate Devitt (DSTG, TASDCRC) and Wing Commander Michael Gan (RAAF, Jericho). Defence participation was organised by Wing Commander Gan, and the academic and scientific contributions and participants were organised by Dr Devitt. Fields of expertise represented by the speakers included ethics of war, ethics of data and AI, autonomous systems in Defence, adaptive autonomy, human factors that affect human-autonomy teaming, and assurance of autonomy—see Appendix C. Speakers and Facilitators at Ethical AI for Defence Workshop. A wide range of military, academic, scientific and industry participants were engaged from both Australia and overseas for the workshop activities—see Appendix D. Organisations in Attendance at the Workshop.

2. METHODOLOGY

The workshop was designed to elicit evidence-based hypotheses regarding ethical AI from a diverse range of perspectives and contexts. The work was conducted using Bayesian epistemology that recommends increasing both the diversity of stakeholders and number of independent evidential interactions on hypotheses to produce more defensible results (Bovens & Hartmann, 2004; Devitt, 2013; Hajek & Hartmann, 2009). This method encourages an inclusive, yet evidence-based approach to ethical AI aiming for more reliable and useful results for Defence. Noting that a single workshop is limited by the number of attendees and contributors it can accommodate, and the fact that it represents only a given moment in time, subsequent research using similar methodologies and appropriate parameters is recommended to ensure that the framework to ensure ethical AI in Defence is robust and defensible.

To achieve the workshop aims, an evidence-based social platform was used[2]. The platform is similar to existing social platforms such as Facebook or Reddit, works on all internet-enabled computers, tablets and smart phones and does not require any specific software to be downloaded. All participants were assumed to have at least a smart phone and therefore could access the platform. Additionally, the digital platform enabled those who were unable to attend the workshop in person to contribute remotely and asynchronously, increasing the inclusivity and diversity of attendees.

Participants were informed that a well-formed hypothesis is a simple proposition that a reasonable person could either agree or disagree with .e.g.:

> Dogs ought to be the only companion animal allowed on domestic flights inside an aeroplane cabin

When forming hypotheses, we encouraged users to use words that imply what is obligatory, permissible, or forbidden, such as:

> Only, most, all, some, many, never, ought, permitted, should, can, should not, cannot, may be, occasionally, sometimes, ought not, in some cases

Participants were not given a definition of artificial intelligence or autonomous systems to guide their thinking. Instead the workshop organisers provided participants with a set of contexts in which AI could be used in Defence, and examples of potential AI and autonomous systems, to frame hypothetical considerations—see Appendix E. Contexts of AI in Defence.

[2] The BetterBeliefs platform used at the workshop was background IP of author Kate Devitt and was used as a free trial for the workshop. Please see section declaration of perceived conflict of interests statement Appendix I.

The contexts were based on the ADF warfighting functions and modified to suit the purposes of the workshop. They were designed to capture all the potential Defence applications of AI including both warfighting and non-warfighting activities.

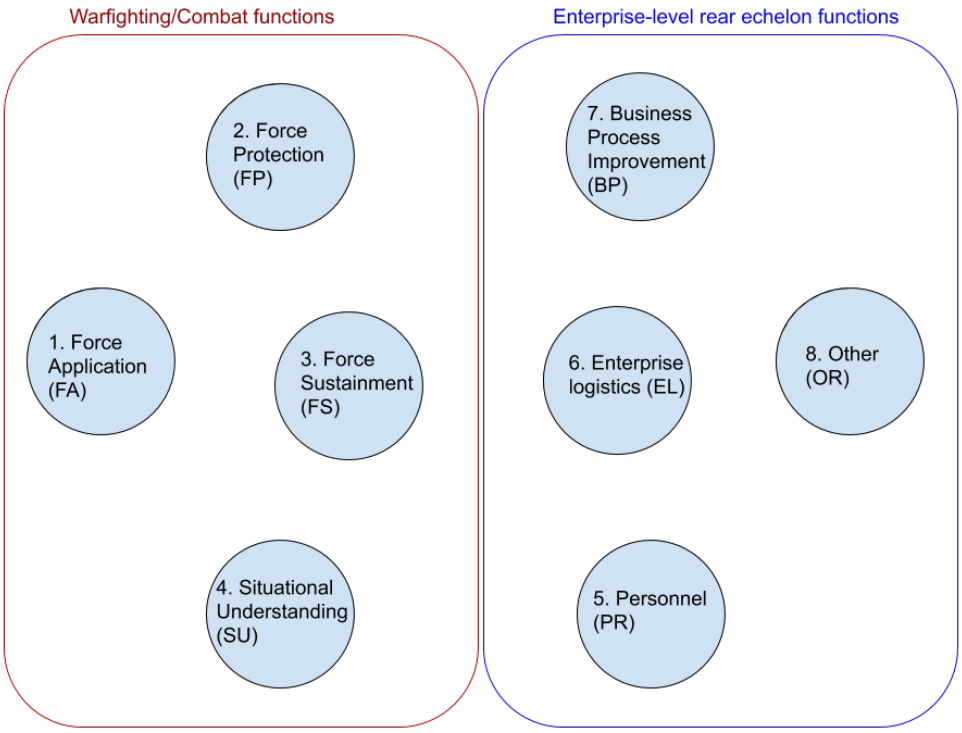


Figure 3 Table lay-out with contexts of AI in Defence during the active sessions of the workshop: Force Application (FA), Force Protection (FP), Force Sustainment (FS), Situational Understanding (SU), Personnel (PR), Enterprise Logistics (EL), Business Process Improvement (BP) and Other (OR)

Users were invited to pick one of the identified contexts, imagine some possible ethical hypothesis for this context and add that hypothesis to the platform.

Users were urged to explore a wide range of ideas and told that they did not need to strongly agree with a hypothesis to add it. Indeed, users were encouraged to:

- Add hypotheses they were sceptical in or *curious* about
- Add radical, unusual, controversial or ‘out there’ hypotheses
- Add hypotheses they would like input and feedback on
- Add hypotheses they feel are *poorly* supported by evidence
- Add hypotheses they believe in, but are not quite sure if they have enough evidence to investigate more thoroughly

- Add hypotheses they have evidence for or against
- Add hypotheses they do not believe.

Users were encouraged to use a wide range of evidence from the internet before, during and after the workshop to back up their ideas. Evidence cited for and against hypotheses ranged from Wikipedia articles, blog posts, magazine and newspaper articles, to reports by reputable institutions and peer reviewed publications. A prize was awarded to the workshop participant who contributed the greatest quality of ideas (as rated by their peers) and evidence to the platform—see Appendix H. Detailed Judging Criteria.

The organisers were aware that there are many biases that the digital platform could reveal including how safe people feel contributing potentially controversial ideas online in different circumstances. Previous use of the platform found that participants at the low end of organisational hierarchies were more interactive, generated more ideas and got greater traction than participants higher in the organisational hierarchy.

To increase participation and to ease the barriers to entry for this workshop the platform was deployed first to organisers and key stakeholders, then opened up to participants of the workshop and for 30 days after the workshop:

1. Prior to the workshop, workshop leaders contributed some initial hypotheses to the platform to provide a starting point for participants and show that ideas do not have to be perfectly worked through or fully formed; (i.e. they are intended to be tentative it is ok to be controversial etc.)
2. Speakers and facilitators were invited to steer the conversation around evidence they might have published on or have preprints on. Organisers offered to help speakers register and log into the platform and walk through via phone.
3. Organisers pre-populated a set of hypotheses modified from Institute of Electrical and Electronics Engineers' (IEEE) *Ethically Aligned Design* (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019)
4. Workshop participants were invited to register onto the platform ahead of the workshop and encouraged to familiarise themselves with the platform by voting hypotheses up or down, ranking evidence, or trying to add a hypothesis
5. Participants used the platform during the workshop around small tables of 4–8 people
6. The platform was available to participants for 30 days after the workshop
7. Participants who wished to discuss hypotheses once data-collection was complete could arrange a teleconference.

Previous use of the platform has shown that less vocal participants at workshop sessions appreciated a safe digital space to formulate their ideas and find evidence for or against them. This is in contrast with socially dominant participants who can disproportionately influence the scope of conversation at events.

The workshop program was divided into oral presentations (see Appendix C. Speakers and facilitators at Ethical AI for Defence Workshop.) and four active sessions (see Table 1). Each active session consisted of a targeted brief and small group (up to 10 people) discussions. During the active sessions, participants were invited to sit at one of eight tables, with each table focusing on a different context for AI within Defence—see Appendix E. Contexts of AI in Defence.

Hypotheses and evidence were categorised into facets and topics and compared to existing government ethical AI frameworks such as *AI Ethics Principles* approved by the Australian Government[3] (Department of Industry Innovation and Science, 2019) and *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence*, approved by the US Department of Defense (Defense Innovation Board, 2019) [4]. Principles from the IEEE's *Ethically Aligned Design* (2019)[5], plus two meta-ethical reviews published by Harvard University (Fjeld, Hilligoss, Achten, Daniel, Feldman & Kagay, 2019)[6] and Nature's *Machine Intelligence* journal (Jobin, Ienca & Vayena, E 2019) provided engineering and scientific guidance to consolidate our framework (see Appendix A. Comparison of Ethical AI Principles). Note: This report refers to these frameworks for information only. It does not seek to recommend a singular set of ethical principles for Defence. This report summarises the outcomes of the workshop and does not represent the views of the Australian Government.

[3] https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

[4] https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF. See also 'DOD Adopts 5 Principles of Artificial Intelligence Ethics' (25 Feb 2020) available from https://www.defense.gov/Explore/News/Article/Article/2094085/dod-adopts-5-principles-of-artificial-intelligence-ethics/.

[5] https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

[6] https://ai-hr.cyber.harvard.edu/primp-viz.html

Table 1 Workshop interaction sessions

Session #	Brief & Activities	Resources
#1 Military Decision Making	Attendees were briefed by Tristan Perez[7] on taxonomy of decision types that a human-AI system might be tasked with (e.g. made by a single decision-maker vs. multi-decision maker; once-off decisions vs. sequential decisions. See Appendix F. A Taxonomy of Decision ProblemsPartcipants were also briefed by Julian Tattersall on the constraints of military decision making[8]. Each table was tasked to • discuss military AI context • come up with examples of human-AI decisions in that context • identify an ethical issue or challenge within those decisions • find some evidence online (URLs) about this ethical issue • propose ethical hypotheses on BetterBeliefs relating to ethical AI for this context.	Appendix F. A Taxonomy of Decision Problems Military decision making see *The Defence Act* (The Australian Government, 1903), *ADDP 5.0 Joint Planning* (Australian Defence Doctrine Publication, 2014, Figure 1.1 The Levels of Conflict) and OODA Loop (e.g. Brehmer, 2005, pp. 1-5)
#2 Human factors	Attendees were briefed by Fiona Kerr on the human factors (cognitive, anthropological and sociological) relating to ethical AI for Defence including the human-autonomy system oversight model. Each table was tasked to • discuss military AI context with respect to cognitive, anthropological and sociological factors that affect decision-making • identify an ethical issue or challenge for decisions emerging from these considerations • find some evidence online (URLs) about this ethical issue	See Endsley (2017)

[7] See Appendix F. A Taxonomy of Decision Problems for details and also French, Maule and Papamichail's (2009) *Decision Behaviour, Analysis and Support*.

[8] Key constraints include constitutional requirements (principle of legality, rule of law) and legislative requirements –*Defence Act* 1903 (Cth). It further looked to limitations and constraints within the Executive Power (Command). Also contrasting Warfighting/combat functions (strategic, operational, tactical) vs. enterprise-level rear echelon functions (Army, Navy, Air Force, Joint and Civilian).

	• propose ethical hypotheses on BetterBeliefs relating to ethical AI in this context that accommodate for cognitive, anthropological and/or sociological factors	
#3 Ethical AI meta-analysis	Attendees were briefed by Derek Leben on meta-analyses of civilian ethical AI principles. Common principles were: promotion of human values, professional responsibility, human control of technology, fairness and non-discrimination, transparency and explainability, safety and security, accountability, privacy, and top-down ethical theories (utilitarianism, contractarianism, Kantianism, virtue ethics, ethics of care, etc.) that might constrain the meaning of principles. Each table was tasked to • discuss military AI context with respect to how bottom-up reasons might challenge the usefulness or validity of existing ethical AI principles from civilian domain • find some evidence online (URLs) to counter established ethical AI principles (add hypotheses as necessary) and add to BetterBeliefs platform • consider top-down theories that could change the way principles could be constructed • propose ethical hypotheses on BetterBeliefs using different top-down theories	See *Principled Artificial Intelligence: A Map of Ethical and Rights Based Approaches* (Fjeld et al., 2019) and *The global landscape of AI ethics guidelines* (Jobin, Ienca, & Vayena, 2019)
#4 Wrap-up	Attendees briefed on BetterBeliefs hypotheses and evidence by the end of the workshop and encouraged to continue to add to the platform until 31 August 2019 when data would be collated.	

3. RESULTS

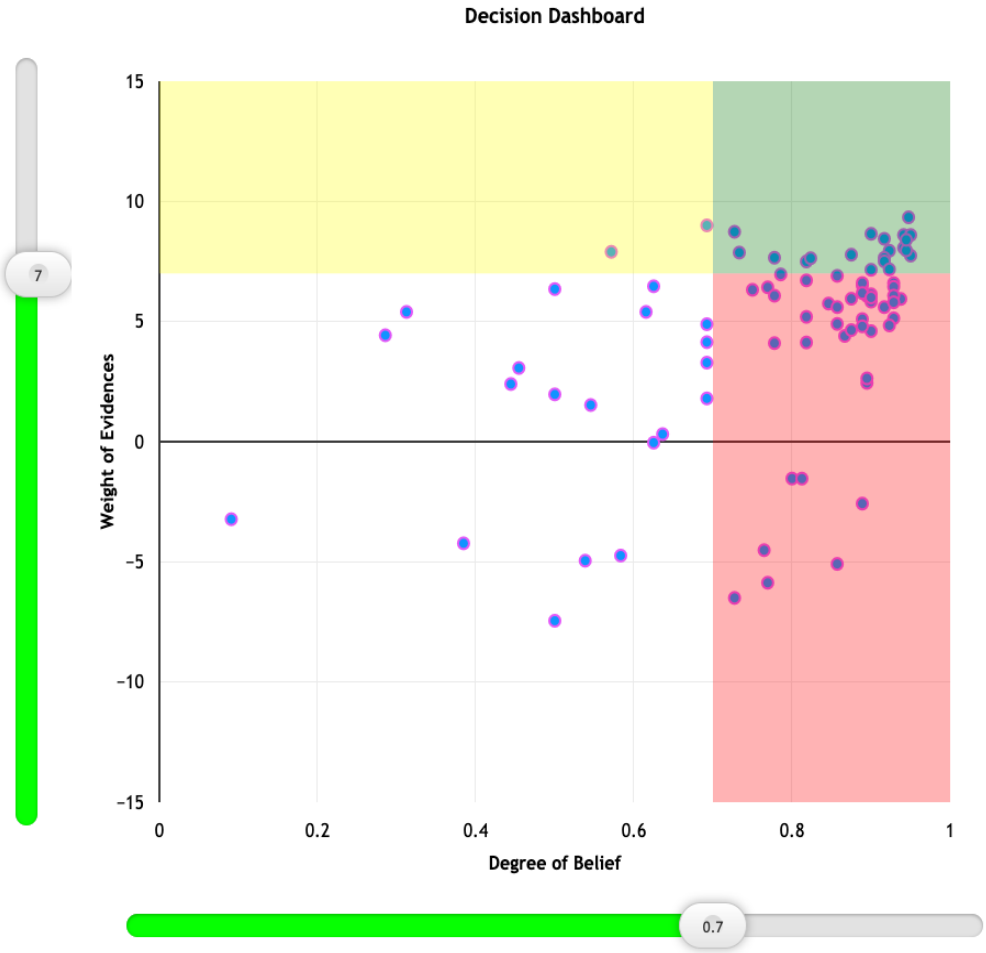


Figure 4 The decision dashboard of 84 hypotheses emerging from Ethical AI for Defence workshop

A total of 56 of 104 attendees used the online, evidence-based social platform. They:

1. *added* 84 ethical AI hypotheses
2. *added* 227 pieces of supporting or refuting evidence
3. *rated* the quality of other users' evidence 637 times
4. *voted* hypotheses up or down 964 times.

The data was patchy (as expected), revealing the limits of the method, but also provided many hypotheses to investigate into the future—see Figure 4. Hypotheses were sorted by how much evidence they had for and against them, the quality of this evidence ('Weight of Evidence-WoE' represented on the Y axis) and how much participants believed in hypotheses ('Degree of Belief-DoB' represented on the X axis). The WoE score could (in theory) keep getting 'weightier' as more evidence accrues, whereas the

DOB score was from 0.0 to 1.0 based on the proportion of users who voted each hypothesis ‘agreed’ or ‘disagree’. Refuting evidence in the algorithm counteracts supporting evidence, meaning that contentious hypotheses have a WoE score closer to ‘O’. Hypotheses in yellow, red and white zones have not been included in this report but offer opportunities for further study. Limits on the digital data collection included the absence of speaker data and the relatively brief opportunity for participants to contribute data to the platform. Hypotheses from the workshop that met a threshold for evidence (WoE > 7), belief (DoB > 0.8), items of evidence (N > 2) and diversity of contributors (N > 11) were ‘greenlit’ forming the basis of topics of this report. Topics were forged from a combination of ‘bottom up’ workshop hypotheses and top-down further consultation with key Defence stakeholders including DSTG, ADF and the TASDCRC and consideration of ethical AI frameworks. Hypotheses were then clustered and forged into five facets for consideration of ethical AI in Defence—see Table 2.

Table 2 Ethical Principles and Topics emerging from the workshop

Facets of Ethical AI for Defence	Topics emerging from the workshop
Responsibility Who is responsible for AI?	Education, 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

The facets, questions, topics and methods are evidence-based results of a single workshop only, rather than an exhaustive review of all ethical AI considerations (there were many more ideas expressed that may be valid under further scrutiny and research). Further workshops are recommended to further explore appropriate frameworks and methods for ethical AI for Defence.

3.1. Responsibility

Who is responsible for AI?

> 1.1 Commanders are appointed to conduct campaigns and operations. They are assigned military forces and have the authority to commit military personnel to battle in potentially life-threatening circumstances. Commanders therefore have a vitally important responsibility. They are accountable for their actions or inaction—*ADDP 00.1 Command and Control AL1* (Department of Defence, 2019a, p.13).

AI offers the opportunity to augment aspects of human decision-making, offering advantages in embedded expertise, larger scale operations, speed, precision and reliability, as well as enhanced patience and vigilance (Scharre & Horowitz, 2018), however it may be unclear who is responsible for decisions or actions in both combat and non-combat operations involving AI. Does the employment of AI in military operations change a commander's responsibility? If so, how? Should programmers or others be responsible for machines?

Two key challenges must be addressed when operating with AI systems, particularly those employing machine learning. Firstly, in order to effectively and ethically employ a given system (AI or not), a commander must sufficiently understand its behaviour and the potential consequences of its operation. Secondly, there is difficulty in identifying any specific individual responsible for a given decision or action.

Some machine learning systems can completely overwrite their initially programmed code based on what they learn from the environment they encounter, which in some circumstances could be an uncontrolled environment. Who is responsible for the decisions made by such machines? Environmentally driven AI autonomy (that is, information-controlled and adaptive) may give Defence advantages, but how ought the ethics of such technologies be managed?

Answering the question of responsibility underpins many of the subsequent concepts in this framework, including *governance* and *traceability*. Participants felt that education is critical to enable a commander to enact their responsibilities, particularly in combat systems.

3.1.1. Education

> The first of four key imperatives is to start educating Defence and other national security personnel about AI—Major General Mick Ryan, Commander Australian Defence College (2018)
>
> … if you want decision-makers to trust the algorithms … you need those decision-makers to be involved in, and capable of understanding, the development of those algorithms, because they are not going to necessarily be involved in the real-time decisions that the algorithms would make—Lt. Gen. Schmidle (Hicks, Hunter, Samp, & Coll, 2017)

Workshop participants considered the importance of education in the role of command. They felt that when Defence teaches leadership and management to military officers, they teach aspects of human behaviour, cognition, and social factors. Thus, for a human to lead and/or manage an AI, they will need to understand the AI. Without understanding AI, the human will be uncomfortable, and the relationship will break down quickly. It is very likely that at least some aspects of AI will be embedded in every defence function and capability. Without early AI education to military personnel, they will likely fail to manage, lead, or interface with AI that they cannot understand and therefore, cannot trust.

3.1.2. Command

In today's information age, humans issue commands to an information environment and that information environment controls industrial age machines. In the context of AI, which exists inside information environments, participants grappled with the question of who is 'in command'? Is it the coders of the algorithm, the person who procured the AI, the person who deployed the AI; or the person who relied on and applied the AI? Responsibility for critical decisions is spread across multiple decision makers from commanders through to designers, acquisition agencies and operators, offering multiple opportunities to exercise authority but also to make mistakes.

Suggestions from participants included that the allocation of ethical and legal responsibility could go across all the nodes/agents in the human-AI network causally relevant for a decision (Floridi, 2016)[9] and that AI could help reduce mistakes and augment human decision makers who bear responsibility (Ekelhof, 2018). In pursuit of accountability for military decisions, the workshop attendees felt it is important that decisions made with the assistance of or by AI are captured by accountability frameworks

[9] Mechanisms to assign responsibility can be located in back propagation from network theory, strict liability from jurisprudence and common knowledge from epistemic logic.

including domestic and international law. The International Committee of the Red Cross (2019) argues that a human-centred approach will help ensure that human beings are ultimately responsible for an AI decision.

It was noted that in high volume and high velocity information environments such as cyber, communications and electronic warfare (EW), decision makers rely increasingly on autonomous systems due to the limits of human processing capacities.

Proactive ethical and legal frameworks may help to ensure fair accountability for humans within AI systems, ensuring operators or individuals are not disproportionately penalised for system-wide and tiered decision-making. Defence can examine legal cases of responsibility in the civilian domain to guide some aspects of the relevant frameworks, e.g. the apportioning of responsibility for the test-driver in an Uber autonomous vehicle accident (Ormsby, 2019). Defence could also consider arguments that humans within complex systems without proactive frameworks risk being caught in ‘moral crumple zones’ (Elish, 2019) where the locus of responsibility falls on human operators rather than the broader system of control within which they operate—see Section 3.2 Governable.

Issues to consider in future research include the potential impact of complexity on decision-makers, how AI malfunctions are managed and how to apportion appropriate levels of responsibility to human decision-makers. Unforeseen complexity-driven factors may put an unreasonable cognitive burden on decision-makers. Apart from the complexity of a situation, a mistake, malfunction, or deliberate corruption of an AI-enabled system that processes and analyses data, information, and intelligence to inform decision-making could cause a mistake that could inform and thus undermine human decision-making in ways that could be risky or destabilizing (Kania, 2017). The complexity and risks associated with AI leads onto the principles of governability and trustworthiness. AI must be capable of operating within a human system of control (see Section 3 Governable) and the competence, integrity and security of an AI-enabled system must be ensured (see Section 4. Trusted). Further research in human decision-making with AI may seek to redefine expectations and obligations of military command when using AI.

3.2. Governance

How is AI controlled?

> 7. Australia has a longstanding and well-articulated position on the use of military force. The application of military force is controlled in accordance with Government direction and must be compliant with domestic and international law. To achieve this, Australia implements a system of control (Department of Defence, 2019b).
>
> Human discretion at some point or at some interface with machine technology is important, but that point of interface will vary—General Angus Campbell, Chief of the Defence Force (Commonwealth, 2019)

AI creators must consider the context in which AI is to be used (see Appendix E. Contexts of AI in Defence) and how AI will be controlled. The point of interface through which control is achieved will vary, depending on the nature of the system and the operational environment. There must be work conducted to understand how humans can be capable of operating ethically within machine-based systems of control.

With regards to the control of lethal autonomous weapons Australia presented a 'non-paper' at the Certain Conventional Weapons meeting Geneva (Department of Defence, 2019b) expressing the legal, policy, technical, and professional forms of controls imposed systematically throughout the 'life' cycle of weapons over nine stages—see Table 3.

Table 3 System of control of weapons (Department of Defence, 2019b)

System of Control
Stage One: Legal and Policy Framework
Stage Two: Design and Development
Stage Three: Testing, Evaluation and Review
Stage Four: Acceptance, Training and Certification
Stage Five: Pre-deployment Selection
Stage Six: Weapon Use Parameters
Stage Seven: Pre-deployment Certification and Training
Stage Eight: Strategic and Military Controls for the Use of Force
Stage Nine: After-Action Evaluation

3.2.1. Effectiveness

Participants suggested that AI systems should be deployed only after demonstrating effectiveness thorough experimentation, simulation, limited live trials etc. Robust testing will be required, allowing for the assessment of AI decision making in relevant scenarios. By presenting varying scenarios, it will be possible to assess the capability of a system to operate in environments with varying levels of risk, dynamics and decision requirements (Ahner, 2016). The IEEE's *Ethically Aligned Design* (2019) argues that creators and operators of autonomous and intelligent systems should provide evidence of the effectiveness and fitness for purpose of autonomous and intelligent systems.

3.2.2. Integration

Participants suggested that system integration would improve the robustness and diversity of decision-making (Abbass, 2019). Automotive vehicle automation provides an example of highly integrated functions/behaviours with human driver cognitive functions such as collision notifications, blind spot monitoring, assured clear distance ahead and so forth. However, it is important to consider individual differences in cognitive abilities to ensure integration fits the operator (Greenwell-Barnden et al., 2019). Social integration of AI is a natural consequence of any use of AI in the society.

3.2.3. Transparency

Transparency refers to an operator's awareness of an autonomous agent's actions, decisions, behaviours, and intention. It has been identified as one factor that could improve human trust in autonomous systems. A certain amount of transparency seems to improve operator performance such as improving situation awareness and reducing workload, however too much transparency can also decrease operator performance (Bhaskara, Skinner, & Loft, 2020; Endsley, 2016). In some contexts there is the need to make the reasoning transparent, but in others there is not. After all, people use technology all the time without knowing how it works and never question it. If an operator can act on transparent information in a timely manner, then transparency can assist with decision-making. However, being overly explanatory may lead to information overload and decision paralysis. More work is needed to ensure the balance between explainable models and maintaining performance (Turek, 2017). IEEE's *Ethically Aligned Design* (2019) argues that the basis of a particular AI decision should always be discoverable, allowing for differences in user need, e.g. the requirements of a legal review team versus an operator making tactical decisions.

3.2.4. Human factors

Participants advocated for cognitive psychology and neurophysiology to be considered in developing AI systems. Human-machine collaboration should be optimised to safeguard against poor decision-making including automation bias and/or mistrust of the system; Too much automation can render human-made decisions sub-optimal (Barnes, Chen, & Hill, 2017). The neurophysiological impacts of technological interaction and intermediation need to be better understood and factored into AI design and use. Complex information systems can lead to cognitive fatigue, distraction (via multi-modal delivery), and performance loss from neural switching. Such factors are not being sufficiently considered when deciding on the appropriate level of AI use, when not to use it, and how to better design human in/on the loop processes and partnerships (Drnec, Marathe, Lukos, & Metcalfe, 2016; Endsley, 2016; Sparrow, Liu, & Wegner, 2011).

Autonomous technologies may enable better situational awareness and a better understanding of the operational environment to allow humans to increase their control. On the other hand, autonomous technologies present fundamental challenges to military structures, the military mind-set, decision-making processes and the relationships between human actors and technologies. If these challenges are not considered carefully, the use of autonomous technologies could result in an unacceptable loss of control. Implementing these technologies gives rise to additional and new challenges with regard to human-machine interfaces, ethics, trust, training, and more (Ekelhof, 2018).

A risk as people start to work with new technology is that they experience it working reliably, rather than learning the technology’s limits. This makes users vulnerable to unknown system errors in different contexts. Reliance on automation may, over time, result in the degradation of humans’ cognitive skills and coordination capabilities’ (Hoffman, Sarter, Johnson, & Hawley, 2018).

The cognitive relationship between human and machine is critical to the system's proper use. After all, machine decision making already affects military decisions and its influence will increase as capabilities develop and the tempo of conflicts increases (Danzig, 2018). There are cases where assigning humans to decisions will not improve decision-making—such as the use of autonomous countermeasures that can respond faster than human reaction times to thwart an attack (Seffers, 2017; United States Navy, 2019; Zender, 2019).

3.2.5. Scope

Participants advised caution about over-reliance or under-reliance of AI. They pointed out that there are many technical issues such as autonomy brittleness, the capacity to deal with emergence, software validation, graceful degradation requirements and learning systems transparency that mean full autonomy is possible only under very high levels of

reliability and robustness, which we have not yet achieved. It is also necessary for the human to have an accurate mental model of the system's capabilities in order to develop sufficient trust in the system to choose to use it (Endsley, 2016).

It is not guaranteed that having a human as the ultimate decision-maker in human-autonomy teams sufficiently manages the systemic risks given increasing system complexity. Humans produce errors of action and inaction. Human intervention can be counterproductive particularly in high tempo environments that demand fast processing and communication capabilities. Human decision-makers can be nuanced, flexible and contextual in ways brittle AIs struggle with, but 'machine complexity confounds human decision makers even when there is ample time for considered judgment' (Danzig, 2018).

No matter how much autonomy is within systems, people will not use it properly if they do not trust it to do what they want it to do (Chiou & Lee, 2016). Consideration of trust and transparency in an AI system could improve the effectiveness of a human-autonomy pairing. Providing confidence in the information or choices being offered by an AI will enhance the decision capability of a human operator (Christensen & Lyons, 2017).

3.2.6. Confidence

> [To help commanders know when to trust the AI and when not to] any information that the machine is telling us should come with a confidence factor—Brig. Gen. Richard Ross Coffman (Freedberg Jr, 2019)

Participants considered whether AI systems that provide advice should also provide a level of confidence in that advice. Confidence reporting needs to be not only in terms of the classification probability object being analysed (e.g. the likelihood the AI's assessment of an object is true), but confidence on whether the class is even contained within the model (e.g. the degree to which the AI is designed and trained to assess the object in question). This is not attainable from the output of the softmax function[10] or similar, but needs to be derived by other means (Gal & Ghahramani, 2016). The prevalence of automation bias is due to users having too high a confidence in the data being presented by AI (Alexander, 2019).

There are also many different sources of uncertainty or *types of ignorance* that are relevant to the human-AI decision making—see Figure 5. 'In many real-world applications the internal decision-making process of AI must be understood in detail' (McLellan,

[10] Softmax regression is a kind of logistic regression that normalizes an input value into a vector of values that follows a probability distribution whose total sums up to 1. It allows a more nuanced interpretation of values rather than a binary 0 or 1 in a neural network model. See https://towardsdatascience.com/softmax-function-simplified-714068bf8156

2016). Understanding how uncertainty is managed by AI is critical; most AI algorithms deal with uncertainty internally in some way, perhaps using a rigorous framework of stochastic probability, or a more heuristic method. Very few attempt to deal with more than one type of ignorance, and conveying these meaningfully to human users without causing information overload is an unsolved problem.

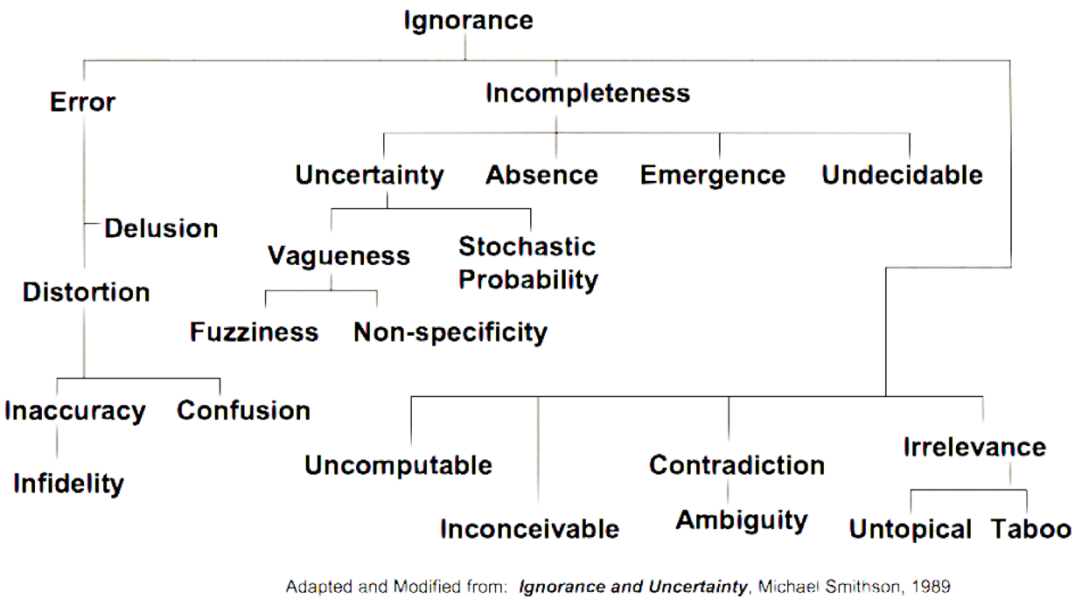


Figure 5 Taxonomy of ignorance and uncertainty by Russell Thomas (Thomas, 2013)

Recognising the true value of data or an algorithm is paramount for warfighters, commanders, developers, and certifiers being able to trust and rely on AI. It may be that confidence levels need to be disclosed for all aspects (nested confidence levels) of decisions made with or by AI. That being said it is unclear what is meant by confidence regarding an AI system.

Do confidence intervals indicate accuracy or reliability? What does a certain confidence level mean? If an AI claims 65% confidence of its decision, does it know enough of its own limitations such that this information is useful to human decision-makers? What are the consequences if the AI is wrong? Of what part of the decision-making process does the disclosure of confidence levels apply to? How will that information be used by human or machine? Participants felt that more work needed to be done to investigate.

In order for confidence levels to be useful, there must be a level of understanding of when we can use (or trust) the AI and when humans will need to intervene (calibrated trust). The information required to effectively enable humans and AI to interact is likely to be broad and varied and ultimately will only be determined through experimentation and real word application (Chen et al., 2018).

Analytic confidence by humans can be broken down along three dimensions: reliability of available evidence, range of reasonable opinion, and responsiveness to new information (Friedman & Zeckhauser, 2018). The 'grey box' process fosters trust and transparency from the human element in the output of the AI system, going some way to ensure that confidence is part of the decision process (Christensen & Lyons, 2017). New developments in AI are modelled as a reduction in the cost of prediction allowing for imperfect decisions and human adjustment until models improve (Agrawal, Gans, & Goldfarb, 2019).

3.2.7. Resilience

Participants highlighted the importance of system resilience, namely that the system exhibits the ability to foresee, contain, and recover from anomalous situations. Hollnagel, Woods, and Leveson (2006) describe how resilience can be achieved in organisations and systems, stating that it can be conceptualised as a combination of the system's ability to prevent something from happening, to prevent something from becoming worse, and to recover from anomalous situations: foresee, contain, and recover.

Participants considered whether AI should be deployed to detect system anomalies to improve the resilience of AI systems themselves, particularly those at risk of cyber-attack (Kh, 2017). This can be supervised machine learning for the detection of expected and unsupervised learning for the detection of uncommon patters. Then probabilistic reasoning should be used on how and when to re-configure the system.

Sutton and Barto (2018) provide an overview of different machine learning techniques and examples of applications which match counterparts associated with detection of anomalies, patterns and decision making. This provides the motivation to the applications of standard AI tools: supervised machine learning, unsupervised machine learning, and probabilistic reasoning.

3.3. Trust

How can AI be trusted?

Human-AI systems in Defence need to be trusted by users and operators, by commanders and support staff and by the military, government and civilian population of a nation. What is 'trust', what is 'trusted' and what is 'trustworthy' are concepts well explored by researchers (Davis, 2019; High-Level Expert Group on Artificial Intelligence, 2019; Hoff & Bashir, 2015; R. R. Hoffman, Johnson, Bradshaw, & Underbrink, 2013; Lee & See, 2004; Schaefer, Chen, Szalma, & Hancock, 2016; Wang, Jamieson, & Hollands, 2009). The High-Level Expert Group on Artificial Intelligence of the European Union 'believe it is essential that trust remains the bedrock of societies, communities,

economies and sustainable development' (2019). They argue that trustworthy AI must be lawful, ethical and robust. The model of trust in this report captures the diverse hypotheses suggested by participants and the context of establishing trust in both technical systems and in people and organisations that develop and deploy them—see Figure 6.

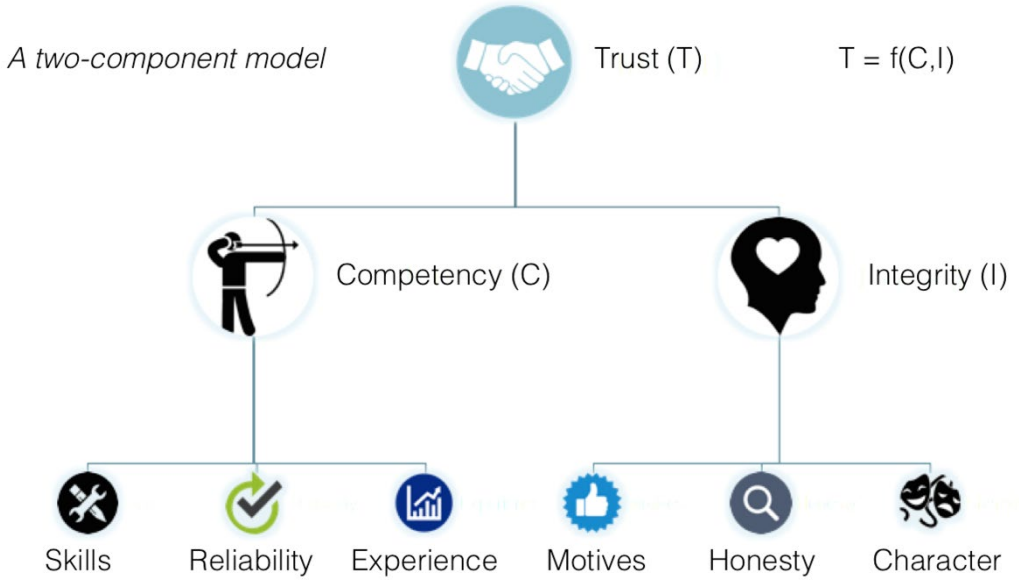


Figure 6 A two-component model of trust incorporating *competence*—skills, reliability and experience—and *integrity*—motives, honesty and character suitable for human-AI systems (Connelly, Crook, Combs, Ketchen, & Aguinis, 2015; Connelly, Miller, & Devers, 2012; Devitt, 2018; Kim, Ferrin, Cooper, & Dirks, 2004).

A two-component model suggests that trust between humans consists of both competency and integrity. Competence comprises of skills, reliability and experience; Integrity comprises of motives, honesty and character. The model is useful to understand human trust, and prompts us to consider how trust might be similar or asymmetric between humans, AI and autonomous systems. Notably, systems do not have intrinsic integrity, but exhibit behaviours and internal processing due to the integrity (or lack therein) of human-AI teams and systems. For example, humans may trust another human with high integrity (e.g. a human driver), but much lower competence than an autonomous counterpart (e.g. a self-driving car). There is a large and increasing body of literature on trust from which future AI projects may draw from.

A human-AI system can be competent and yet not have exactly the right skills to succeed in a specific context, or it fails to do a task having reached the limit of its experience. Competence improves when human-AI systems learn more skills (e.g. operator training), become more reliable (better test and evaluation) and more experienced (e.g. data

training). Integrity comprises motives, honesty and character. We trust a human-AI system that intends to be ethical, is transparent about their actions and embodies a culture that, regardless of competence, inclines them to take responsibility for their actions, be thoughtful and empathetic to others and other 'positive' traits. This two-factor model of trust combines ability and ethics.

The model can explain why humans might continue to use a technology (trusting its reliability) even when they do not trust the manufacturers. For example, services such as Google Maps are trusted by users to guide their journeys without users knowing anything about the underpinning algorithms. Users trust systems for many reasons including the system's reliability and predictability, and because people trust experts, their peers communities, organisations, government institutions, etc.

Thus, users may trust corporations such as Google to be good at mapping the world, even if they do not trust Google to not abuse their position of information power under the auspices of surveillance capitalism (Zuboff, 2019). Operators will hold multiple levels of trust in the systems they are using depending on what aspect of trust is under scrutiny. In some cases, users may develop a reliance on low integrity technology that they can predict easily, such as using the known flight path of an adversary's drone to develop countermeasures. Users may also depend on technologies because of convenience rather than trust. Finally individual differences exist in the propensity to trust, highlighting that trust is a relational rather than an objective property.

Trust is a complex and active research area. The authors offer the model as a framework to interpret the results of the workshop rather than the definitive, or the only valid model of trust.

3.3.1. Sovereign Capability

Participants considered the impact of Australia's potential reliance on overseas suppliers. Such reliance might expose Defence to anti-competitive, proprietary systems and platforms which are encumbered by International Traffic in Arms Regulations (ITAR). One author (Dingle, 2019) has raised a similar concern in relation to the lack of spare parts for the F35 Joint Strike Fighter. Workshop participants agreed that there is a risk that lack of investment in sovereign AI could impact our ability to achieve sovereign decision superiority, should this be Defence's objective.

To meet the needs of sovereign capability with regards to AI, the National Security Science and Technology Interdepartmental Committee was established in March 2017 (Defence, 2018). The committee endorsed six national security science and technology priorities, all with the potential to be improved with appropriate applications of AI: cybersecurity, intelligence, border security and ID management, investigative support

and forensic science, preparedness, protection, prevention and incident response, and technology foresighting (Callinan, 2019).

3.3.2. Safety

Participants indicated that AI systems should be safe. The safety case could be demonstrated through experimentation, simulation, limited live trials etc. AI testing in varying scenarios will help assess the capability of a system to operate in environments with varying dynamics, decision requirements and levels of risk. (Ahner, 2016). To be safe, AI should avoid negative side effects while pursuing its goals, avoid 'reward' hacking, have scalable oversight so that actions can be checked/reviewed, be able to explore safely, and be robust to 'distributional shift', that is, safe in different contexts to those it was trained in (Amodei et al., in-press). Some of the considerations of participants accord with the Australian ethical AI principles of reliability and safety that 'throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose' (Department of Industry Innovation and Science, 2019).

> Program testing can be used to show the presence of bugs, but never to show their absence—E.W. Dijkstra (1970)

However, it was noted that test and evaluation will have limited guarantees for adaptive AI systems operating in unforeseen contexts, indicating that levels of risk need to be understood in the design and deployment of these systems.

3.3.3. Supply Chain

Participants noted that AI generated by unsecure supply chains can contain backdoors or be vulnerable to hacking. Recent large-scale cyber-attacks have been the result of a breach from within a vendor or supply chain (Langcaster, 2018).

Better data transparency might ensure the provenance of suppliers. Participants considered whether AI could be used in acquisition decisions to validate and verify the veracity of the origins of the componentry from suppliers. If there are gaps in the data, AI can flag that further authentication needs to occur. Increasingly there are practical ethical templates to assist with AI procurement (The Institute for Ethical AI & Machine Learning, 2019). AI acquisition should perhaps be more like art purchasing, where authentication requires the history of certification to follow it (Dutton, 2003). See also Sroufe and Curkovic's (2008) discussion of supply chains.

Participants felt that all aspects of developing AI for the military should be scrutinised for its ethicality, not just the end product for military application. 'Cradle to grave' assessment of the ethics of AI development was felt to increase the capacity for ethical assurance in AI use for the military.

To this end, unjust biases can be identified and mitigated against in the AI algorithm from the datasets to learning protocols and interpretative layer. Some participants argued that AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups—in accordance with Australia's civilian ethical AI principles (Department of Industry Innovation and Science, 2019) and the Department of Defence ethical AI principles—see Appendix A. But participants also noted that Defence would have different obligations than the civilian domain concerning the data supply chain and whether that supply chain would be or should be transparent given security considerations. There are increasingly available tools to assist AI developers check for unfair discrimination, e.g. IBM is trying to make AI systems more transparent and biases more visible with the AI Fairness 360 toolkit. Utilising the toolkit will help identify and explain the limits and biases of training data, model bias in tests and data bias in the evaluation of deployed systems (Celis, Huang, Keswani, & Vishnoi, 2019; IBM Research Trusted AI, 2019; Lockwood, 2019; Speicher et al., 2018). IBM's toolkit is designed to improve trust in AI by demonstrating the systemic disadvantages to unprivileged groups and conversely systemic advantage to privileged groups. In a Defence context, *jus in bello* ethical principles must be abided by, particularly appropriate discrimination of combatants from non-combatants and proportionality (Coates, 2016).

3.3.4. Test & Evaluation

Participants considered the requirements of operational test and evaluation of AI before being brought into service (including Article 36 review/s for weapons under Additional Protocol 1 of the *Geneva Conventions*) (International Committee of the Red Cross, 1977).[11] An identified risk of AI is undesirable consequences from unintended combinations of legitimate rules and/or patterns. Traditional test and evaluation (T&E) defines the desired system response for all anticipated operating conditions, but the condition-response matrix for AI is intractably large, preventing engineers from fully enumerating system requirements (Scheidt, Hibbitts, Chen, Bekker, & Paxton, 2017).

Recent developments in AI such as DeepMind's AlphaGo, AlphaZero and MuZero are not programmed with explicit responses to situations encountered by the machine, but instead can solve a problem and produce a decision within a domain that was not explicitly encoded in the software. MuZero is particularly impressive as it is not even coded with the rules of the games it was able to learn to play (Schrittwieser et al., 2019). This gives AI the potential for high performance in challenging and complex domains,

[11] See the Australian Article 36 Review Process https://unog.ch/80256EDD006B8954/(httpAssets)/46CA9DABE945FDF9C12582FE00380420/$file/2018_GGE+LAWS_August_Working+paper_Australia.pdf and the *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977* https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/INTRO/470

without prior knowledge of underlying dynamics. Such AI will be challenging to test and evaluate under existing Article 36 for weapons under Additional Protocol 1 of the *Geneva Conventions (*1977).

Many AI applications rely on deep learning algorithms, especially for sensor-input recognition. Deep learning systems, however, are still prone to error even under ideal conditions. This shortcomings can be targeted by adversarial AI technologies—imagery or interference—which is targeted not at the operator, but at the AI classification algorithms at the heart of many future systems (Tramèr et al., 2018).

Participants pointed out the value of testing and evaluating AI under significant adversarial scenarios. Potentially AI T&E should be more rigorous as the second and third order effects are more likely to be unknown with new technology. Regardless, participants suggested iterative testing throughout AI design and application (Burton et al., 2020).

3.3.5. Misuse and Risks

Participants pointed out that potential misuses and risks for AI may be categorically different and/or more extensive depending on the anticipated level of autonomy and planned contexts of use (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019). For example, highly autonomous systems must be sufficiently resilient and adaptive to threats to operate in their intended context of use. As AI capabilities become more powerful and widespread there will be an expansion of existing threats due to lower costs, an expansion of actors who can carry out attacks, an increase in the rate of those attacks and an increased set of potential targets. New threats may arise from the additional capabilities afforded by AI; and malicious actors will be able to exploit new AI system vulnerabilities. Attacks enabled by the growing use of AI are likely to be finely targeted and difficult to attribute (Brundage et al., 2018).

Cyber capabilities that change system behaviour, either deliberately or as a by-product of malware, will be a significant threat to trust in autonomy (Dowse, 2018). Cyber mitigation will be key to maintaining the trust and integrity of autonomous systems. Systems must be resilient or able to defend themselves from attack, including protecting their communications feeds[12]. The ability to take control of systems has been demonstrated in commercial vehicles, including ones that still require drivers but have an ‘internet of things’ connection. In a worst-case scenario, systems could be re-tasked to operate on behalf of opposing forces, e.g. a US CIA RQ170 surveillance drone was captured by Iran

[12] In 2009 a Predator's unencrypted video feed was reportedly intercepted (although not hacked or controlled) by insurgents using a $26USD piece of commercial off-the-shelf (COTS) software.

after they took control and landed it. This provided an opportunity for technical exploitation (CNN Wire Staff, 2011).

Participants felt that AI should be developed cognisant of the risks of cyber interference and incorporate processes and systems to maintain cyber hygiene (Australian Cybersecurity Centre, 2018).The availability of ‘dirty data’ to train AI (see the MIT project 'Norman' that demonstrates how data can produce a psychopathic AI [Yanardag, Cebrian, & Rahwan, 2018]) presents a method to produce unintended behaviours. Participants suggested methods to identify when data provided has negatively influenced the behaviour and countermeasures to mitigate the risk and/or correct the behaviour.

A RAND Corporation Report (Winkelman et al., 2019) points to the liabilities and responsibilities when autonomous vehicles are hacked. Uninhabited aerial vehicles have many vulnerabilities with complex ICT architecture and multiple attack surfaces (A. Kim, Wampler, Goppert, Hwang, & Aldridge, 2012).

It should be asked whether AI creators could defend against all potential risks and misuses, as there are unknown unknowns that cannot be foreseen by even the most scrupulous organisations. A methodology for predicting unknown unknowns can reduce the risks of developing AI (Kim, 2012). In order to develop malicious AI, developers need access to the models, weightings, data and so forth that enable them to access and modify AI for nefarious reasons. For this reason, serious consideration must be given to when withholding publication of or limiting the release of AI documentation and code is justified. AI researchers and developers should consider a wider range of factors in weighing obligations for responsible publication including potential accidents, misuses, harms (and means of limiting harms) (Crootof, 2019).

Many data science tools make it difficult or impossible to assess the true accuracy of a model. It is not sufficient to validate models you need to validate the data preparation for model performance and model building including parameter optimisation and feature/s engineering (Rapidminer, 2018)

The Australian ethical AI principles recommend that decisions made by AI systems should be contestable. In a civilian domain, this means that when an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system (Contestibility; Department of Industry Innovation and Science, 2019). Workshop participants felt that Defence should consider how the use of AI systems could or should be contested within military decision-making and communicate any divergences between civilian and military ethical AI decision-making to the Australian public.

3.3.6. Authority Pathway

Participants considered the role of AI to assist tactical decision makers making ethical decisions. For example, AI could be used to help decision makers before and at the trigger point to make more ethical and correct judgements. AI paired with interactive interfaces can be built to build ethical awareness, habits, reasoning and actions (Staines, Formosa, & Ryan, 2019). AI is assisting medical decision makers in this regard (Shortliffe, 1987). An AI might be able to take all available information into account and process it efficiently to assist a human-on-the-loop or in-the-loop to determine whether objects are combatants or non-combatants potentially reducing the risk of civilian death.

Consider the Vincennes incident where a civilian Airbus A300 was mistaken for a F-14 Tomcat (Linnan, 1991). In this case, multiple conflicting sources of information and human biases lead to the loss of the civilian craft[13]. The use of AI, programmed to prioritise abidance with international law, that integrates multiple sources of data, and that is trusted by tactical decision makers might have prevented such a tragedy. Ethically-driven software might have helped the Commanding Officer receive disconfirming evidence and change their actions. Or a tool might combine multiple lines of evidence and present scenarios for the commander rating the likelihood that the A300 was a civilian aircraft rather than an F-14 fighter. However, the decision-support tool would also need to align with operational requirements. There may not be time to evaluate multiple scenarios. A clear presentation within the decision-making temporal envelop would ensure that the commander could see alternate hypotheses to the coordinated attack scenario and make a more informed decision as to the best course of action.

AI programmed with ethical and legal considerations could be incorporated into manned-unmanned teams (MUM-T) configurations such as US AFRL Loyal Wingman program. The idea being that a manned platform pairs with an unmanned off-board aircraft operating as a wingman or scout (Fawkes & Menzel, 2018).

[13] In this case the computer readouts confirm the civilian flight via IFF Mode Three civilian transponder signal, but multiple personnel recalled identification of Flight 655 as an F-14; some even remembered observing IFF Mode Two Military signals. Crew on the Vincennes also reported the aircraft as descending (as though a fighter) rather than ascending (as confirmed by U.S.S. Sides). The A300 departed the airport 20 min late which confused the crew of the Vincennes and the A300 did not respond to the multiple requests for information from the Vincennes to identify itself as commercial rather than having military intent. One reason why the disconfirming evidence from computer readouts and U.S.S Sides may not have changed the attack behaviour is that the Vincennes Commanding officer believed the Iranians were conducting a coordinated attack similar to that displayed during Operation Praying Mantis. The Vincennes was certainly under surface attack from speedboats during the incident, so the Commanding Officer's hypothesis was plausible in the situation.

3.3.7. Data Subjects

Workshop participants expressed a concern that the data of Defence personnel working in rear echelon functions might be affected by AI in HR areas such as posting and promotion, disciplinary and performance management, recruitment and retention—see Appendix E. Contexts of AI in Defence. Workshop participants were cognisant of the potentially different circumstances Defence personnel faced versus civilians with regards to data privacy. On the one hand, in the civilian domain, the Australian ethical AI principles state, 'Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection' (Department of Industry Innovation and Science, 2019). Participants noted there are some Defence uses of AI where national defence requires information to be secure and not available to data subjects. Still, Defence personnel and their data are to be treated ethically and participants felt that further consideration of the impact of AI systems on personnel was important.

Participants noted that the use of 'big data' and statistical research tools could cause harm to personnel through the use of AI. As larger pools of data are pulled into personnel research tasks there is a heightened risk that individuals could be harmed. AI programs that have access to both anonymous surveys and system level demographic data could attribute negative comments to an individual by using mosaic theory techniques Participants wondered whether this risk to the individual is acceptable when compared to the business intelligence gained by the organisation. Mosaic theory is based on research within financial or surveillance contexts but would be used to attribute all data provided by an individual (Davidowitz, 2014; Kerr, 2012; Kugler & Strahilevitz, 2016).

3.4. Law

How can AI be used lawfully?

The ADF has a strong record of compliance with applicable legal frameworks, so AI developers should be cognisant of the legal obligations within their anticipated use of the technology. Law within a Defence context has specific ethical considerations that must be understood. International humanitarian law (IHL) (*lex specialis*) and international human rights law (*lex generalis*) were forged from ethical theories in just war theory *jus ad bellum* governing the resort to force, *jus in bello* regulating the conduct of parties engaged in lawful combat (Coates, 2016), *jus post bellum* regarding obligations after combat and *jus ad vim* concerning the use of force short of war (Galliott, 2019). The legal frameworks that accompany Defence activities are human-centred, which should mean that AI compliance with them will produce more ethical outcomes (Liivoja & McCormack, 2016). Using AI to augment human decision-making could lead to better humanitarian outcomes.

There are many national laws that potentially apply to the military use of AI ranging from the *Privacy Act* and *Copyright Act*, the *Public Service Act*, *Public Governance, Performance and Accountability Act* and *Archives Act* through to the *Crimes Act* and *Criminal Code Amendment and Cybercrime Act*. In addition, there are many policies and directives that may apply, some of which have the force of law. In military contexts there will also typically be an extant set of rules called the 'rules of engagement', which among other things specify the conditions that must be met in order to fire upon a target.

Legal compliance may be able to be 'built into' AI algorithms, but this relies on them being sufficiently unambiguous and well specified that they can be encoded as rules that a computer can interpret and meets stakeholder expectations. In practice, laws are not always that clear, even to humans. In addition they can have many complicated conditions and have many interconnections to other laws. Further work is needed to clarify how AI can best enable abidance with applicable laws.

3.4.1. Protected Symbols and Surrender

Workshop participants argued that Defence AI might be used to recognise protected symbols and signs of surrender. The thought was that such AI may reduce the number of operational accidents from human error which have significant negative political and humanitarian impact. Examples of human errors include mistaken attacks on medical facilities, such as the US attack on a Médecins Sans Frontières (MSF) hospital in Afghanistan in 2015 and in multiple attacks on hospitals in Yemen by the Saudi-led coalition (Lewis, 2019). Despite both reporting their location to military forces and displaying a red crescent sign, hospitals were attacked by military forces in the mistaken belief that they were military targets. Analysis of these inadvertent attacks reveals patterns of human errors both in deconfliction (since these structures were on the no strike list) and in identification (since attacks failed to identify either the nature of medical facilities or the red crescent symbol marking the structure) of medical facilities. AI technology may enable greater protections against human errors (Oakford, 2018).

However, some parties to a conflict, in particular, non-state actors, have been known to misuse protected emblems (such as the red cross or red crescent) in order to gain a tactical advantage (such as the use of ambulances as vehicle borne improvised explosive devices in Iraq and Afghanistan). An example of a misuse of a protected symbol is falsifying protected symbols to enable them to be targeted by AI—an action that constitutes an act of perfidy. Similarly, falsely representing non-protected objects as protected objects to fool an AI would also constitute an act of perfidy. Human-AI systems that seek to improve abidance with IHL must be embedded in an information environment that anticipates deception, disinformation and misinformation with regards to protected objects.

Humans make many errors in conflicts, for example civilians being misidentified as hostile forces (Kolenda, Reid, Rogers, & Retzius, 2016). If all weapon systems recognised protected systems the incidence of even intentional incidents could be reduced. A minimally just AI, or 'MinAI' could be used today in all forms of existing conventional weapons to prevent unintended harm (Scholz & Galliott, 2018). MinAI deals with what is ethically impermissible and is contrasted with MaxAI, an ethical machine guided by both acceptable and non-acceptable actions e.g. MinAI includes the use of machine learning to detect a red cross and diverting an unintended strike, or to stopping a surface-to-air missile (SAM) from striking a passenger aircraft carrying innocent civilians (in the case of the loss of life of flight MH17), which no SAM should be permitted to do.

3.4.2. De-escalation

Participants felt that AI and autonomy might assist in the de-escalation of conflicts. For example, shooting down an unmanned drone has reduced ethical risk, given no harm or loss of life for human operators, providing a new calculus of actions in the achievement of military objectives. The shooting down of a drone may still provoke an escalation in a retaliatory use of force, but to a lesser extent than the shooting down of a crewed aircraft. The un-attributable nature of some cyber and EW systems brings new challenges to both force escalation and de-escalation. Participants felt that AI could be used to increase situational awareness for commanders to enable them to manage escalation and de-escalation with better information, knowledge and understanding of the conflict with manned and unmanned systems.

The consideration of AI and force escalation ties into a bigger question of proportionality and the use of unmanned systems. The dynamics of escalation and deterrence using these systems is evolving and needs to be better understood (Schaus & Johnson, 2018).

An illustration of a recent de-escalation related to unmanned systems is President Trump's decision to call off a strike on Iran in retaliation of shooting down a US drone on 20 June 2019. Though the strike was a legal action, Trump claims he called off the strike because he was informed that 150 people would have likely been killed in the strike. On Twitter Trump said, 'We were cocked & loaded to retaliate last night on 3 different sights when I asked, how many will die' (Trump, 2019, as cited in Chappell, 2019). Ten minutes before the strike was to begin Trump decided that the strike was not 'proportionate to shooting down an unmanned drone' (Trump, 2019, as cited in Chappell, 2019). Regardless of whether this narrative of decision-making is in fact how the decision was made, the example provides a case where a leader explained their de-escalation decisions using an ethical calculus where shooting down an unmanned (though expensive) drone is not thought to warrant the projected loss of life. Also note that though the strike was aborted, a less than lethal use of force remained open.

3.5. Traceability

How are the actions of AI recorded?

There are legislative requirements for Defence to record its decision-making. However, the increasing use of AI within human-AI systems means the manner of records must be considered. Records can represent the systems involved, the causal chain of events, and the humans and AIs that were part of decisions.

Participants felt that information needs to be accessible and explanatory; the training and expertise of humans must be open to scrutiny; and the background theories and assumptions, training, test and evaluation process of AIs must be retained. Information on AI systems should be available and understandable by auditors. That being said, just as some aspects of human decision-making can be inscrutable and some aspects of the decisions of AIs may remain opaque. It will be up to organisations that certify or acquire AI systems to determine the required levels of explanation.

When decisions lead to expected outcomes or positive outcomes, the factors that lead to those decisions may not come under scrutiny. However, when low likelihood and/or negative outcomes occur, participants felt that organisations should be able to 'rewind' the decision process to understand what occurred and what lessons might be learned. Noting that decisions made under uncertainty will always have a chance of producing negative outcomes, even if the decision-making process is defensible and operators are acting appropriately.

3.5.1. Explainability

Participants supported the consideration of human oversight, understanding and explainability of AI—noting that these concepts are complex and variously interpretable. Participants discussed the lessons in explainability of AI from the autonomous Manoeuvring Characteristics Augmentation System (MCAS) on the Boeing 737 Max which caused hundreds of civilian deaths. How MCAS was allowed to be deployed on the aircraft is a story of many human errors—where operators and testing authorities were not sufficiently informed with regards to the autonomous systems on board the 737 Max aircraft (Campbell, 2019).

In order to examine what it means for AI to be 'understood', DARPA has invested in Explainable AI (XAI) (Turek, 2019) to produce more explainable models while not reducing prediction accuracy. These will enable human users to understand, appropriately trust, and effectively manage human-AI systems. The path is not easy, as some explanations can decrease scepticism and increase automation bias (Heaven, 2020). There is no one shared definition of what constitutes a sufficient explanation for

AI. However, models of explainability from the social sciences may assist endeavours to produce truly explainable AI (Miller, 2019).

3.5.2. Accountability

Australian domestic legislation imposes an obligation on Commonwealth Departments to record and retain records relating to certain decisions. It is likely that decisions by AI will be captured by this, or similar, legislative requirements, see the *Archives Act* (Australian Government, 2016). Participants felt that human-AI logistics systems should be able to output explanations of the decisions being made in accordance with legislative obligations. Appropriate transparency in decision making improves the ability to educate the system, provide feedback on outcomes and build trust between the AI system and human operators.

Participants felt that evidence of the operation of AI systems ought to be intelligible, technically transparent to experts (see Section 3.2.3. Transparency) and explainable to stakeholders (e.g. citizens and consumers) so that they can meaningfully consent or challenge their use and operations (Blacklaws, 2018). For example, Article 22 of the *General Data Protection Regulation* (GDPR) provides some safeguards for data subjects against automated decision-making that might have legal or other significant consequences on the individual (European Parliament and of the Council, 2016).[14] Individual consent is more than merely legal or technical check boxing, it requires a human-centred, ongoing social contract (see the section titled 'Consent in a digital age' [The British Academy & The Royal Society, 2017, pp. 36-37]).

Participants suggested that Human-AI systems ought to be able to provide evidence of how a decision was made. But identifying exactly who has to provide evidence (e.g. the AI, the developer, the users of the AI, etc.) must still be investigated. Potential reviews of decisions must be supported by evidence of the AI decision process and testing to support the decision making by the AI enabled system. AI must demonstrate an evidence-based rather than an *ad hoc* decision process (Wilkinson, 2019).

The basis of any particular AI decision in Defence should be retained according to legislative requirements. No matter how an AI is deployed in Defence, its data, training, theoretical underpinning, decision-making models and actions should be recorded and auditable by the appropriate levels of government and, where appropriate, made available to the public.

[14] For more information, see ‘What is GDPR, the EU’s new data protection law?’ available at: https://gdpr.eu/what-is-gdpr/

4. METHOD FOR DEVELOPING AI ETHICALLY IN DEFENCE

There are many benefits to increasing AI and autonomous systems capabilities in Defence, including removing humans from high-threat environments, reducing sustainment costs, achieving greater mass on the battlefield, exploiting asymmetric advantage, accelerating capability development timelines and capitalising on advances made in the civil sector[15].

There was consistent agreement during and after the workshop that an effective and practical methodology would best support Defence and industry in developing AI systems. A method for ethical AI means assessing ethical compliance from design to deployment, requiring repeated testing, prototyping, and reviewing for technological and ethical limitations. Developers already must produce risk documentation for technical issues. Similar documentation for ethical risks ensures developers identify, acknowledge and attempt to mitigate ethical risks early in the design process and throughout T&E (Daniels & Williams, 2020; Vallor, 2018).

AI projects involving machine ethics should specify what ethical framework/s they are basing their de-risking strategies upon, e.g. consequentialism, Kantian, virtue ethics, ethics of care and so forth (for more on machine ethics see Cave, Nyrup, Vold, & Weller, 2019; Leben, 2019; Tavani, 2015). This workshop did not focus on machine ethics and the topic was noted by participants would be of value in subsequent research and engagement.

Three tools have been developed by the workshop organisers to assist Defence and industry in developing AI systems for Defence. The three tools (currently under internal review) are:

1. An **AI Checklist** for the development of ethical AI systems
2. An **Ethical AI Risk Matrix** to describe identified risks and proposed treatment—see Appendix B.
3. For larger programs, a data item descriptor (DID) for contractors to develop a formal **Legal and Ethical Assurance Program Plan** (LEAPP) to be included in project documentation for all programs where an ethical risk assessment is above a certain threshold—see Appendix G. Data Item Description DID-ENG-SW-LEAPP.

[15] See Marcus Hellyer's (2019) report, Accelerating autonomy: Autonomous systems and the Tiger helicopter replacement

4.1. Ethical AI for Defence Checklist

The main components of the checklist are:

A. Describe the military context in which the AI will be employed[16]

B. Explain the types of decisions supported by the AI[17].

C. 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[18]

D. Explain framework/s to be used[19]

E. Employ subject matter experts to guide AI development[20]

F. Employ appropriate verification and validation techniques to reduce risk.[21]

4.2. Ethical AI Risk Matrix

Create an Ethical AI Risk Matrix (see Appendix B. Ethical AI Risk Matrix), with detail for each project activity:

- Define the activity you are undertaking
- Indicate the ethical facet and topic the activity is intended to address.
- Estimate the risk to the project objectives if issue is not addressed?
- Define specific actions you will undertake to support the activity
- Provide a timeline for the activity
- Define action and activity outcomes
- Identify the responsible party(ies)
- Provide the status of the activity.

[16] See Appendix E. Contexts of AI in Defence.

[17] See Appendix F. A Taxonomy of Decision Problems, *The Defence Act* (1903), and 'Critical Decision Analysis' in Cognitive Task Analysis Methods (Stanton, Salmon, & Rafferty, 2013).

[18] See topics particularly under the Governance and Trusted sections

[19] For examples see Appendix A.

[20] For example, use consultants, contractors or hire employees with relevant expertise in military ethics, decision science, law, human factors, and data science to assist with AI project conceptualisation and planning

[21] Seek out best practice in autonomy and intelligent system test and evaluation methods to accelerate certification and assurance for acquisition, adoption and social license.

4.3. Legal and Ethical Assurance Program Plan (LEAPP)

For AI programs where an ethical risk assessment is above a certain threshold, a more comprehensive legal and ethical program plan should be provided. The Legal and Ethical Assurance Program Plan (LEAPP) describes a contractor's plan for assuring that software acquired under the contract meets the Commonwealth's legal and ethical assurance (LEA) requirements.

The draft Data Item Description (DID) at Appendix G. Data Item Description DID-ENG-SW-LEAPP provides guidance to contractors developing legal and ethical assurance programs for complex Defence AI systems. The LEAPP provides Defence with visibility into the contractor's legal and ethical planning, supports progress and risk assessment and provides input into Defence's internal planning, including weapons reviews under Article 36 of Additional Protocol 1. The DID will be distributed for review and comment by Defence and industry stakeholders before it is considered for Defence contracts.

4.4. Summary

There are many benefits to increasing AI and autonomous systems capabilities in Defence, including removing humans from high-threat environments, reducing capability costs and achieving asymmetric advantage. However, significant work needs to be undertaken to ensure that introduction of the technology does not result in adverse outcomes. To explore how to achieve ethical AI in Defence, a workshop was held in Canberra from 30 July to 1 August 2019. A total of 104 people from 45 organisations attended, including representatives from government, civil society, universities and defence industry. The workshop resulted in the identification of five facets of ethical AI in Defence, 20 evidence-based topics to be explored when considering AI and a method for ensuring ethical AI in Defence. This report conveys pragmatic methods to ethically de-risk Defence AI projects, but methods are also pertinent to de-risking the ethics of autonomous systems, semi-autonomous, manned-unmanned teaming and human-autonomy teaming. This report focuses on the outcomes of the workshop for further consideration and does not represent the views of the Australian Government. Tools suggested to ethically de-risk projects include: Ethical AI Checklist, Ethical AI Risk Matrix and LEAPP) (for larger acquisitions). *A Method for Ethical AI in Defence* aims to practically ensure accountability for a) considering ethical risks, b) assigning person/s to each risk and c) making humans accountable for decisions on how ethics are de-risked. The outputs of the workshop are a small part of a substantial and ongoing investment in appropriate methodologies, frameworks and theories to guide the development, evaluation, deployment and adaptation of ethical AI and autonomous systems across

Defence and the Trusted Autonomous Systems Defence Cooperative Research Centre (TASDCRC). Outputs will support the development of Defence policy, doctrine, research and AI project management. The first of these outputs are this report, and the accompanying brochure and poster, which can all be downloaded from http://www.dst.defence.gov.au/ethicalAI.

5. CONTRIBUTORS

The following individuals contributed to the report including hypotheses and evidence that informed the creation of ethical AI facets, topics and methodology: Hussein Abbass, Eugene Aidman, Andrew Back, Christopher Bailey, Saba Bazargan-Forward, Trent Beilken, Adella Bhaskara, Robert Bolia, Stephen Bornstein, Glenn Burgess, Dragana Calic, Massimiliano Cappuccio, Darren Carruthers, Jessica Casben Fell, Wygene Chong, Mal Christie, Susan Cockshell, Nikki Coleman, Emily Defina, Harley Dennett, Kate Devitt, Bradley Donnelly, Piers Duncan, Shane Dunn, Heather Emery, Peter Francis, Michael Gan, Michelle Gee, Antonio Giardina, Alex Gibbs, Damian Gilchrist, Anne Goyne, Chris Gyngell, Marcus Hellyer, Rachel Horne, Paul Jones, Dale Lambert, Derek Leben, Larry Lewis, Scott Lowe, Glenn Logan, Fiona Kerr, Ian Koegelenberg, Dean Lewis, Mark Lilley, Duncan MacIntosh, Luke Marsh, Ryan Messina, Samantha Murray, Tyson Nicholas, Simon Ng, Tristan Perez, Vanessa Pigrum, Helen Pongracic, Carmine Pontecorvo, Daniel Pope, Travis Reddy, Jerome Reid, Ben Rice, Morgan Saletta, Jason Scholz, Alison Spark, Julian Tattersall, Anh Tu, Naomi van der Linden, Tim van Gelder, Samuel White, Sarah-Jane White, Kate Yaxley and Kath Ziesing.

6. REFERENCES

Abbass, H. A. (2019). Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust. *Cognitive Computation, 11*(2), 159-171.

Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy, 47*.

Ahner, D. K. (2016). *Test and Evaluation of Autonomous Systems*. 33rd International Test and Evaluation Symposium. Retrieved from https://www.itea.org/conference-proceedings/33rd-international-test-and-evaluation-symposium-proceedings-2016/

Alexander, D. (2019, 10 May). Is our reliance on technology creating a new dark age? *Interesting Engineering*. Retrieved from https://interestingengineering.com/is-our-reliance-on-technology-creating-a-new-dark-age

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (in-press). *Concrete problems in AI safety.* arXiv preprint, arXiv:*1606.06565*. https://arxiv.org/abs/1606.06565

Australian Cybersecurity Centre. (2018). *Cyber Hygiene*. Australian Signals Directorate. Retrieved from https://www.cyber.gov.au/advice/cyber-hygiene

Archives Act 1983, Act No. 79 of 1983 C.F.R. (2016).

Barnes, M. J., Chen, J. Y., & Hill, S. (2017). *Humans and Autonomy: Implications of Shared Decision Making for Military Operations*. US Army Research Laboratory. Retrieved from: https://apps.dtic.mil/docs/citations/AD1024840

Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent Transparency: A Review of Current Theory and Evidence. *IEEE Transactions on Human-Machine Systems*, 1-10. doi:10.1109/THMS.2020.2965529

Blacklaws, C. (2018). Algorithms: transparency and accountability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2128).

Bovens, L., & Hartmann, S. (2004). *Bayesian epistemology*. Oxford: Oxford University Press.

Brehmer, B. (2005). *The Dynamic OODA Loop: Amalgamating Boyd's OODA Loop and the Cybernetic Approach to Command and Control*. 10th International Command and Control Research and Technology Symposium: The Future of C2.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Filar, B. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*. Report published by Future of Humanity Institute, University of Oxford, Centre for the Study of Existential Risk, University of Cambridge, Center for a New American Society, Electronic Frontier Foundation, and OpenAI. Retrieved from https://arxiv.org/pdf/1802.07228.pdf

Burton, S., Habli, I., Lawton, T., McDermid, J., Morgan, P., & Porter, Z. (2020). Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective. *Artificial Intelligence, 279*, 103201. doi: https://doi.org/10.1016/j.artint.2019.103201

Callinan, M. (2019). *Defence and security R&D: A sovereign strategic advantage*. Australian Strategic Policy Institute. Retrieved fromhttps://www.aspi.org.au/report/defence-and-security-rd-sovereign-strategic-advantage

Campbell, D. (2019, 2 May). Redline: The many human errors that brought down the Boeing 737 Max. *The Verge*. Retrieved from https://www.theverge.com/2019/5/2/18518176/boeing-737-max-crash-problems-human-error-mcas-faa

Cave, S., Nyrup, R., Vold, K., & Weller, A. (2019). Motivations and Risks of Machine Ethics. *Proceedings of the IEEE, 107*(3), 562-574. doi:10.1109/JPROC.2018.2865996

Celis, L. E., Huang, L., Keswani, V., & Vishnoi, N. K. (2019). *Classification with fairness constraints: A meta-algorithm with provable guarantees.* Proceedings of the Conference on Fairness, Accountability, and Transparency.

Chappell, B. (2019, 21 June). Trump says he called off strike on Iran because he didn't see it as 'proportionate'. *NPR*. Retrieved from https://www.npr.org/2019/06/21/734683701/trump-reportedly-orders-strike-on-iran-then-calls-off-attack-plan

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and

human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science, 19*(3), 259-282. doi:10.1080/1463922X.2017.1315750

Chiou, E. K., & Lee, J. D. (2016). Cooperation in Human-Agent Systems to Support Resilience: A Microworld Experiment. *Human Factors: The Journal of Human Factors and Ergonomics Society, 58*(6), 846-863. doi:10.1177/0018720816649094

Christensen, J. C., & Lyons, J. B. (2017). Trust between Humans and Learning Machines: Developing the Gray Box. *Mechanical Engineering, 139*(06), S9-S13. doi:10.1115/1.2017-Jun-5

CNN Wire Staff. (2011). Obama says U.S. has asked Iran to return drone aircraft. *CNN*. Retrieved from https://edition.cnn.com/2011/12/12/world/meast/iran-us-drone/index.html

Coates, A. J. (2016). *The ethics of war* (2nd ed.). Manchester: Manchester University Press.

Commonwealth of Australia (2019, October 23) *Foreign Affairs, Defence And Trade Legislation Committee,* Senate. Retrieved from https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id%3A%22committees%2Festimate%2F53068544-efe7-4494-a0f2-2dbca4d2607b%2F0000%22

Connelly, B. L., Crook, T. R., Combs, J. G., Ketchen, D. J., & Aguinis, H. (2015). Competence- and Integrity-Based Trust in Interorganizational Relationships: Which Matters More? *Journal of Management*. doi:10.1177/0149206315596813

Connelly, B. L., Miller, T., & Devers, C. E. (2012). Under a cloud of suspicion: trust, distrust, and their interactive effect in interorganizational contracting. *Strategic Management Journal, 33*(7), 820-833.

Crootof, R. (2019, 24 October). Artificial Intelligence research needs responsible publication norms. *Lawfare*. Retrieved from https://www.lawfareblog.com/artificial-intelligence-research-needs-responsible-publication-norms

Daniels, O., & Williams, B. (2020). Day zero ethics for military AI. *War on the Rocks.* Retrieved from https://warontherocks.com/2020/01/day-zero-ethics-for-military-ai/

Danzig, R. (2018). *Technology Roulette: Managing Loss of Control as Many Militaries Pursue Technological Superiority*. Center for a New American Security. Retrieved from https://www.cnas.org/publications/reports/technology-roulette

Davidowitz, A. S. (2014). Abandoning the Mosaic Theory: Why the Mosaic Theory of Securities Analysis Constitutes Illegal Insider Trading and What to Do about It. *Wash. UJL & Pol'y, 46*, 281.

Davis, S. E. (2019). *Individual Differences in Operators' Trust in Autonomous Systems: A Review of the Literature* (DST-Group-TR-3587). Joint and Operations Analysis Division, Defence Science and Technology Group, Department of Defence.

Defense Innovation Board. (2019). *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*. Retrieved from https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF

Department of Defence (2014). *ADDP 5.0 Joint Planning*. (ADDP 5.0). Retrieved from https://www.defence.gov.au/adfwc/Documents/DoctrineLibrary/ADDP/ADDP_5_0.pdf

Department of Defence, (2013) *Australian Defence Doctrine Publication 3.0 Campaigns and Operations Edition 2 AL1*, (p 3.1).

Department of Defence. (2018). NATIONAL SECURITY SCIENCE AND TECHNOLOGY. Policy and priorities. Retrieved from https://www.dst.defence.gov.au/sites/default/files/attachments/documents/NS%20S%26T%20policy%20and%20priorities.pdf

Department of Defence. (2019a). *ADDP 00.1 Command and Control AL1*.

Department of Defence. (2019b). *Australia's System of Control and applications for Autonomous Weapon Systems*. Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, 25–29 March 2019 and 20–21 August 2019, Geneva. https://www.unog.ch/80256EDD006B8954/(httpAssets)/16C9F75124654510C12583C9003A4EBF/$file/CCWGGE.12019WP.2Rev.1.pdf

Department of Defence, (2019c) *Australian Defence Doctrine Publication 00.2 Command and Control Edition 2 AL1*, (p. 1.2).

Department of Industry Innovation and Science. (2019). *AI Ethics Principles*. Retrieved from https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles.

Devitt, S. K. (2013). *Homeostatic epistemology: Reliability, coherence and coordination in a Bayesian virtue epistemology.* (Ph.D.), Rutgers The State

University of New Jersey – New Brunswick, Retrieved from http://eprints.qut.edu.au/62553/

Devitt, S. K. (2018). Trustworthiness of autonomous systems. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of Trusted Autonomy* (pp. 161-184). Cham: Springer International Publishing.

Dijkstra, E. W. (1970). On the reliability of mechanisms. In *Notes on Structured programming* (2nd ed., pp. EWD249-247). Retrieved from http://www.cs.utexas.edu/users/EWD/ewd02xx/EWD249.PDF

Dingle, Sarah. (2019, July 25). US officials reveal there aren't enough spare parts for the F-35 Joint Strike Fighter. *ABC News*. Retrieved from https://www.abc.net.au/news/2019-07-25/there-arent-enough-spare-parts-for-the-joint-strike-fighter/11337686

Dowse, A. (2018). The Need for Trusted Autonomy in Military Cyber Security. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of Trusted Autonomy* (pp. 203-213). Cham: Springer International Publishing.

Drnec, K., Marathe, A. R., Lukos, J. R., & Metcalfe, J. S. (2016). From Trust in Automation to Decision Neuroscience: Applying Cognitive Neuroscience Methods to Understand and Improve Interaction Decisions Involved in Human Automation Interaction. *Frontiers in Human Neuroscience, 10*(290). doi:10.3389/fnhum.2016.00290

Dutton, D. (2003). Authenticity in Art. In J. Levinson (Ed.), *The Oxford Handbook of Aesthetics* (pp. 258-274).

Ekelhof, M. A. (2018). Lifting the fog of targeting. *Naval War College Review, 71*(3), 61-95.

Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction. *Engaging Science, Technology, and Society, 5*, 40-60.

Endsley, M. R. (2016). From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors, 59*(1), 5-27. doi:10.1177/0018720816681350

European Parliament and of the Council. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. Retrieved from https://eur-lex.europa.eu/eli/reg/2016/679/oj

Fawkes, A. J., & Menzel, M. (2018). The future role of artificial intelligence. *The Journal of the Joint Air Power Competence Centre,* 27.

Fjeld, J., Hilligoss, H., Achten, N., Daniel, M. L., Feldman, J., & Kagay, S. (2019). *Principled Artificial Intelligence: A Map of Ethical and Rights Based Approaches.* Retrieved from https://ai-hr.cyber.harvard.edu/primp-viz.html

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, & McDaniel, P. (2018). *Ensemble Adversarial Training: Attacks and Defenses*. 6th International Conference on Learning Representations, Vancouver, Canada. https://openreview.net/forum?id=rkZvSe-RZ&noteId=rkZvSe-RZ

Floridi, L. (2016). 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, 374*(2083), 20160112.

Freedberg Jr, S. J. (2019). The Art of Command, The Science of AI. *Breaking Defence*. Retrieved from https://breakingdefense.com/2019/11/the-art-of-command-the-science-of-ai/

French, S., Maule, J., & Papamichail, N. (2009). *Decision Behaviour, Analysis and Support*. Cambridge: Cambridge University Press.

Friedman, J. A., & Zeckhauser, R. (2018). Analytic Confidence and Political Decision-Making: Theoretical Principles and Experimental Evidence From National Security Professionals. *Political Psychology, 39*(5), 1069-1087. doi:10.1111/pops.12465

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a bayesian approximation: Representing model uncertainty in deep learning.* The 33rd International Conference on Machine Learning (ICML 2016) 19-24 Jun 2016, New York.

Galliott, J. (Ed.) (2019). *Force Short of War in Modern Conflict: Jus Ad Vim*. Edinburgh: Edinburgh University Press.

GDPR.eu (2020). *Art. 22 GDPR. Automated individual decision-making, including profiling*. Retrieved from https://gdpr.eu/article-22-automated-individual-decision-making/

Greenwell-Barnden, J. N., Bender, A., Loft, S., Bowden, V., Whitney, S. J., Lipp, O. V., & Visser, T. A. W. (2019). *One size fits one: The benefits of customizing automation to accommodate differences in operator multitasking*. Defence Human Science Symposium: Human Sciences Impact for the Warfighter, University of Canberra.

Hajek, A., & Hartmann, S. (2009). Bayesian epistemology. In J. Dancy, E. Sosa, & M. Steup (Eds.), *A Companion to Epistemology* (pp. 93-105). Chicester: John Wiley & Sons, Ltd.

Heaven, D. (2020). Why asking an AI to explain itself can make things worse. *Artificial Intelligence / Machine Learning, MIT Technology Review.* Retrieved from https://www.technologyreview.com/s/615110/why-asking-an-ai-to-explain-itself-can-make-things-worse/

Hellyer, M. (2019) Accelerating autonomy: Autonomous systems and the Tiger helicopter replacement. *Australian Strategic Policy Institute.* Retrieved from https://www.aspistrategist.org.au/defence-should-accelerate-australias-adoption-of-autonomous-systems/

Hicks, K., Hunter, A. P., Samp, L. S., & Coll, G. (2017). *Assessing the Third Offset Strategy*. Center for Strategic & International Studies. Retrieved from https://www.csis.org/analysis/assessing-third-offset-strategy

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407-434. doi:10.1177/0018720814547570

Hoffman, R. R., Johnson, M., Bradshaw, J. M., & Underbrink, A. (2013). Trust in Automation. *IEEE Intelligent Systems, 28*(1), 84-88. doi:10.1109/MIS.2013.24

Hoffman, R. R., Sarter, N., Johnson, M., & Hawley, J. K. (2018). Myths of automation and their implications for military procurement. *Bulletin of the Atomic Scientists, 74*(4), 255-261. doi:10.1080/00963402.2018.1486615

Hollnagel, E., Woods, D. D., & Leveson, N. (2006). *Resilience engineering: Concepts and precepts.* Ashgate Publishing, Ltd.

IBM Research Trusted AI. (2019). AI Fairness 360 Open Source Toolkit. Retrieved from https://aif360.mybluemix.net/

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (EADe1)*. Retrieved from https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/autonomous-systems.html

International Committee of the Red Cross. (1977) *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), 8 June 1977.* Retrieved from https://ihl-databases.icrc.org/applic/ihl/ihl.nsf/INTRO/470?OpenDocument

International Committee of the Red Cross. (2019). *Artificial intelligence and machine learning in armed conflict: A human-centred approach*. Retrieved from https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389-399. doi:10.1038/s42256-019-0088-2

Kania, E. B. (2017). *Battlefield singularity: Artificial intelligence, military revolution, and China's future military power*. Center for a New American Security. Retrieved from https://www.cnas.org/publications/reports/battlefield-singularity-artificial-intelligence-military-revolution-and-chinas-future-military-power

Kerr, O. S. (2012). The mosaic theory of the Fourth Amendment. *Mich. L. Rev., 111*, 311.

Kh, R. (2017, 1 Dec). How AI is the Future of Cybersecurity. *Info Security*. Retrieved from https://www.infosecurity-magazine.com/next-gen-infosec/ai-future-cybersecurity/

Kim, A., Wampler, B., Goppert, J., Hwang, I., & Aldridge, H. (2012). *Cyber attack vulnerabilities analysis for unmanned aerial vehicles*. American Institute of Aeronautics and Astronautics. Retrieved from https://arc.aiaa.org/doi/abs/10.2514/6.2012-2438

Kim, P. H., Ferrin, D. L., Cooper, D., & Dirks, K. T. (2004). Removing the Shadow of Suspicion: The Effects of Apology Versus Denial for Repairing Competence- Versus Integrity-Based Trust Violations. *Journal of Applied Psychology, 89*(1), 104-118. doi:10.1037/0021-9010.89.1.104

Kim, S. D. (2012). *Characterizing unknown unknowns*. PMI Global Congress 2012—North America, Vancouver, British Columbia, Canada. https://www.pmi.org/learning/library/characterizing-unknown-unknowns-6077

Kolenda, C. D., Reid, R., Rogers, C., & Retzius, M. (2016, June). The strategic costs of civilian harm: Applying lessons from Afganistan to current and future conflicts. *Open Society Foundations*. Retrieved from

https://www.opensocietyfoundations.org/publications/strategic-costs-civilian-harm

Kugler, M. B., & Strahilevitz, L. J. (2016). Actual Expectations of Privacy, Fourth Amendment Doctrine, and the Mosaic Theory. *The Supreme Court Review, 2015*(1), 205-263.

Langcaster, B. (2018). Cyber Security in your Supply Chain. APMG International. Retrieved from https://apmg-international.com/article/cyber-security-your-supply-chain

Leben, D. (2019). *Ethics for robots: How to design a moral algorithm*. New York: Routledge.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society, 46*(1), 50-80. doi:10.1518/hfes.46.1.50_30392

Lewis, L. (2019). *Protecting Medical Care in Conflict: A Solvable Problem*. Retrieved from CNA, Arlington Virginia: https://humanrightscommission.house.gov/sites/humanrightscommission.house.gov/files/documents/Protecting%20Medical%20Care%20in%20Conflict%20-%20Lewis.pdf

Liivoja, R., & McCormack, T. (Eds.). (2016). *Routledge Handbook of the Law of Armed Conflict*. New York, NY: Routledge.

Linnan, D. K. (1991). Iran Air Flight 655 and Beyond: Free Passage, Mistaken Self-Defense, and State Responsibility. *Yale J. Int'l L., 16*, 245.

Lockwood, T. (2019, January 8). Artificial intelligence can now explain its own decision-making. *IoT for All*. Retrieved from https://www.iotforall.com/artificial-intelligence-can-explain-own-decision-making/

McLellan, C. (2016, December 1). Inside the black box: Understanding AI decision-making. *ZDNet*. Retrieved from https://www.zdnet.com/article/inside-the-black-box-understanding-ai-decision-making/

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1-38.

Oakford, S. (2018, August 17). One American's failed quest to protect civilians in Yemen. *The Atlantic*. Retrieved from https://www.theatlantic.com/international/archive/2018/08/yemen-saudi-airstrike-school-bus/567799/

Ormsby, G. (2019, March 11). Uber fatality unveils AI accountability issues. *Lawyers Weekly*. Retrieved from https://www.lawyersweekly.com.au/biglaw/25213-uber-fatality-unveils-ai-accountability-issues

Rapidminer. (2018). How to correctly validate machine learning models. *Rapidminer*.

Ryan, M. (2018, 20 April). Building a brilliant ADF. *The Strategist*. Retrieved from https://www.aspistrategist.org.au/building-brilliant-adf/

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors, 58*(3), 377-400. doi:10.1177/0018720816634228

Scharre, P., & Horowitz, M. C. (2018). *Artificial Intelligence: What Every Policymaker Needs to Know*. Center for a New American Security.

Schaus, J., & Johnson, K. (2018). *Unmanned aerial systems' influences on conflict escalation dynamics*. Center for Strategic and International Studies. Retrieved from https://www.csis.org/analysis/unmanned-aerial-systems-influences-conflict-escalation-dynamics

Scheidt, D. H., Hibbitts, C. A., Chen, M. H., Bekker, D. L., & Paxton, L. J. (2017). *On the need for Artificial Intelligence and Advanced Test and Evaluation Methods for Space Exploration*. Planetary Science Vision 2050 Workshop. https://pdfs.semanticscholar.org/351d/ca793a251bd5d8e7a9fc08c6f29357209c05.pdf

Scholz, J., & Galliott, J. (2018). *AI in Weapons: the moral imperative for minimally-just autonomy*. International Conference on Science and Innovation for Land Power, Adelaide. https://www.dst.defence.gov.au/sites/default/files/basic_pages/documents/ICSILP18Wed1600_Scholz_et_al-AI_in_Weapons.pdf

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., & Graepel, T. (2020). Mastering atari, go, chess and shogi by planning with a learned model. *Nature*. 588, pp. 604–609. Retrieved from https://www.nature.com/articles/s41586-020-03051-4

Seffers, G. I. (2017). *Smarter AI for Electronic Warfare.* Armed Forces Communications and Electronics Association. Retrieved from https://www.afcea.org/content/smarter-ai-electronic-warfare

Shortliffe, E. (1987). Computer Programs to Support Clinical Decision Making-Reply. *JAMA : the journal of the American Medical Association, 258*, 61-66. doi:10.1001/jama.1987.03400170060016

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips. *Science, 333*(6043), 776-778. doi:10.1126/science.1207745

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). *A unified approach to quantifying algorithmic unfairness: Measuring individual and group unfairness via inequality indices*. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, United Kingdom.

Sroufe, R., & Curkovic, S. (2008). An examination of ISO 9000: 2000 and supply chain quality assurance. *Journal of operations management, 26*(4), 503-520.

Staines, D., Formosa, P., & Ryan, M. (2019). Morality Play: A Model for Developing Games of Moral Expertise. *Games and Culture, 14*(4), 410-429.

Stanton, N., Salmon, P. M., & Rafferty, L. A. (2013). Cognitive Task Analysis Methods. *Human factors methods: a practical guide for engineering and design*. Ashgate Publishing, Ltd.

Sutton, R., & Barto, A. (2018). *Reinforcement learning: an introduction, ser. Adaptive computation and machine learning series*. Cambridge, Massachusetts: The MIT Press.

Tavani, H. T. (2015). Levels of Trust in the Context of Machine Ethics. *Philosophy & Technology, 28*(1), 75-90. doi:10.1007/s13347-014-0165-8

The Australian Government. (1903). *The Defence Act*. (C1903A00020 No. 20). Retrieved from https://www.legislation.gov.au/Series/C1903A00020.

The British Academy, & The Royal Society. (2017). *Data management and use: Governance in the 21st century*. Retrieved from https://royalsociety.org/-/media/policy/projects/data-governance/data-management-governance.pdf

The Institute for Ethical AI & Machine Learning. (2019). *The AI-RFX Procurement Framework: Practical templates to support AI procurement*. Retrieved from https://ethical.institute/rfx.html

Thomas, R. C. (2013). *The rainforest of ignorance and uncertainty*. Retrieved from https://exploringpossibilityspace.blogspot.com/2013/07/the-rainforest-of-ignorance-and.html

Turek, M. (2017). *Explainable Artificial Intelligence (XAI)*. DARPA. Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence

Turek, M. (2019). *Explainable Artificial Intelligence (XAI)*. DARPA. Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence

United States Navy. (2019). AEGIS Weapon System (AWS). *United States Navy Fact File.* Retrieved from https://www.navy.mil/navydata/fact_display.asp?cid=2100&tid=200&ct=2

Vallor, S. (2018). *An Ethical Toolkit for Engineering/Design Practic*e. Markkula Center for Applied Ethics. Retrieved from https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and reliance on an automated combat identification system. *Human Factors, 51*(3), 281-291. doi:10.1177/0018720809338842

Wilkinson, D. (2019). The counter-intuitive side of evidence-based practice. *The Oxford Review.* Retrieved from https://www.oxford-review.com/counter-intuitive-side-evidence-based-practice/

Winkelman, Z., Buenaventura, M., Anderson, J. M., Beyene, N. M., Katkar, P., & Baumann, G. C. (2019). *When Autonomous Vehicles Are Hacked, Who Is Liable?* RAND Corporation.

Yanardag, P., Cebrian, M., & Rahwan, I. (2018). *Norman: Worlds first psychopath AI*. MIT Media Lab, Scalable Cooperation. Retrieved from http://norman-ai.mit.edu/

Zender, A. (2019). Anti-ship missile defence with artificial intelligence. *Beyond the Horizon ISSG*. Retrieved from https://behorizon.org/anti-ship-missile-defense-with-artificial-intelligence/

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.* PublicAffairs.

APPENDIX A. COMPARISON OF ETHICAL AI FRAMEWORKS

Facets of Ethical AI & Topics emerging from the workshop	**Australian Government*'s AI Ethics Principles* (Department of Industry Innovation and Science, 2019)**	**IEEE's *Ethically Aligned Design* Principles (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2019)**	**US Defense Ethical AI Principles (Defense Innovation Board, 2019)**	***Principled Artificial Intelligence: A Map of Ethical and Rights Based Approaches* (Fjeld et al., 2019)**	**The global landscape of AI ethics guidelines (Jobin et al., 2019).**
RESPONSIBILITY: Who is responsible for AI? **Education** **Command**	Human, social and environmental wellbeing: Throughout their lifecycle, AI systems should benefit individuals, society and the environment Human-centred values: Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals	Human rights: Autonomous and Intelligent Systems (A/IS) shall be created and operated to respect, promote, and protect internationally recognized human rights. Well-being: A/IS creators shall adopt increased human well-being as a primary success criterion for development.	RESPONSIBLE: Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of DoD AI systems	Promotion of human values Professional responsibility	Responsibility

GOVERNANCE: How is AI controlled? **Effectiveness** **Integration** **Transparency** **Human Factors** **Scope** **Confidence** **Resilience** **Empowerment**	Transparency and explainability: There should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them	Effectiveness: A/IS creators and operators shall provide evidence of the effectiveness and fitness for purpose of A/IS Transparency: The basis of a particular A/IS decision should always be discoverable Competence: A/IS creators shall specify and operators shall adhere to the knowledge and skill required for safe and effective operation.	GOVERNABLE: DoD AI systems should be designed and engineered to fulfil their intended function while possessing the ability to detect and avoid unintended harm or disruption, and for human or automated disengagement or deactivation of deployed systems that demonstrate unintended escalatory or other behaviour	Human Control of Technology Transparency	Transparency
TRUST: How can AI be trusted? **Sovereign Capability** **Safety**	Reliability and safety: Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose Fairness: Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or	Data Agency: A/IS creators shall empower individuals with the ability to access and securely share their data, to maintain people's capacity to	EQUITABLE: DoD should take deliberate steps to avoid unintended bias in the development and deployment of combat or non-combat AI systems that would inadvertently cause harm to persons.	Fairness and non-discrimination Safety and Security Privacy	Justice and fairness Non-maleficence Privacy

Supply Chain **Test & Evaluation** **Misuse and risks** **Authority pathway** **Data subjects**	result in unfair discrimination against individuals, communities or groups Privacy protection and security: Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data Contestability: When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system	have control over their identity Awareness of Misuse: A/IS creators shall guard against all potential misuses and risks of A/IS in operation.	RELIABLE: DoD AI systems should have an explicit, well-defined domain of use, and the safety, security, and robustness of such systems should be tested and assured across their entire life cycle within that domain of use		
LAW: How can AI be used lawfully? **Protected Symbols and Surrender** **De-escalation**	No equivalent		No equivalent	No equivalent	No equivalent
TRACEABLILITY: How are the actions of AI recorded?	Accountability: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI	Accountability: A/IS shall be created and operated to provide an unambiguous rationale for all decisions made.	TRACEABLE: DoD's AI engineering discipline should be sufficiently advanced such that technical experts possess	Accountability Explainability	

Explainability **Accountability**	systems, and human oversight of AI systems should be enabled	an appropriate understanding of the technology, development processes, and operational methods of its AI systems, including transparent and auditable methodologies, data sources, and design procedure and documentation

APPENDIX B. ETHICAL AI RISK MATRIX

Table 4 Ethical AI risk matrix with advice on how to complete each section

Activities	Ethical Issue to be addressed	Risks	Actions	Timeline	Outcome	Assignee	Status
Define the activity you are undertaking	*Indicate the ethical issue the activity is intended to address*	*Estimate the risk to the project objectives if issue is not addressed*	*Define specific actions you will undertake to support the activity*	*Provide a timeline the activity*	*Define action and activity outcomes*	*Identify the responsible party/ies*	*Provide a status update*

Table 5 An example of a completed Ethical AI Risk Matrix

Activities	Ethical Issue to be addressed	Risks	Actions	Timeline	Outcome	Assignee	Status
AI project activity	Identify and mitigate against unjust biases in the AI algorithm.	High	Evaluate bias and conduct de-biasing[22]	Q1 + Q2	Reduction of risk to AI project, e.g. increase stakeholder buy-in	Stakeholder #1	50% complete

[22] e.g. https://github.com/EthicalML/explainability-and-bias

APPENDIX C. SPEAKERS AND FACILITATORS AT ETHICAL AI FOR DEFENCE WORKSHOP

GPCAPT Jerome Reid, Director, Plan Jericho, Royal Australian Air Force

Mr Bob Bolia, Research Leader Aerospace Effectiveness, Aerospace Division, Defence Science & Technology Group (DST)

Prof Jason Scholz, CEO Trusted Autonomous Systems Defence Cooperative Research Centre (TASDCRC)

Dr Larry Lewis, Vice President and Director of the Center for Autonomy and Artificial Intelligence, Centre for Naval Analyses, USA

WGCDR Julian Tattersall, Australian Defence Force.

CHAP Nikki Coleman, Royal Australian Air Force

Dr Susan Cockshell, Group Leader Human and Autonomous Decision Superiority, Defence Science & Technology Group

Dr Fiona Kerr, Founder & Director, Neurotech Institute

Dr Jai Galliott Research Leader Values in Defence & Security Technology Group, Australian Defence Force Academy, University of New South Wales, Co-lead Ethics and Law of Trusted Autonomous Systems Activity, TASDCRC)

A/Prof Saba Bazargan, Department of Philosophy, University of California, San Diego, USA

Dr Derek Leben, Ethics of autonomous systems, University of Pittsburgh, USA

A/Prof Seth Lazar, Project Lead Humanising Machine Intelligence, ANU

Ms Ellen Broad, Data ethics, Data61

Professor Jacob Hohwy, Principle Investigator Cognition & Philosophy Lab, Monash University

Dr Tim van Gelder Research Lead The Swarm Project, University of Melbourne

Dr Shane Dunn, Scientific Advisor, Joint Division, Defence Science & Technology Group

WGCDR Julian Tattersall, Royal Australian Air Force

Dr Tristan Perez, Research Lead Assurance of Autonomy, Trusted Autonomous Systems Defence Cooperative Research Centre

Mr Rick Shaw, Partner of Consulting and part of the Actuaries practice, DeLoitte

A/Prof Rain Liivoja, Law, University of Queensland, Co-lead Ethics and Law of Trusted Autonomous Systems Activity, TASDCRC

APPENDIX D. ORGANISATIONS IN ATTENDANCE AT THE WORKSHOP

- Australian Army
- Attorney General's Department
- Australian Defence Force Academy
- Australian Defence Magazine
- Australian Maritime Safety Authority
- Australian National University
- Australian Strategic Policy Institute
- BAE Systems
- Boeing Phantom Works
- Centre for Defence Leadership & Ethics
- CNA Analysis & Solutions
- Cyborg Dynamics Engineering
- Dalhousie University
- Data61/CSIRO
- Deakin University
- Defence AI Centre
- Defence Force Recruiting
- Defence Legal
- Defence People Group
- Defence Science & Technology Group
- Defence Signal & Cyber Command
- DefendTex
- Deloitte
- Department of Foreign Affairs & Trade
- Defence Science and Technology Group
- FAL Lawyers
- International Committee of the Red Cross
- Joint Capabilities Joint Information Warfare Information Warfare Joint Intelligence
- The Mandarin
- Monash University
- Neurotech Institute
- Royal Australian Air Force
- Royal Australian Navy
- Royal Melbourne Institute for Technology
- Skyborne Technologies
- Thales
- The Cranlana Centre for Ethical Leadership
- Trusted Autonomous Systems Defence Cooperative Research Centre
- University of California San Diego
- University of Melbourne
- University of New South Wales
- University of Pennsylvania
- University of Pittsburgh
- University of Queensland
- University of Technology Sydney

APPENDIX E. CONTEXTS OF AI IN DEFENCE

The vast number of potential military AI applications made it necessary to develop an effective taxonomy for the Ethical AI for Defence workshop. The taxonomy was required to ensure the workshop addressed the widest range of applications of AI in Defence from warfighting to business applications, and to avoid exclusively focusing on narrow applications such as autonomous weapons systems. It was also designed to identify if different applications of AI in Defence should warrant different treatment in relation to ethical issues.

To ensure relevance, the ADF warfighting functions were selected as the starting point for the taxonomy. These were referred to as the *Contexts for AI in Defence*. The contexts were designed to capture all the potential Defence applications of AI, and customised to take into consideration practical limitations such as the background and number of participants. Contexts should not necessarily be used as a universal taxonomy for AI in Defence.

The ADF warfighting functions are defined as 'capabilities and activities conducive to [military] success at the operational level' (Defence, 2013, p. 3.1) where 'each function is a set of related joint capabilities and activities grouped together to help joint commanders integrate, synchronise and direct campaigns and operations' (Defence, 2019c, p. 1.2). The ADF currently recognises 6 warfighting functions: command, situational understanding, force generation and sustainment, force projection, force protection and force application.

For the purpose of the workshop, *force application*, *force protection* and *situational understanding* were retained, and *force generation and sustainment* was subdivided into *force sustainment* and three enterprise-level contexts: personnel, enterprise logistics and business process improvement. This was done to emphasise the role of AI in the non-warfighting functions in the ADF. With the addition of an 'Other' category this resulted in 8 contexts for AI in Defence.

After some discussion, the decision was made to incorporate *force projection* in the workshop *force sustainment* context, and to not include the *command* warfighting function. Not incorporating the *command* function was somewhat controversial, as there is significant evidence that it will be the function most impacted by AI. However, given the command function has significant overlaps with all the other functions, it was excluded to avoid duplication and confusion, particularly with non-military participants.[23] General

[23] The warfighting functions within ADF doctrine are currently under review. It is anticipated that they will be renamed the 'joint functions' and the 6 existing functions will remain unchanged with the possible addition of

feedback on the taxonomy during the workshop was positive. Participants with military expertise were invited to share their in-depth knowledge of each domain with participants. The resulting 8 contexts were subsequently divided into combat/warfighting and enterprise level/rear echelon contexts, which were used in the active discussion sessions and the online platform. The Contexts of AI in Defence are listed below:

E.1. Combat/Warfighting

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

an additional function capturing capabilities and activities in the information domain. The current approved (2012) warfighting functions were be used for the purpose of the Ethical AI for Defence Workshop.

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

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

Tag	Situational Understanding (SU)
Description	The accurate interpretation of a situation and the likely actions of groups and individuals within it. Situational Understanding enables timely and accurate decision making.
AI examples	AI that enables or supports Intelligence, Surveillance and Reconnaissance (ISR) activities including: object recognition and categorisation of still and full motion video removal of unwanted sensor data identification of enemy deception activities anomaly detection and alerts monitoring of social media and other open-source media channels optimisation of collection assets AI that fuses data and disseminates intelligence to strategic, operational and tactical decision makers Decision support tools Battle Management Systems AI that supports Command and Control functions Algorithms used to predict likely actions of groups and individuals AI used to assess individual and collective behaviour and attitudes

E.2. Enterprise-level and Rear Echelon Functions

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

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

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

APPENDIX F. A TAXONOMY OF DECISION PROBLEMS

Content contributed by Tristan Perez (reformatted for report), see also French, S., Maule, J., & Papamichail, N. (2009). *Decision Behaviour, Analysis and Support*. Cambridge: Cambridge University Press

Decision-maker/s	**Type of Decision**	**Example/s**
Single decision-maker	*Single-stage* once-off decisions	A decision as to whether continue with current mission objectives or consider alternatives given changes in the operational conditions.
		A decision about deploying a particular type of weapon towards a hostile asset
	Multi-stage sequential decisions in time	Management of a supply chain to support a replenishment of supplies for a mission over number of days or months
		Motion control of a network of autonomous systems to deliver un-interruptible communications for C2
		Missile guidance towards a fixed target
Multi-decision maker	*Decisions under conflict* Games Cooperative vs. non-cooperative iterated vs. non-iterated Zero sum vs non-zero sum Two vs N players	Once-off games, e.g.
		Two governments negotiating over a contested land or sea area
		Sequential games, e.g.
		Two aircraft/marine craft in a pursue and evade situation
		Multiple autonomous systems avoiding collisions while seeking to attain individual mission goals
		Managing a network of military assets during engagement
	Consensus decisions social choice	A resolution of the UN Security Council
		A number of countries developing guidelines for the conduct of trials of autonomous systems at the International Maritime Organisation Meeting
		A group of manned assets and group of AS deciding how to engage with a hostile asset
		A jury deciding for guilt or innocence
		Prime minister and council decision to escalating war

APPENDIX G. DATA ITEM DESCRIPTION DID-ENG-SW-LEAPP

This is an example of a possible Legal and Ethical Assurance Program Plan (LEAPP), and is not an official Defence document nor has the content been approved for official use.

DATA ITEM DESCRIPTION

1. DID NUMBER: DID-ENG-SW-LEAPP

2. TITLE: LEGAL AND ETHICAL ASSURANCE PROGRAM PLAN FOR ARTIFICIAL INTELLIGENCE SYSTEMS

3. DESCRIPTION AND INTENDED USE

3.1 The Legal and Ethical Assurance Program Plan (LEAPP) describes the Contractor's plan for assuring that Software acquired under the contract that is categorised as Artificial Intelligence (AI) meets the Commonwealth's Legal and Ethical Assurance (LEA) requirements.

3.2 For Contractors acquiring and/or supplying Software classified as AI under the Contract, the LEAPP is expected to describe the approach, plans and procedures to be applied to the management of the AI Software being acquired and/or supplied. This would typically include the monitoring and review of Subcontractors developing AI Software, the Configuration Management of acquired AI Software, and the integration and Verification of this AI Software with other elements being supplied under the Contract.

3.3 The Commonwealth uses the LEAPP:

a. to provide visibility into the Contractor's technical planning;

b. for progress and risk assessment purposes; and

c. to provide input into the Commonwealth's own planning.

4. INTER-RELATIONSHIPS

4.1 The LEAPP is subordinate to the following data items, where these data items are required under the Contract:

a. Software Management Plan (SMP);

b. Integrated Support Plan (ISP);

c. Configuration Management Plan (CMP); and

d. Verification and Validation Plan (V&VP).

5. APPLICABLE DOCUMENTS

5.1 The following documents form a part of this DID to the extent specified herein:

XXXX *Australian Defence Framework for Ethical AI*

6. PREPARATION INSTRUCTIONS

6.1 Generic Format and Content

6.1.1 The data item shall comply with the general format, content and preparation instructions contained in the CDRL clause entitled 'General Requirements for Data Items'.

6.1.2 The data item shall include a traceability matrix that defines how each specific content requirement, as contained in this DID, is addressed by sections within the data item.

6.2 Specific Content

6.2.1 Description of AI Functionality

6.2.1.1 The LEAPP shall describe the relevant context and environments that the AI software will be required to function in.

6.2.1.2 The LEAPP shall describe the nature of decisions that the AI Software will be making or supporting.

6.2.2 Integration with human operators

6.2.2.1 The LEAPP shall describe how the AI software integrates with human operators to ensure effectiveness and legal and ethical decision-making in the anticipated contexts.

6.2.3 Countermeasures against misuse

6.2.3.1 The LEAPP shall describe the countermeasures within the AI software to prevent misuse.

6.2.4 Ethical Frameworks and subject matter experts

6.2.4.1 The LEAPP shall describe scientific and/or academic ethical frameworks used to develop the AI Software.

6.2.4.2 The LEAPP shall identify the legal and ethical subject matter experts used by the Contractor to guide AI Software development.

6.2.5 Verification and Validation of LEA aspects of AI

6.2.5.1 The LEAPP shall describe LEA Verification and Validation (V&V) of AI software of as an integrated effort within the Contractor's V&V program.

6.2.5.2 The LEAPP shall identify design milestones at which LEA tests are to be performed to assess compatibility among human performance requirements, personnel aptitude and skill requirements, training requirements, and equipment design aspects of personnel equipment and Software interfaces.

6.2.5.3 The LEAPP shall identify major V&V objectives and describe the V&V methods to be applied for the LEA program.

6.2.6 Legal and Ethical Assurance in Subcontractor Efforts

6.2.6.1 The LEAPP shall define how all work conducted by Subcontractors shall be scoped, managed and monitored to ensure the Contract objectives are met.

6.2.6.2 The LEAPP shall define how the Subcontractor documentation relating to legal and ethical assurance will be controlled and integrated into the overall project documentation.

6.2.7 Expectations of the Contractor

6.2.7.1 The LEAPP shall identify the expectations of the Contractor with respect to the Commonwealth in order to ensure the AI LEA l objectives are met.

6.2.8 Legal and Ethical Assurance in System Analysis

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

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

a. Who is responsible for AI?

b. How is AI controlled?;

c. How can AI be trusted?;

d. How can AI be used lawfully?;

e. How are the actions of AI recorded?

6.2.9 Derivation of Personnel and Training Requirements

6.2.9.1 The LEAPP shall describe the methods by which the Contractor shall ensure that operator and maintainer Personnel and Training requirements are based upon LEA requirements developed from system analysis data.

6.2.10 Legal and Ethical Assurance Working Group

6.2.10.1 Where the SOW requires the Contractor to establish a LEA Working Group (LEAWG), the LEAPP shall include a plan for the LEAWG, including:

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

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

c. arrangements for the conduct of LEAWG meetings.

APPENDIX H. DETAILED JUDGING CRITERIA

At the conclusion of the workshop, a prize was awarded to the most scientific and collaborative user of the BetterBeliefs platform. This was measured by 1) the *quality* and *quantity* of their evidence and 2) by their rating of evidence suggested by others. A live digital leaderboard was available to participants at the workshop.

To be eligible for the prize, users had to: add multiple pieces of evidence, get their evidence evaluated by other users and rate the quality of evidence suggested by others.

Points were assigned out of a maximum of 100 pts. The user with the highest number of points won the prize.

CRITERIA #1 Quality of Evidence (50 pts) The quality of evidence added by a user was measured by the way *other* users rated its quality. 30 points was assigned to the average quality of a user's evidence (star rating averaged over all evidence added when ratings >1). E.g. 5 stars = 30 pts, 4.5 stars = 27 pts, 4 stars = 24 pts, 3.5 stars = 21 pts, 3 stars = 18 pts, 2.5 stars = 15 pts, 2 stars = 12 pts, 9 pts = 1.5 stars, 1 star = 6 pts, 0.5 star = 3 pts.

20 pts was assigned to the *number* of unique users who rated the evidence. Points were assigned by ranking users from most ratings by others to least ratings by others, then apportioning points based on quartile rank: Q1 = 20 pts, Q2 = 15 pts, Q3 = 10 pts, Q4 = 5 pts.

CRITERIA #2 Quantity of evidence (30 pts) The total number of items of evidence added by a user. Points were assigned by ranking users from most items of evidence added to least items of evidence added, then apportioning points based on their quartile rank: Q1 = 30 pts, Q2 = 22.5 pts, Q3 = 15 pts, Q4 = 7.5 pts.

CRITERIA #3 Rate other people's evidence (20 pts) The total number of unique items of evidence rated by each user. Points were assigned by ranking users from most number of unique items of evidence rated to least number of unique items of evidence rated and then apportioning points based on their quartile rank: Q1 = 20 pts, Q2 = 15 pts, Q3 = 10 pts, Q4 = 5 pts.

APPENDIX I. DECLARATION OF PERCEIVED CONFLICT OF INTERESTS FOR DR. S. KATE DEVITT

- Kate Devitt co-designed and led the build of the BetterBeliefs social platform (BB) while an academic researcher at Queensland University of Technology (QUT) 2017-2018 with other QUT staff: Ms Tamara Pearce, Distinguished Professor Kerrie Mengersen, Dr Alok Chowdhury, under a Commercial Research agreements with Expedia Inc., Queensland Fire & Emergency Services, The World Hospital Congress 2018 and investment by QUT Bluebox.
- BB is an instantiation of theories from Kate's PhD thesis in philosophy grounded in Bayesian & Virtue epistemology to improve collective evidence-based decision-making under uncertainty. However, the platform draws on transdisciplinary research from Philosophy, Business Innovation, Design, Bayesian Statistics, Cognitive Science and Information Technology from all four contributing researchers.
- The BB IP is owned by QUT
- As of Friday 28 June 2019 the company 'BetterBeliefs Pty. Ltd.' has a licensing agreement with QUT to use the BB IP.
- Kate Devitt is CEO of BetterBeliefs Pty. Ltd.
- Kate Devitt is a permanent employee of Defence Science and Technology Group (DSTG) Aerospace Division (AD) Human Factors Group as a 'Social & Ethical Robotic Researcher' and declared background IP to her supervisor Dr Helen Pongracic upon employment at DSTG 26 November 2018. She discussed and had approved the specific use of BetterBeliefs for this workshop with Dr Craig Rogers (then Director Commercialisation and Intellectual Property, Technology Partnerships Office, DSTG).
- BB is being used as a trial tool 2 July – 31 August 2019 at no cost to DSTG, Jericho & TASDCRC as a trial business process for fast evidence-based ideation for physical and virtual workshop participants to capture workshop data.
- Kate Devitt used the tool during the workshop to improve the effectiveness and efficiency of communication and documentation around *Ethical AI for Defence* as a DSTG staff member and not for commercial gain.
- Kate Devitt declared her interest in the BB platform to participants in the workshop (email, in person, & over the phone) and made clear to them that she was using the tool as a business tool only, as a DSTG staff member and not for commercial gain.
- Workshop participants who wished to discuss commercial aspects of the platform were directed to BetterBeliefs Chief Operating Officer and Business Development Manager Ms Tamara Pearce tamara@betterbeliefs.com.au. Kate Devitt commits to recusing herself from any commercial conversations regarding the BetterBeliefs platform during her duties

to DSTG. Any commercial activities undertaken by Kate Devitt with BetterBeliefs are managed by application for secondary employment as per DSTG policy guidelines.

- Ms Tamara Pearce manages the agreement with Jericho for trial use of the platform including information security, and information management (including take-down of the platform post-workshop).

DEFENCE SCIENCE AND TECHNOLOGY GROUP DOCUMENT CONTROL DATA		IMM/CAVEAT (OF DOCUMENT)
TITLE A Method for Ethical AI in Defence		SECURITY CLASSIFICATION Document (O) Title (O)
AUTHOR(S) Kate Devitt, Michael Gan, Jason Scholz and Robert Bolia		PRODUCED BY Defence Science and Technology Group Department of Defence PO Box 7931 Canberra BC ACT 2610
DSTG NUMBER DSTG-TR-3786	REPORT TYPE Technical Report	DOCUMENT DATE January 2021
TASK NUMBER	TASK SPONSOR Plan Jericho Air Force Trusted Autonomous Systems Defence Cooperative Research Centre	RESEARCH DIVISION Aerospace Division
MAJOR SCIENCE AND TECHNOLOGY CAPABILITY Aerospace Systems 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 Ethics, artificial intelligence, artificial intelligence systems, autonomous operations, philosophy		

This is a report on the outcomes of a workshop only and does not represent an official position of Defence. It represents views expressed by participants and stakeholders of the workshop.